DocCode – SCORE

SCORE Placeholder Sheet for IFW Content

Application Number: 17460145 Document Date: 08/27/2021

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

Doc Code: TRACK1.REQ
Document Description: TrackOne Request

PTO/AIA/424 (04-14)

CERTIFICATION AND REQUEST FOR PRIORITIZED EXAMINATION UNDER 37 CFR 1.102(e) (Page 1 of 1)

First Named Inventor:	Maxim Serebryakov	Nonprovisional Application Number (if known):	
Title of Invention:	Real-Time Accent Conversion Model		

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS PRIORITIZED EXAMINATION FOR THE ABOVE-IDENTIFIED APPLICATION.

1. The processing fee set forth in 37 CFR 1.17(i)(1) and the prioritized examination fee set forth in 37 CFR 1.17(c) have been filed with the request. The publication fee requirement is met because that fee, set forth in 37 CFR 1.18(d), is currently $0. The basic filing fee, search fee, and examination fee are filed with the request or have been already been paid. I understand that any required excess claims fees or application size fee must be paid for the application.

2. I understand that the application may not contain, or be amended to contain, more than four independent claims, more than thirty total claims, or any multiple dependent claims, and that any request for an extension of time will cause an outstanding Track I request to be dismissed.

3. The applicable box is checked below:

I. ☑ Original Application (Track One) - Prioritized Examination under § 1.102(e)(1)

i. (a) The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a). This certification and request is being filed with the utility application via EFS-Web.

---OR---

(b) The application is an original nonprovisional plant application filed under 35 U.S.C. 111(a). This certification and request is being filed with the plant application in paper.

ii. An executed inventor's oath or declaration under 37 CFR 1.63 or 37 CFR 1.64 for each inventor, **or** the application data sheet meeting the conditions specified in 37 CFR 1.53(f)(3)(i) is filed with the application.

II. ☐ Request for Continued Examination - Prioritized Examination under § 1.102(e)(2)

i. A request for continued examination has been filed with, or prior to, this form.
ii. If the application is a utility application, this certification and request is being filed via EFS-Web.
iii. The application is an original nonprovisional utility application filed under 35 U.S.C. 111(a), or is a national stage entry under 35 U.S.C. 371.
iv. This certification and request is being filed prior to the mailing of a first Office action responsive to the request for continued examination.
v. No prior request for continued examination has been granted prioritized examination status under 37 CFR 1.102(e)(2).

Signature /Brandon J. Kennedy/	Date August 27, 2021
Name (Print/Typed) Brandon J. Kennedy	Practitioner Registration Number 67894

Note: *This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required.**

☑ *Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76.
This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.

Secrecy Order 37 CFR 5.2:

☐ Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)

Inventor Information:

Inventor 1 Remove

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Maxim		Serebryakov	

Residence Information (Select One) ● US Residency Non US Residency Active US Military Service

City	Palo Alto	State/Province	CA	Country of Residence	US

Mailing Address of Inventor:

Address 1	2819 Diavila Ct.		
Address 2			
City	Pleasanton	State/Province	CA
Postal Code	94588	Country	US

Inventor 2 Remove

Legal Name

Prefix	Given Name	Middle Name	Family Name	Suffix
	Shawn		Zhang	

Residence Information (Select One) ◉ US Residency Non US Residency Active US Military Service

City	Pleasanton	State/Province	CA	Country of Residence	US

Mailing Address of Inventor:

Address 1	2819 Diavila Ct.		
Address 2			
City	Pleasanton	State/Province	CA
Postal Code	94588	Country	US

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button. Add

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

☐ **An Address is being provided for the correspondence Information of this application.**

Customer Number	130085		
Email Address	ls3_docketing@cardinal-ip.com	Add Email	Remove Email

Application Information:

Title of the Invention	Real-Time Accent Conversion Model		
Attorney Docket Number	Sanas-00003	**Small Entity Status Claimed**	☒
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	4	**Suggested Figure for Publication (if any)**	

Filing By Reference:

Only complete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32).
Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

Please Select One:	● Customer Number	US Patent Practitioner	○ Limited Recognition (37 CFR 11.9)
Customer Number	130085		

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, 365(c), or 386(c) or indicate National Stage entry from a PCT application. Providing benefit claim information in the Application Data Sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78.
When referring to the current application, please leave the "Application Number" field blank.

Prior Application Status	Pending		Remove
Application Number	Continuity Type	Prior Application Number	Filing or 371(c) Date (YYYY-MM-DD)
	Claims benefit of provisional	63185345	2021-05-06

Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the **Add** button. Add

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55. When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX)[i] the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(i)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

Remove

Application Number	Country[i]	Filing Date (YYYY-MM-DD)	Access Code[i] (if applicable)

Additional Foreign Priority Data may be generated within this form by selecting the **Add** button. Add

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

☐ This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013.
NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

Authorization or Opt-Out of Authorization to Permit Access:

When this Application Data Sheet is properly signed and filed with the application, applicant has provided written authority to permit a participating foreign intellectual property (IP) office access to the instant application-as-filed (see paragraph A in subsection 1 below) and the European Patent Office (EPO) access to any search results from the instant application (see paragraph B in subsection 1 below).

Should applicant choose not to provide an authorization identified in subsection 1 below, applicant **must opt-out** of the authorization by checking the corresponding box A or B or both in subsection 2 below.

NOTE: This section of the Application Data Sheet is **ONLY** reviewed and processed with the **INITIAL** filing of an application. After the initial filing of an application, an Application Data Sheet cannot be used to provide or rescind authorization for access by a foreign IP office(s). Instead, Form PTO/SB/39 or PTO/SB/69 must be used as appropriate.

1. Authorization to Permit Access by a Foreign Intellectual Property Office(s)

A. Priority Document Exchange (PDX) - Unless box A in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the World Intellectual Property Organization (WIPO), and any other foreign intellectual property office participating with the USPTO in a bilateral or multilateral priority document exchange agreement in which a foreign application claiming priority to the instant patent application is filed, access to: (1) the instant patent application-as-filed and its related bibliographic data, (2) any foreign or domestic application to which priority or benefit is claimed by the instant application and its related bibliographic data, and (3) the date of filing of this Authorization. See 37 CFR 1.14(h)(1).

B. Search Results from U.S. Application to EPO - Unless box B in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the EPO access to the bibliographic data and search results from the instant patent application when a European patent application claiming priority to the instant patent application is filed. See 37 CFR 1.14(h)(2).

The applicant is reminded that the EPO's Rule 141(1) EPC (European Patent Convention) requires applicants to submit a copy of search results from the instant application without delay in a European patent application that claims priority to the instant application.

2. Opt-Out of Authorizations to Permit Access by a Foreign Intellectual Property Office(s)

☐ A. Applicant **DOES NOT** authorize the USPTO to permit a participating foreign IP office access to the instant application-as-filed. If this box is checked, the USPTO will not be providing a participating foreign IP office with any documents and information identified in subsection 1A above.

☐ B. Applicant **DOES NOT** authorize the USPTO to transmit to the EPO any search results from the instant patent application. If this box is checked, the USPTO will not be providing the EPO with search results from the instant application.

NOTE: Once the application has published or is otherwise publicly available, the USPTO may provide access to the application in accordance with 37 CFR 1.14.

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

Applicant 1 Remove

If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.

Clear

● Assignee	Legal Representative under 35 U.S.C. 117	Joint Inventor
Person to whom the inventor is obligated to assign.	Person who shows sufficient proprietary interest	

If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:

Name of the Deceased or Legally Incapacitated Inventor:

If the Applicant is an Organization check here. ☒

Organization Name	Sanas.ai Inc.

Mailing Address Information For Applicant:

Address 1	2819 Diavila Ct.		
Address 2			
City	Pleasanton	**State/Province**	CA
Country	US	Postal Code	94588
Phone Number		Fax Number	
Email Address			

Additional Applicant Data may be generated within this form by selecting the Add button. Add

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

Assignee 1

Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication. An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication.

Remove

If the Assignee or Non-Applicant Assignee is an Organization check here. ☐

Prefix	**Given Name**	Middle Name	**Family Name**	Suffix

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1			
Address 2			
City		**State/Province**	
Country i		Postal Code	
Phone Number		Fax Number	
Email Address			

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button. Add

Signature:

Remove

NOTE: This Application Data Sheet must be signed in accordance with 37 CFR 1.33(b). **However, if this Application Data Sheet is submitted with the INITIAL filing of the application and either box A or B is not checked in subsection 2 of the "Authorization or Opt-Out of Authorization to Permit Access" section, then this form must also be signed in accordance with 37 CFR 1.14(c).**

This Application Data Sheet **must** be signed by a patent practitioner if one or more of the applicants is a **juristic entity** (e.g., corporation or association). If the applicant is two or more joint inventors, this form must be signed by a patent practitioner, **all** joint inventors who are the applicant, or one or more joint inventor-applicants who have been given power of attorney (e.g., see USPTO Form PTO/AIA/81) on behalf of **all** joint inventor-applicants.

See 37 CFR 1.4(d) for the manner of making signatures and certifications.

Signature	/Brandon J. Kennedy/			Date (YYYY-MM-DD)	2021-08-27
First Name	Brandon	Last Name	Kennedy	Registration Number	67894

Additional Signature may be generated within this form by selecting the Add button. Add

PTO/AIA/14 (02-18)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76	Attorney Docket Number	Sanas-00003
	Application Number	
Title of Invention	Real-Time Accent Conversion Model	

This collection of information is required by 37 CFR 1.76. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 23 minutes to complete, including gathering, preparing, and submitting the completed application data sheet form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1 The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3 A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent CooperationTreaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82A (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

TRANSMITTAL FOR POWER OF ATTORNEY TO ONE OR MORE REGISTERED PRACTITIONERS

NOTE: This form is to be submitted with the Power of Attorney by Applicant form (PTO/AIA/82B) to identify the application to which the Power of Attorney is directed, in accordance with 37 CFR 1.5, unless the application number and filing date are identified in the Power of Attorney by Applicant form. If neither form PTO/AIA/82A nor form PTO/AIA82B identifies the application to which the Power of Attorney is directed, the Power of Attorney will not be recognized in the application.

Application Number	
Filing Date	August 27, 2021
First Named Inventor	Maxim Serebryakov
Title	Real-Time Accent Conversion Model
Art Unit	
Examiner Name	
Attorney Docket Number	Sanas-00003

SIGNATURE of Applicant or Patent Practitioner

Signature	/Brandon J. Kennedy/	Date (Optional)	2021-08-27
Name	Brandon J. Kennedy	Registration Number	67,894
Title (if Applicant is a juristic entity)	Attorney of Record		
Applicant Name (if Applicant is a juristic entity)	Sanas.ai Inc.		

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. If more than one applicant, use multiple forms.

☑ *Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 130085

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

Sanas.ai Inc.

☐ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☑ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature		Date (Optional)	05.12.2021
Name	Maxim Serebryakov		
Title	Chief Executive Officer		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☑ Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of: Sanas.ai Inc. Serial No.: TBD Filing Date: August 26, 2021 Title: Real-Time Accent Conversion Model	Confirmation No.: TBD Examiner: TBD Group Art Unit: TBD

Commissioner for Patents
P.O. Box 1450
Alexandria VA 22313-1450

GENERAL AUTHORIZATIONS UNDER 37 C.F.R. §§ 1.25(b) and 1.136(a)(3)

In this or any related application filed pursuant to 37 C.F.R. § 1.53, for the entire pendency thereof, and with respect to Deposit Account No. 506632, the Commissioner is hereby generally authorized:

(a) under 37 C.F.R. § 1.25(b), subject to the provisions of 37 C.F.R. § 1.311(b), to charge all fees set forth in 37 C.F.R. §§ 1.16 to 1.18; and

(b) under 37 C.F.R. § 1.136(a)(3) to treat any future reply requiring an extension of time as incorporating a request therefor, and specifically to charge any fee that may be due in connection with such a request.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: August 27, 2021

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Registration No. 67,894
(312) 754-9315
kennedy@ls3ip.com

Lee Sullivan Shea & Smith LLP
656 W. Randolph St., Floor 5W
Chicago, IL 60661
(312)754-9602

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	Real-Time Accent Conversion Model

As the below named inventor, I hereby declare that:

This declaration is directed to:

- [■] The attached application, or
- [ ] United States application or PCT international application number _____ filed on _____.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Shawn Zhang Date (Optional) :_____

Signature: _____

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

PTO/AIA/01 (06-12)
Approved for use through 11/30/2020. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DECLARATION (37 CFR 1.63) FOR UTILITY OR DESIGN APPLICATION USING AN APPLICATION DATA SHEET (37 CFR 1.76)

Title of Invention	Real-Time Accent Conversion Model

As the below named inventor, I hereby declare that:

This declaration is directed to:

- [■] The attached application, or
- [ ] United States application or PCT international application number _____ filed on _____.

The above-identified application was made or authorized to be made by me.

I believe that I am the original inventor or an original joint inventor of a claimed invention in the application.

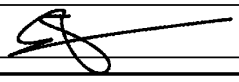
I hereby acknowledge that any willful false statement made in this declaration is punishable under 18 U.S.C. 1001 by fine or imprisonment of not more than five (5) years, or both.

WARNING:

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to the USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

LEGAL NAME OF INVENTOR

Inventor: Maxim Serebryakov Date (Optional): 08/27/21

Signature: Maxim Serebryakov, 

Note: An application data sheet (PTO/SB/14 or equivalent), including naming the entire inventive entity, must accompany this form or must have been previously filed. Use an additional PTO/AIA/01 form for each additional inventor.

This collection of information is required by 35 U.S.C. 115 and 37 CFR 1.63. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 1 minute to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

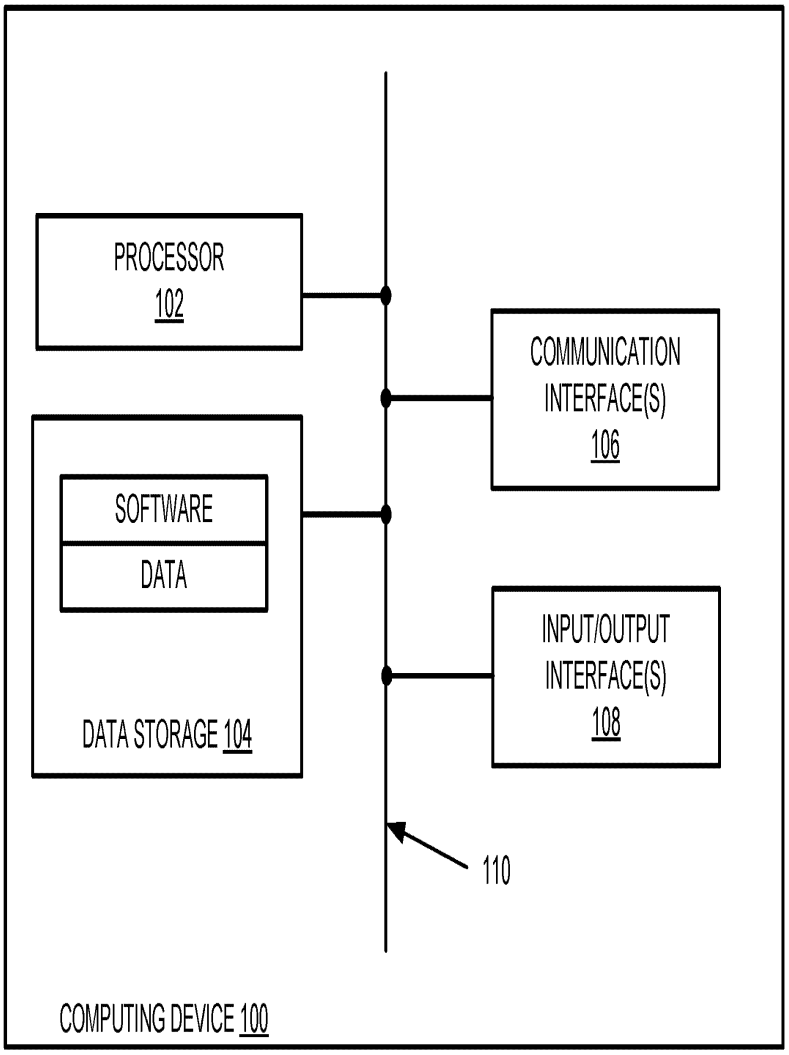

PROCESSOR
102
COMMUNICATION
INTERFACE(S)
106
SOFTWARE
DATA
INPUT/OUTPUT
INTERFACE(S)
108
DATA STORAGE 104
110
COMPUTING DEVICE 100

FIG. 1

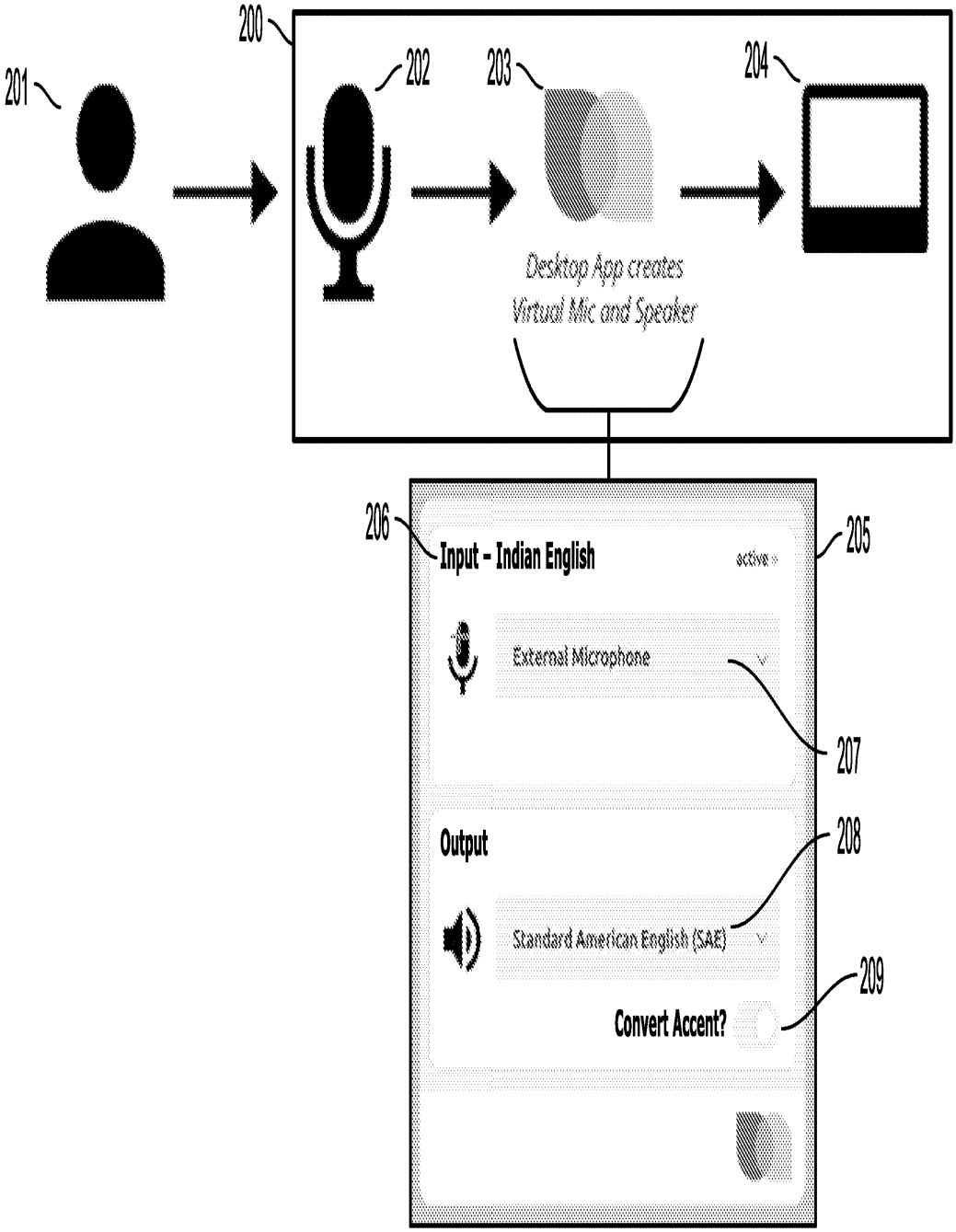

200
201
202
203
204
Desktop App creates
Virtual Mic and Speaker
206
205
Input - Indian English
active
External Microphone
207
Output
208
Standard American English (SAE)
209
Convert Accent?

FIG. 2

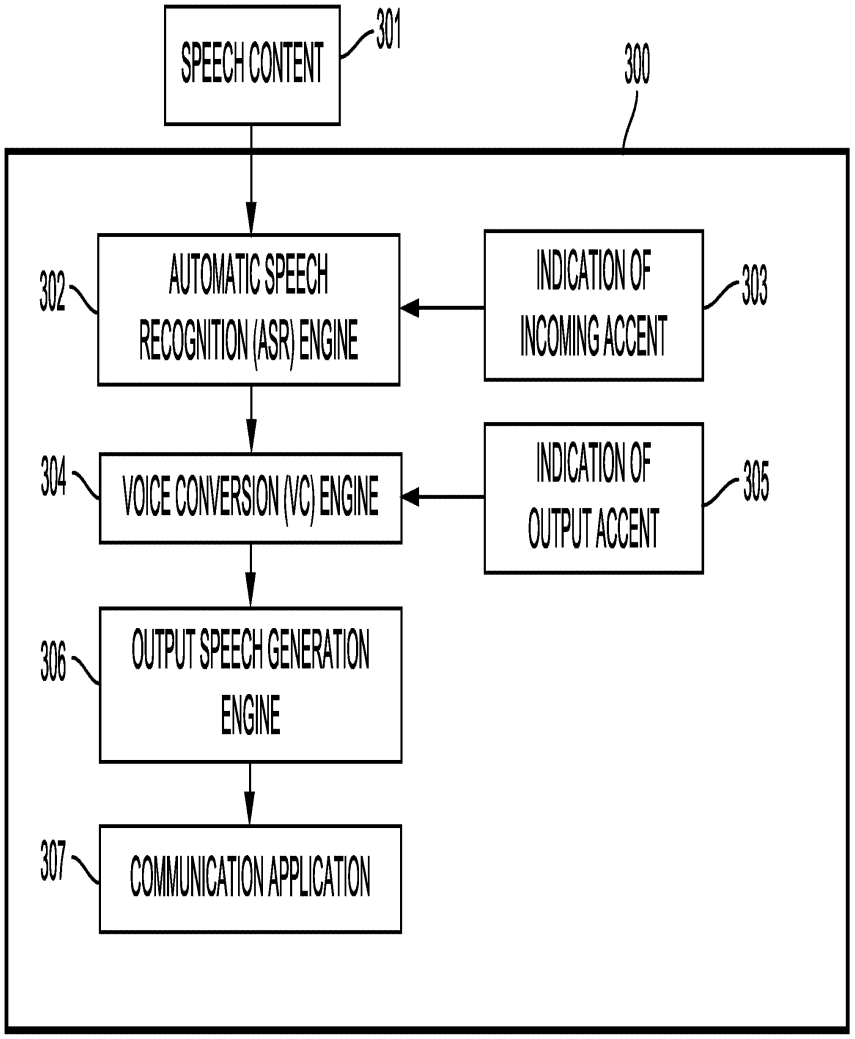

SPEECH CONTENT
301
300
AUTOMATIC SPEECH RECOGNITION (ASR) ENGINE
302
INDICATION OF INCOMING ACCENT
303
VOICE CONVERSION (VC) ENGINE
304
INDICATION OF OUTPUT ACCENT
305
OUTPUT SPEECH GENERATION ENGINE
306
COMMUNICATION APPLICATION
307

FIG. 3

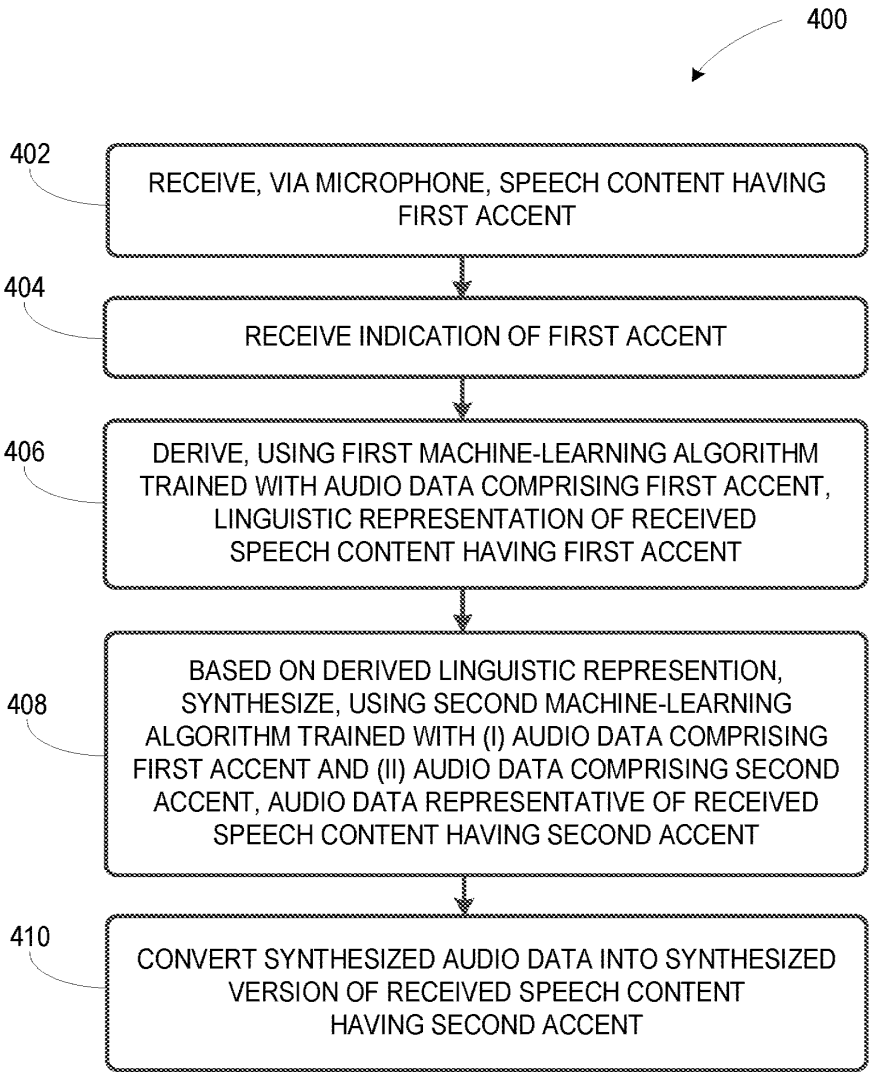

400
402
RECEIVE, VIA MICROPHONE, SPEECH CONTENT HAVING FIRST ACCENT
404
RECEIVE INDICATION OF FIRST ACCENT
406
DERIVE, USING FIRST MACHINE-LEARNING ALGORITHM TRAINED WITH AUDIO DATA COMPRISING FIRST ACCENT, LINGUISTIC REPRESENTATION OF RECEIVED SPEECH CONTENT HAVING FIRST ACCENT
408
BASED ON DERIVED LINGUISTIC REPRESENTION, SYNTHESIZE, USING SECOND MACHINE-LEARNING ALGORITHM TRAINED WITH (I) AUDIO DATA COMPRISING FIRST ACCENT AND (II) AUDIO DATA COMPRISING SECOND ACCENT, AUDIO DATA REPRESENTATIVE OF RECEIVED SPEECH CONTENT HAVING SECOND ACCENT
410
CONVERT SYNTHESIZED AUDIO DATA INTO SYNTHESIZED VERSION OF RECEIVED SPEECH CONTENT HAVING SECOND ACCENT

FIG. 4

Application for a United States Patent

United States Patent and Trademark Office

Attorney Docket No.: Sanas-00003

Title: Real-Time Accent Conversion Model

Applicant: Sanas.ai Inc.

IPR2026-00272
Krisp Technologies EX1002 Page 24

CROSS-REFERENCE TO RELATED APPLICATIONS

[1] This application claims the benefit of priority under 35 U.S.C. § 119 to U.S. Provisional Patent App. No. 63/185,345, filed on May 6, 2021, which is incorporated herein by reference in its entirety.

BACKGROUND

[2] Software applications are used on a regular basis to facilitate communication between users. As some examples, software applications can facilitate text-based communications such as email and other chatting/messaging platforms. Software applications can also facilitate audio and/or video-based communication platforms. Many other types of software applications for facilitating communications between users exist.

[3] Software applications are increasingly being relied on for communications in both personal and professional capacities. It is therefore desirable for software applications to provide sophisticated features and tools which can enhance a user's ability to communicate with others and thereby improve the overall user experience. Thus, any tool that can improve a user's ability to communicate with others is desirable.

IPR2026-00272
Krisp Technologies EX1002 Page 25

OVERVIEW

[4] One of the oldest communication challenges faced by people around the world is the barrier presented by different languages. Further, even among speakers of the same language, accents can sometimes present a communication barrier that is nearly as difficult to overcome as if the speakers were speaking different languages. For instance, a person who speaks English with a German accent may have difficulty understanding a person who speaks English with a Scottish accent.

[5] Today, there are relatively few software-based solutions that attempt to address the problem of accent conversion between speakers of the same language. One type of approach that has been proposed involves using voice conversion methods that attempt to adjust the audio characteristics (e.g., pitch, intonation, melody, stress) of a first speaker's voice to more closely resemble the audio characteristics of a second speaker's voice. However, this type of approach does not account for the different pronunciations of certain sounds that are inherent to a given accent, and therefore these aspects of the accent remain in the output speech. For example, many accents of the English language, such as Indian English and Irish English do not pronounce the phoneme for the digraph "th" found in Standard American English (SAE), instead replacing it with a "d" or "t" sound (sometimes referred to as th-stopping). Accordingly, a voice conversion model that only adjusts the audio characteristics of input speech does not address these types of differences.

[6] Some other approaches have involved a speech-to-text (STT) conversion of input speech as a midpoint, followed by a text-to-speech (TTS) conversion to generate the output audio content. However, this type of STT-TTS approach cannot capture many of the nuances of input speech that can provide information beyond the meaning of the words themselves, such as the prosody or emotion of the speaker. Further, a STT-TTS approach generally involves a degree of latency (e.g.,

IPR2026-00272
Krisp Technologies EX1002 Page 26

up to several seconds) that makes it impractical for use in real-time communication scenarios such as an ongoing conversation (e.g., a phone call).

[7] To address these and other problems with existing solutions for performing accent conversion, disclosed herein is new software technology that utilizes machine-learning models to receive input speech in a first accent and then output a synthesized version of the input speech in a second accent, all with very low latency (e.g., 300 milliseconds or less). In this way, accent conversion may be performed by a computing device in real time, allowing two users to verbally communicate more effectively in situations where their different accents would have otherwise made such communication difficult.

[8] Accordingly, in one aspect, disclosed herein is a method that involves a computing device (i) receiving an indication of a first accent, (ii) receiving, via at least one microphone, speech content having the first accent, (iii) receiving an indication of a second accent, (iv) deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent, (v) based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (a) audio data comprising the first accent and (b) audio data comprising the second accent, audio data representative of the received speech content having the second accent, and (vi) converting the synthesized audio data into a synthesized version of the received speech content having the second accent.

[9] In another aspect, disclosed herein is a computing device that includes at least one processor, a communication interface, a non-transitory computer-readable medium, and program instructions stored on the non-transitory computer-readable medium that are executable by the at

IPR2026-00272
Krisp Technologies EX1002 Page 27

least one processor to cause the computing device to carry out the functions disclosed herein, including but not limited to the functions of the foregoing method.

[10] In yet another aspect, disclosed herein is a non-transitory computer-readable storage medium provisioned with software that is executable to cause a computing device to carry out the functions disclosed herein, including but not limited to the functions of the foregoing method.

[11] One of ordinary skill in the art will appreciate these as well as numerous other aspects in reading the following disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[12] **FIG. 1** depicts an example computing device that may be configured to carry out one or more functions of a real-time accent conversion model.

[13] **FIG. 2** depicts a simplified block diagram of a computing device configured for real-time accent conversion.

[14] **FIG. 3** depicts a simplified block diagram of a computing device and an example data flow pipeline for a real-time accent conversion model.

[15] **FIG. 4** depicts an example flow chart that may be carried out to facilitate using a real-time accent conversion model.

IPR2026-00272
Krisp Technologies EX1002 Page 28

DETAILED DESCRIPTION

[16] The following disclosure refers to the accompanying figures and several example embodiments. One of ordinary skill in the art should understand that such references are for the purpose of explanation only and are therefore not meant to be limiting. Part or all of the disclosed systems, devices, and methods may be rearranged, combined, added to, and/or removed in a variety of manners, each of which is contemplated herein.

I. EXAMPLE COMPUTING DEVICE

[17] FIG. 1 is a simplified block diagram illustrating some structural components that may be included in an example computing device 100, on which the software technology discussed herein may be implemented. As shown in FIG. 1, the computing device may include one or more processors 102, data storage 104, a communication interface 106, one or more input/output (I/O) interfaces 108, all of which may be communicatively linked by a communication link 110 that may take the form of a system bus, among other possibilities.

[18] The processor 102 may comprise one or more processor components, such as general-purpose processors (e.g., a single- or multi-core microprocessor), special-purpose processors (e.g., an application-specific integrated circuit or digital-signal processor), programmable logic devices (e.g., a field programmable gate array), controllers (e.g., microcontrollers), and/or any other processor components now known or later developed. In line with the discussion above, it should also be understood that processor 102 could comprise processing components that are distributed across a plurality of physical computing devices connected via a network, such as a computing cluster of a public, private, or hybrid cloud.

[19] In turn, data storage 104 may comprise one or more non-transitory computer-readable storage mediums that are collectively configured to store (i) software components including program instructions that are executable by processor 102 such that computing device 100 is

IPR2026-00272
Krisp Technologies EX1002 Page 29

configured to perform some or all of the disclosed functions and (ii) data that may be received, derived, or otherwise stored, for example, in one or more databases, file systems, or the like, by computing device 100 in connection with the disclosed functions. In this respect, the one or more non-transitory computer-readable storage mediums of data storage 104 may take various forms, examples of which may include volatile storage mediums such as random-access memory, registers, cache, etc. and non-volatile storage mediums such as read-only memory, a hard-disk drive, a solid-state drive, flash memory, an optical-storage device, etc. In line with the discussion above, it should also be understood that data storage 104 may comprise computer-readable storage mediums that are distributed across a plurality of physical computing devices connected via a network, such as a storage cluster of a public, private, or hybrid cloud. Data storage 104 may take other forms and/or store data in other manners as well.

[20] The communication interface 106 may be configured to facilitate wireless and/or wired communication between the computing device 100 and other systems or devices. As such, communication interface 106 may communicate according to any of various communication protocols, examples of which may include Ethernet, Wi-Fi, Controller Area Network (CAN) bus, serial bus (e.g., Universal Serial Bus (USB) or Firewire), cellular network, and/or short-range wireless protocols, among other possibilities. In some embodiments, the communication interface 106 may include multiple communication interfaces of different types. Other configurations are possible as well.

[21] The I/O interfaces 108 of computing device 100 may be configured to (i) receive or capture information at computing device 100 and/or (ii) output information for presentation to a user. In this respect, the one or more I/O interfaces 108 may include or provide connectivity to input components such as a microphone, a camera, a keyboard, a mouse, a trackpad, a touchscreen, or a

IPR2026-00272
Krisp Technologies EX1002 Page 30

stylus, among other possibilities. Similarly, the I/O interfaces 108 may include or provide connectivity to output components such as a display screen and an audio speaker, among other possibilities.

[22] It should be understood that computing device 100 is one example of a computing device that may be used with the embodiments described herein, and may be representative of the computing devices 200 and 300 shown in FIGs. 2-3 and discussed in the examples below. Numerous other arrangements are also possible and contemplated herein. For instance, other example computing devices may include additional components not pictured or include less than all of the pictured components.

II. EXAMPLE FUNCTIONALITY

[23] Turning to FIG. 2, a simplified block diagram of a computing device configured for real-time accent conversion is shown. As described above, the disclosed technology is generally directed to a new software application that utilizes machine-learning models to perform real-time accent conversion on input speech that is received by a computing device, such as the computing device 200 shown in FIG. 2. In this regard, the accent-conversion application may be utilized in conjunction with one or more other software applications that are normally used for digital communications.

[24] For example, as shown in FIG. 2, a user 201 of the computing device 200 may provide speech content that is captured by a hardware microphone 202 of the computing device 200. In some embodiments, the hardware microphone 202 shown in FIG. 2 might be an integrated component of the computing device 200 (e.g., the onboard microphone of a laptop computer or smartphone). In other embodiments, the hardware microphone 202 might take the form of a wired

IPR2026-00272
Krisp Technologies EX1002 Page 31

or wireless peripheral device (e.g., a webcam, a dedicated hardware microphone) that is connected to an I/O interface of the computing device 200. Other examples are also possible.

[25] The speech content may then be passed to the accent-conversion application 203 shown in FIG. 2. In some implementations, the accent-conversion application 203 may function as a virtual microphone that receives the captured speech content from the hardware microphone 202 of the computing device 200, performs accent conversion as discussed herein, and then routes the converted speech content to a digital communication application 204 (e.g., Zoom, Skype, Viber, Telegram, etc.) that would normally receive input speech content directly from the hardware microphone 202. Advantageously, the accent conversion may be accomplished locally on the computing device 200, which may tend to minimize the latency associated with other applications that may rely on cloud-based computing.

[26] FIG. 2 shows one possible example of a virtual microphone interface 205 that may be presented by the accent-conversion application 203. For example, the virtual microphone interface 205 may provide an indication 206 of the input accent of the user 201, which may be established by the user 201 upon initial installation of the accent-conversion application 203 on computing device 200. As shown in FIG. 2, the virtual microphone interface 205 indicates that the user 201 speaks with an Indian English accent. In some implementations, the input accent may be adjustable to accommodate users with different accents than the user 201.

[27] Further, the virtual microphone interface 205 may include a drop-down menu 207 or similar option for selecting the input source from which the accent-conversion application 203 will receive speech content, as the computing device 200 might have multiple available options to use as an input source. Still further, the virtual microphone interface 205 may include a drop-down menu 208 or similar option for selecting the desired output accent for the speech content. As

IPR2026-00272
Krisp Technologies EX1002 Page 32

shown in FIG. 2, the virtual microphone interface 205 indicates that the incoming speech content will be converted to speech having a SAE accent. The converted speech content is then provided to the communication application 204, which may process the converted speech content as if it had come from the hardware microphone 202.

[28] Still further, in some implementations, the virtual microphone interface 205 may include a toggle 209 or similar control that may be used to turn the accent conversion functionality of the accent-conversion application 203 on or off. When the conversion functionality is toggled off, the accent-conversion application 203 may act as a pass-through for the input speech. In this way, the user 201 may avoid the hassle of reconfiguring input devices to remove the virtual microphone for conversations where accent conversion is not needed, allowing the user 201 to easily move between conversations with and without accent conversion engaged.

[29] Advantageously, the accent-conversion application 203 may accomplish the operations above, and discussed in further detail below, at speeds that enable real-time communications, having a latency as low as 50-700 ms (e.g., 200 ms) from the time the input speech received by the accent-conversion application 203 to the time the converted speech content is provided to the communication application 204. Further, the accent-conversion application 203 may process incoming speech content as it is received, making it capable of handling both extended periods of speech as well as frequent stops and starts that may be associated with some conversations. For example, in some embodiments, the accent-conversion application 203 may process incoming speech content every 160 ms. In other embodiments, the accent-conversion application 203 may process the incoming speech content more frequently (e.g., every 80 ms) or less frequently (e.g., every 300 ms).

IPR2026-00272
Krisp Technologies EX1002 Page 33

[30] Turning now to FIG. 3, a simplified block diagram of a computing device 300 and an example data flow pipeline for a real-time accent conversion model are shown. For instance, the computing device 300 may be similar to or the same as the computing device 200 shown in FIG. 2. At a high-level, the components of the real-time accent conversion model that operate on the incoming speech content 301 include (i) an automatic speech recognition (ASR) engine 302, (ii) a voice conversion (VC) engine 304, and (iii) an output speech generation engine 306. As one example, the output speech generation engine may be embodied in a vocoder 306.

[31] FIG. 3 will be discussed in conjunction with FIG. 4, which depicts a flow chart 400 that includes example operations that may be carried out by a computing device, such as the computing device 300 of FIG. 3, to facilitate using a real-time accent conversion model.

[32] At block 402, the computing device 300 may receive speech content 301 having a first accent. For instance, as discussed above with respect to FIG. 2, a user such as user 201 may provide speech content 301 having an Indian English accent, which may be captured by a hardware microphone of the computing device 300. In some implementations, the computing device 300 may engage in pre-processing of the speech content 301, including converting the speech content 301 from an analog signal to a digital signal using an analog-to-digital converter (not shown), and/or down-sampling the speech content 301 to a sample rate (e.g., 16 kHz) that will be used by the ASR engine 302, among other possibilities. In other implementations, one or more of these pre-processing actions may be performed by the ASR engine 302.

[33] The ASR engine 302 includes one or more machine learning models (e.g., a neural network, such as a recurrent neural network (RNN), a transformer neural network, etc.) that may be trained using previously captured speech content from many different speakers having the first accent. Continuing the example above, the ASR engine 302 may be trained with previously

IPR2026-00272
Krisp Technologies EX1002 Page 34

captured speech content from a multitude of different speakers, each having an Indian English accent. For instance, the captured speech content used as training data may include transcribed content in which each of the speakers read the same script (e.g., a script curated to provide a wide sampling of speech sounds, as well as specific sounds that are unique to the first accent). Thus, the ASR engine 302 may align and classify each frame of the captured speech content according to its monophone and triphone sounds, as indicated in the corresponding transcript. As a result of this frame-wise breakdown of the captured speech across multiple speakers having the first accent, the ASR engine 302 may develop a learned linguistic representation of speech having an Indian English accent that is not speaker-specific.

[34] On the other hand, the ASR engine 302 may also be used to develop a learned linguistic representation for an output accent that is only based on speech content from a single, representative speaker (e.g., a target SAE speaker) reading a script in the output accent, and therefore is speaker specific. In this way, the synthesized speech content that is generated having the target accent (discussed further below) will tend to sound like the target speaker for the output accent. In some cases, this may simplify the processing required to perform accent conversion and generally reduce latency.

[35] Further, it should be understood that the captured speech content that is used to train the ASR engine 302 does not necessarily need to be limited to captured speech content having the first accent. Rather, the ASR engine 302 discussed herein may be trained using captured speech content having a diversity of accents, which may enable the ASR engine 302 to develop a learned linguistic representation of not only the first accent, but also the second accent as well as additional accents. In this way, the accent-conversion application 203 noted above may utilize a single ASR engine

IPR2026-00272
Krisp Technologies EX1002 Page 35

302 that is capable of receiving and converting speech content having various different input accents.

[36] In some implementations, the speech content collected from the multiple Indian English speakers as well as the target SAE speaker for training the ASR engine 302 may be based on the same script, also known as parallel speech. In this way the transcripts used by the ASR engine 302 to develop a linguistic representation for speech content in both accents are the same, which may facilitate mapping one linguistic representation to the other in some situations. Alternatively, the training data may include non-parallel speech, which may require less training data. Other implementations are also possible, including hybrid parallel and non-parallel approaches.

[37] It should be noted that the learned linguistic representations developed by the ASR engine 302 and discussed herein may not be recognizable as such to a human. Rather, the learned linguistic representations may be encoded as machine-readable data (e.g., a hidden representation) that the ASR engine 302 uses to represent linguistic information.

[38] In practice, the ASR engine 302 may be individually trained with speech content including multiple different accents, across different languages, and may develop a learned linguistic representation for each one. Accordingly, at block 404, the computing device 300 may receive an indication of the Indian English accent associated with the received speech content 301, so that the appropriate linguistic representation is used by the ASR engine 302. As noted above, this indication of the incoming accent, shown by way of example as block 303 in FIG. 3, may be established at the time the accent-conversion application is installed on the computing device 300 and might not be changed thereafter. As another possibility, the accent-conversion application may be adjusted to indicate a different incoming accent, such that the ASR engine 302 uses a different learned linguistic representation to analyze the incoming speech content 301.

IPR2026-00272
Krisp Technologies EX1002 Page 36

[39] At block 406, the ASR engine 302 may derive a linguistic representation of the received speech content 301, based on the learned linguistic representation the ASR engine 302 has developed for the Indian English accent. For instance, the ASR engine 302 may break down the received speech content 301 by frame and classify each frame according to the sounds (e.g., monophones and triphones) that are detected, and according to how those particular sounds are represented and inter-related in the learned linguistic representation associated with an Indian English accent.

[40] In this way, the ASR engine 302 functions to deconstruct the received speech content 301 having the first accent into a derived linguistic representation with very low latency. In this regard, it should be noted that the ASR engine 302 may differ from some other speech recognition models that are configured predict and generate output speech, such as a speech-to-text model. Accordingly, the ASR engine 302 may not need to include such functionality.

[41] The derived linguistic representation of the received speech content 301 may then be passed to the VC engine 304. Similar to the indication of the incoming accent 303, the computing device 300 may also receive an indication of the output accent, shown by way of example as block 305 in FIG. 3, so that the VC engine 304 can apply the appropriate mapping and conversion from the incoming accent to the output accent. For instance, the indication of the output accent may be received based on a user selection from a menu, such as the virtual microphone interface 205 shown in FIG. 2, prior to receiving the speech content 301 having the first accent.

[42] Similar to the ASR engine 302, the VC engine 304 includes one or more machine learning models (e.g., a neural network) that use the learned linguistic representations developed by the ASR engine 302 as training inputs to learn how to map speech content from one accent to another. For instance, the VC engine 304 may be trained to map an ASR-based linguistic representation of

IPR2026-00272
Krisp Technologies EX1002 Page 37

Indian English speech to an ASR-based linguistic representation of a target SAE speaker. In training the VC engine 304, it is necessary to align the Indian English speech to the SAE speech during this mapping. One possible way to accomplish this is by using individual monophones and/or triphones within the training data as a possible heuristic to better determine the alignments. Like the learned linguistic representations themselves, the learned mapping between the two representations may be encoded as machine-readable data (e.g., a hidden representation) that the VC engine 304 uses to represent linguistic information.

[43] Accordingly, at block 408, the VC engine 304 may utilize the learned mapping between the two linguistic representations to synthesize, based on the derived linguistic representation of the received speech content 301, audio data that is representative of the speech content 301 having the second accent. The audio data that is synthesized in this way may take the form of a set of mel spectrograms. For example, the VC engine 304 may map each incoming frame in the derived linguistic representation to an outgoing target speech frame.

[44] In this way, the VC engine 304 functions to reconstruct acoustic features from the derived linguistic representation into audio data that is representative of speech by a different speaker having the second accent, all with very low latency. Advantageously, because the VC engine 304 works at the level of encoded linguistic data and does not need to predict and generate output speech as a midpoint for the conversion, it can function more quickly than alternatives such as a STT-TTS approach. Further, the VC engine 304 may more accurately capture some of the nuances of voice communications, such as brief pauses or changes in pitch, prosody, or the emotion of the speaker, all of which can convey important information and which may be lost if the speech content were converted to text first and then back to speech.

IPR2026-00272
Krisp Technologies EX1002 Page 38

[45] At block 410, the output speech generation engine 306 may convert the synthesized audio data into output speech, which may be a synthesized version of the received speech content 301 having the second accent. As noted above, the output speech may further have the voice identity of the target speaker whose speech content, having the second accent, was used to train the ASR engine 302. In some examples, the output speech generation engine 306 may take the form of a vocoder 306 or similar component that can rapidly process audio under the real-time conditions contemplated herein. The output speech generation engine 306 may include one or more additional machine learning algorithms (e.g., a neural network, such as a generative adversarial network, one or more Griffin-Lim algorithms, etc.) that learn to convert the synthesized audio data into waveforms that are able to be heard. Other examples are also possible.

[46] As shown in FIG. 3, the output speech generation engine 306 may pass the output speech to a communication application 307 operating on the computing device 300. The communication application 307 may then transmit the output speech to one or more other computing devices, cause the computing device 300 to play back the output speech via one or more speakers, and/or store the output speech as an audio data file, among numerous other possibilities.

[47] Although the examples discussed above involve a computing device 300 that utilizes the accent-conversation application for outgoing speech (e.g., situations where the user of computing device 300 is the speaker), it is also contemplated that the accent-conversion application may be used by the computing device 300 in the opposite direction as well, for incoming speech content where the user is a listener. For instance, rather than being situated as a virtual microphone between a hardware microphone and the communication application 307, the accent-conversion application may be deployed as a virtual speaker between the communication application 307 and a hardware speaker of the computing device 300, and the indication of the incoming accent 303

IPR2026-00272
Krisp Technologies EX1002 Page 39

and the indication of the output accent 305 shown in FIG. 3 may be swapped. In some cases, these two pipelines may run in parallel such that a single installation of the accent-conversion application is performing two-way accent conversion between users. In the context of the example discussed above, this arrangement may allow the Indian English speaker, whose outgoing speech is being converted to an SAE accent, to also hear the SAE speaker's responses in Indian English accented speech (e.g., synthesized speech of a target Indian English speaker).

[48] As a further extension, the examples discussed above involve an ASR engine 302 that is provided with an indication of the incoming accent. However, in some embodiments it may be possible to use the accent-conversion application discussed above in conjunction with an accent detection model, such that the computing device 300 is initially unaware of one or both accents that may be present in a given communication. For example, an accent detection model may be used in the initial moments of a conversation to identify the accents of the speakers. Based on the identified accents, the accent-conversion application may determine the appropriate learned linguistic representation(s) that should be used by the ASR engine 302 and the corresponding learned mapping between representations that should be used by the VC engine 304. Additionally, or alternatively, the accent detection model may be used to provide a suggestion to a user for which input/output accent the user should select to obtain the best results. Other implementations incorporating an accent detection model are also possible.

[49] FIG. 4 includes one or more operations, functions, or actions as illustrated by one or more of blocks 402-410, respectively. Although the blocks are illustrated in sequential order, some of the blocks may also be performed in parallel, and/or in a different order than those described herein. Also, the various blocks may be combined into fewer blocks, divided into additional blocks, and/or removed based upon the desired implementation.

IPR2026-00272
Krisp Technologies EX1002 Page 40

[50] In addition, for the example flow chart in FIG. 4 and other processes and methods disclosed herein, the flow chart shows functionality and operation of one possible implementation of present embodiments. In this regard, each block may represent a module, a segment, or a portion of program code, which includes one or more instructions executable by one or more processors for implementing logical functions or blocks in the process.

[51] The program code may be stored on any type of computer readable medium, for example, such as a storage device including a disk or hard drive. The computer readable medium may include non-transitory computer readable medium, for example, such as computer-readable media that stores data for short periods of time like register memory, processor cache and Random-Access Memory (RAM). The computer readable medium may also include non-transitory media, such as secondary or persistent long-term storage, like read only memory (ROM), optical or magnetic disks, compact disc read only memory (CD-ROM), for example. The computer readable media may also be any other volatile or non-volatile storage systems. The computer readable medium may be considered a computer readable storage medium, for example, or a tangible storage device. In addition, for the processes and methods disclosed herein, each block in FIG. 4 may represent circuitry and/or machinery that is wired or arranged to perform the specific functions in the process.

III. CONCLUSION

[52] Example embodiments of the disclosed innovations have been described above. Those skilled in the art will understand, however, that changes and modifications may be made to the embodiments described without departing from the true scope and spirit of the present invention, which will be defined by the claims.

IPR2026-00272
Krisp Technologies EX1002 Page 41

[53] Further, to the extent that examples described herein involve operations performed or initiated by actors, such as "humans," "operators," "users," or other entities, this is for purposes of example and explanation only. Claims should not be construed as requiring action by such actors unless explicitly recited in claim language.

IPR2026-00272
Krisp Technologies EX1002 Page 42

CLAIMS

We Claim:

1. A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 43

2. The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

3. The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

4. The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication

IPR2026-00272
Krisp Technologies EX1002 Page 44

of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent.

7. The computing device of claim 1, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

8. The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

IPR2026-00272
Krisp Technologies EX1002 Page 45

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

10. A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent;

IPR2026-00272
Krisp Technologies EX1002 Page 46

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent.

11. The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

12. The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 47

13. The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

14. The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

15. The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent.

16. The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

IPR2026-00272
Krisp Technologies EX1002 Page 48

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

17. The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. A method carried out by a computing device, the method comprising:

IPR2026-00272
Krisp Technologies EX1002 Page 49

receiving an indication of a first accent;

receiving an indication of a second accent;

receiving, via at least one microphone, speech content having the first accent;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent.

20. The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 50

ABSTRACT

Techniques for real-time accent conversion are described herein. An example computing device receives an indication of a first accent and a second accent. The computing device further receives, via at least one microphone, speech content having the first accent. The computing device is configured to derive, using a first machine-learning algorithm trained with audio data including the first accent, a linguistic representation of the received speech content having the first accent. The computing device is configured to, based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data including the second accent, audio data representative of the received speech content having the second accent. The computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 51

Electronic Patent Application Fee Transmittal	
Application Number:	
Filing Date:	
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Brandon J. Kennedy
Filer:	Brandon Jacob Kennedy
Attorney Docket Number:	Sanas-00003
Filed as Small Entity	

Filing Fees for Track I Prioritized Examination - Nonprovisional Application under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
UTILITY FILING FEE (ELECTRONIC FILING)	4011	1	80	80
UTILITY SEARCH FEE	2111	1	350	350
UTILITY EXAMINATION FEE	2311	1	400	400
REQUEST FOR PRIORITIZED EXAMINATION	2817	1	2100	2100
Pages:				
Claims:				
Miscellaneous-Filing:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
PUBL. FEE- EARLY, VOLUNTARY, OR NORMAL	1504	1	0	0
PROCESSING FEE, EXCEPT PROV. APPLS.	2830	1	70	70
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
		Total in USD ($)		**3000**

Electronic Acknowledgement Receipt	
EFS ID:	43628518
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Brandon J. Kennedy
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	27-AUG-2021
Filing Date:	
Time Stamp:	23:03:40
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$3000
RAM confirmation Number	E20218QN04084885
Deposit Account	506632
Authorized User	Brandon Kennedy
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: 37 CFR 1.16 (National application filing, search, and examination fees) 37 CFR 1.17 (Patent application and reexamination processing fees)	

37 CFR 1.19 (Document supply fees)

37 CFR 1.20 (Post Issuance fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	TrackOne Request	Sanas-00003_TrackOneRequest.pdf	114630 fe8d98c0c2032bc82904dc34114dad34196858a9	no	2
Warnings:					
Information:					
2	Application Data Sheet	Sanas-00003_ADS.pdf	1255911 f574efbb5d8aa6559be0e84c4a3234b363325eac	no	8
Warnings:					
Information:					
3	Power of Attorney	Sanas-00003_POA.pdf	230765 e4674e9d363eee9ee414ba80c3e4d6a5d1f782f4	no	3
Warnings:					
Information:					
4	Authorization for Extension of Time all replies	Sanas-00003_GeneralAuthorization.pdf	84000 9dab60cb7160b6b5f8994994a8d0139996059b5e	no	1
Warnings:					
Information:					
5	Oath or Declaration filed	Sanas-00003_Declaration-Zhang-Executed.pdf	132427 0cc3e7f7c1736d4d23dbf23c8381358cf2e5a231	no	2
Warnings:					
Information:					
6	Oath or Declaration filed	Sanas-00003_Declaration-Serebryakov-Executed.pdf	195342 b67ab1d9d48ab72db68802ec4380bfc17a1434c1	no	2
Warnings:					
Information:					

7	Drawings-only black and white line drawings	Sanas-00003_Figures.pdf	127538 85d8fc3d910327f9b8a6e324bc442a6cf9e06bb5	no	4
Warnings:					
Information:					
8		Sanas-00003_Specification.pdf	173631 f0ae1039d327b2b3343d4274f982b503c5de79dc	yes	28
	Multipart Description/PDF files in .zip description				
	Document Description		**Start**	**End**	
	Specification		1	19	
	Claims		20	27	
	Abstract		28	28	
Warnings:					
Information:					
9	Fee Worksheet (SB06)	fee-info.pdf	50582 e221dfc372105d4846cd3c5ef474cea76850c931	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	2364826		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Doc code: IDS
Doc description: Information Disclosure Statement (IDS) Filed

PTO/SB/08a (02-18)
Approved for use through 11/30/2020. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (**Not for submission under 37 CFR 1.99)**		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	TBD
	Attorney Docket Number	Sanas-00003

U.S.PATENTS Remove

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	10163451	B2	2018-12-25	DIRAC et al.	
	2	10614826	B2	2020-04-07	HUFFMAN et al.	

If you wish to add additional U.S. Patent citation information please click the Add button. Add

U.S.PATENT APPLICATION PUBLICATIONS Remove

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button. Add

FOREIGN PATENT DOCUMENTS Remove

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2] i	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]
	1							

If you wish to add additional Foreign Patent Document citation information please click the Add button Add

NON-PATENT LITERATURE DOCUMENTS Remove

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (**Not for submission under 37 CFR 1.99)**		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	TBD
	Attorney Docket Number	Sanas-00003

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1		

If you wish to add additional non-patent literature document citation information please click the Add button Add

EXAMINER SIGNATURE

Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	TBD
	Attorney Docket Number	Sanas-00003

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

See attached certification statement.

The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Brandon J. Kennedy/	Date (YYYY-MM-DD)	2021-08-31
Name/Print	Brandon J. Kennedy	Registration Number	67894

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IPR2026-00272
Krisp Technologies EX1002 Page 60

Electronic Acknowledgement Receipt

EFS ID:	43650418
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	31-AUG-2021
Filing Date:	
Time Stamp:	15:45:39
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Information Disclosure Statement (IDS) Form (SB08)	Sanas-00003_IDS.pdf	1034215 7f1ee89a21f3dd6db6521615bd0f3de7550ffffb2	no	4

Warnings:

Information:	
Total Files Size (in bytes):	1034215

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875

Application or Docket Number: 17/460,145

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	80		N/A	
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A	350		N/A	
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	400		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	20 minus 20 =	*	x 50 =	0.00	OR		
INDEPENDENT CLAIMS (37 CFR 1.16(h))	3 minus 3 =	*	x 240 =	0.00			
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).			0.00			
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				0.00			
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	830		TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	SMALL ENTITY RATE($)	SMALL ENTITY ADDITIONAL FEE($)	OR	OTHER THAN SMALL ENTITY RATE($)	OTHER THAN SMALL ENTITY ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)	OR	RATE($)	ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
17/460,145	08/27/2021	2658	830	Sanas-00003	20	3

CONFIRMATION NO. 3314

130085
Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

FILING RECEIPT

OC000000128253416

Date Mailed: 09/10/2021

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Maxim Serebryakov, Palo Alto, CA;
Shawn Zhang, Pleasanton, CA;

Applicant(s)

Sanas.ai Inc., Pleasanton, CA;

Power of Attorney: The patent practitioners associated with Customer Number 130085

Domestic Priority data as claimed by applicant

This appln claims benefit of 63/185,345 05/06/2021

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.
Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

IPR2026-00272
Krisp Technologies EX1002 Page 64

If Required, Foreign Filing License Granted: 09/08/2021

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 17/460,145**

Projected Publication Date: 11/10/2022

Non-Publication Request: No

Early Publication Request: No
**** SMALL ENTITY ****
Title

Real-Time Accent Conversion Model

Preliminary Class

704

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

IPR2026-00272
Krisp Technologies EX1002 Page 65

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 10/01/2021

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
10/01/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 67

Decision Granting Request for Prioritized Examination (Track I)	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.	
	Examiner ANNA PAGONAKIS	**Art Unit** OPET	**AIA (FITF) Status** Yes

1. THE REQUEST FILED 27 August 2021 IS **GRANTED** .

The above-identified application has met the requirements for prioritized examination

A. ☑ for an original nonprovisional application (Track I).
B. ☐ for an application undergoing continued examination (RCE).

2. **The above-identified application will undergo prioritized examination**. The application will be accorded special status throughout its entire course of prosecution until one of the following occurs:

A. filing a **petition for extension of time** to extend the time period for filing a reply;

B. filing an **amendment to amend the application to contain more than four independent claims, more than thirty total claims** , or a multiple dependent claim;

C. filing a **request for continued examination** ;

D. filing a notice of appeal;

E. filing a request for suspension of action;

F. mailing of a notice of allowance;

G. mailing of a final Office action;

H. completion of examination as defined in 37 CFR 41.102; or

I. abandonment of the application.

Telephone inquiries with regard to this decision should be directed to ANNA PAGONAKIS at (571) 270-3505. In his/her absence, calls may be directed to Petition Help Desk at (571) 272-3282.

/ANNA PAGONAKIS/ Petitions Examiner, OPET	

IPR2026-00272
Krisp Technologies EX1002 Page 68

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 10/25/2021

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
10/25/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 69

Office Action Summary	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 08/27/2021.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-20 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-20 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☑ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☑ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Specification

2. The use of the terms "Zoom," "Skype," "Viber," and "Telegram," which are trade names or marks used in commerce, have been noted in this application. The terms should be accompanied by the generic terminology; furthermore the terms should be capitalized wherever it appears or, where appropriate, include a proper symbol indicating use in commerce such as ™, SM , or ® following the term.

Although the use of trade names and marks used in commerce (i.e., trademarks, service marks, certification marks, and collective marks) are permissible in patent applications, the proprietary nature of the marks should be respected and every effort made to prevent their use in any manner which might adversely affect their validity as commercial marks.

Claim Objections

3. Claims 6 and 15 objected to because of the following informalities: the last limitation reads "receive the indication of the first accent receive the indication of the second accent". Appropriate correction is required.

Claim Rejections - 35 USC § 102

4. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory

basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

5. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (a)(1) the claimed invention was patented, described in a printed publication, or in public use, on sale, or otherwise available to the public before the effective filing date of the claimed invention.

6. Claim(s) 1-2, 4, 6, 8, 10-11, 13, 15, 17, and 19 is/are rejected under 35 U.S.C. 102(a)(1) as being anticipated by Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac.

Regarding claim 1, Dirac teaches:

A computing device comprising:

at least one microphone (Fig. 6 element 621, col. 8 lines 16-32, where a microphone is used);

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used);

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used); and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) such that the computing device is configured to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via the at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 2, Dirac teaches:

The computing device of claim I, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second

accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac teaches:

The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac teaches:

The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent (col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent (col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 8, Dirac teaches:

The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used), wherein the non-transitory computer- readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received

speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 11, Dirac teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac teaches:

The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent (col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent (col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 17, Dirac teaches:

The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

receiving an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Claim Rejections - 35 USC § 103

7. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

8. Claims 3 and 12 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Zhao et al. (Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *INTERSPEECH* (pp. 2843-2847).), hereinafter referred to as Zhao.

Regarding claim 3, Dirac teaches:

The computing device of claim 1,

Dirac does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

Regarding claim 12, Dirac teaches:

The non-transitory computer-readable medium of claim 10,

Dirac does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

9. Claims 5 and 14 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac teaches:

The computing device of claim 1,

Dirac does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

Regarding claim 14, Dirac teaches:

The non-transitory computer-readable medium of claim 10,

Dirac does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

10. Claims 7 and 16 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Sajjan et al. (Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.), hereinafter referred to as Sajjan.

Regarding claim 7, Dirac teaches:

The computing device of claim 1, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone (col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent) sound detected in the frame based on the learned linguistic representation for the first accent.

Dirac does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

(ii) triphone (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used) sound detected in the frame

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

Regarding claim 16, Dirac teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a learned linguistic representation for the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone based on the learned linguistic representation for the first accent (col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Dirac does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

ii) triphone sound detected in the frame (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used)

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

11. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac teaches:

The computing device of claim 1,

Dirac does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by including the delay of Leff (Leff para [0093]) in the conversion of Dirac (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac teaches:

The non-transitory computer-readable medium of claim 10,

Dirac does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that,

when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by including the delay of Leff (Leff para [0093]) in the conversion of Dirac (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed

with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac teaches:

The method of claim 19,

Dirac does not teach:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by including the delay of Leff (Leff para [0093]) in the conversion of Dirac (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed

with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

12. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2021/0074264 A1 para [0048] teaches both monophone and triphone training.

13. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20210074264-A1	03-2021	Liang; Hao	G06N20/00	1/1
*	B	US-20160203827-A1	07-2016	LEFF; Julian	G06T13/40	704/207
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.
	V	Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In INTERSPEECH (pp. 2843-2847).
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021								
	1	✓								
	2	✓								
	3	✓								
	4	✓								
	5	✓								
	6	✓								
	7	✓								
	8	✓								
	9	✓								
	10	✓								
	11	✓								
	12	✓								
	13	✓								
	14	✓								
	15	✓								
	16	✓								
	17	✓								
	18	✓								
	19	✓								
	20	✓								

IPR2026-00272
Krisp Technologies EX1002 Page 91

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	**Date**	**Examiner**
G10L13/02	10/20/2021	BB
G10L15/02	10/20/2021	BB

CPC Combination Sets - Searched*		
Symbol	**Date**	**Examiner**

US Classification - Searched*			
Class	**Subclass**	**Date**	**Examiner**

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	**Date**	**Examiner**
Performed inventor search in PE2E, EAST	10/20/2021	BB
Performed text search	10/20/2021	BB
Performed text limited CPC search	10/20/2021	BB

Interference Search			
US Class/CPC Symbol	**US Subclass/CPC Group**	**Date**	**Examiner**

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 92

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	0	((serebryakov zhang).in. ("sanas.ai").asn.) and (accent and synthesi$4 and conver$4 and machine near3 learn$3).clm.	US-PGPUB; USPAT	OR	ON	2021/10/18 11:51
S2	2	("10163451" "10614826").pn.	US-PGPUB; USPAT	OR	ON	2021/10/18 11:51
S3	7	(voice accent) near5 conver$4 and frame same monophone same triphone	US-PGPUB; USPAT	OR	ON	2021/10/19 23:37
S4	0	frame same monophone same triphone same accent	US-PGPUB; USPAT	OR	ON	2021/10/19 23:40
S5	0	frame same (monophone mono-phone phone) same (triphone tri-phone) same accent	US-PGPUB; USPAT	OR	ON	2021/10/19 23:41
S6	10	frame same (monophone mono-phone phone phoneme) same (triphone tri-phone) same accent	US-PGPUB; USPAT	OR	ON	2021/10/19 23:41
S7	12,738	(voice accent) near5 conver$4 and (time latency delay lag) same (ms millisecond milli-second)	US-PGPUB; USPAT	OR	ON	2021/10/20 11:24
S8	229	(voice accent) near5 conver$4 and (time latency delay lag) same (ms millisecond milli-second) same synthesiz$4	US-PGPUB; USPAT	OR	ON	2021/10/20 11:24
S9	16	(voice accent) near5 conver$4 same (time latency delay lag) same (ms millisecond milli-second) same synthesiz$4	US-PGPUB; USPAT	OR	ON	2021/10/20 11:24
S10	8	((voice accent) near5 conver$4).ab. and real near2 time and (time latency delay lag) same (ms millisecond milli-second) same synthesiz$4	US-PGPUB; USPAT	OR	ON	2021/10/20 11:26
S11	48	((voice accent) near5 conver$4).ab. and real near2 time and (time latency delay lag) near9 synthesiz$4	US-PGPUB; USPAT	OR	ON	2021/10/20 11:27
S12	565	((voice accent) near5 conver$4).ab. and real near2 time and (time latency delay lag) same (second millisecond ms)	US-PGPUB; USPAT	OR	ON	2021/10/20 11:34

S13	231	((voice accent) near5 conver$4).ab. and real near2 time same (time latency delay lag) same (second millisecond ms)	US-PGPUB; USPAT	OR	ON	2021/10/20 11:35
S14	43	((voice accent) near5 conver$4).ab. and real near2 time same (time latency delay lag) same (milli-second millisecond ms)	US-PGPUB; USPAT	OR	ON	2021/10/20 11:36
S15	2	(US-20210074264-$ or US-20160203827-$).did.	US-PGPUB	OR	ON	2021/10/20 12:58
S16	5	S14 and (G10L13/02 G10L15/02).cpc.	US-PGPUB; USPAT	OR	ON	2021/10/20 13:05

EAST Search History (Interference)

<This search history is empty>

10/20/2021 1:10:20 PM
C:\Users\bblankenagel\Documents\EAST\Workspaces\17460145.wsp

Doc code: IDS
Doc description: Information Disclosure Statement (IDS) Filed

PTO/SB/08a (02-18)
Approved for use through 11/30/2020. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT
(Not for submission under 37 CFR 1.99)

Application Number	17460145
Filing Date	2021-08-27
First Named Inventor	Maxim Serebryakov
Art Unit	2658
Examiner Name	TBD
Attorney Docket Number	Sanas-00003

U.S.PATENTS Remove

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
/B.B./	1	10163451	B2	2018-12-25	DIRAC et al.	
/B.B./	2	10614826	B2	2020-04-07	HUFFMAN et al.	

If you wish to add additional U.S. Patent citation information please click the Add button. Add

U.S.PATENT APPLICATION PUBLICATIONS Remove

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button. Add

FOREIGN PATENT DOCUMENTS Remove

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2] i	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]
	1							

If you wish to add additional Foreign Patent Document citation information please click the Add button Add

NON-PATENT LITERATURE DOCUMENTS Remove

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	TBD
	Attorney Docket Number	Sanas-00003

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1		

If you wish to add additional non-patent literature document citation information please click the Add button Add

EXAMINER SIGNATURE

Examiner Signature	/BRYAN S BLANKENAGEL/	Date Considered	10/20/2021

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	TBD
	Attorney Docket Number	Sanas-00003

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

See attached certification statement.

The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Brandon J. Kennedy/	Date (YYYY-MM-DD)	2021-08-31
Name/Print	Brandon J. Kennedy	Registration Number	67894

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

IPR2026-00272
Krisp Technologies EX1002 Page 98

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

BIB DATA SHEET

CONFIRMATION NO. 3314

SERIAL NUMBER	FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
17/460,145	08/27/2021 RULE	704	2658	Sanas-00003

APPLICANTS
Sanas.ai Inc., Pleasanton, CA;

INVENTORS
Maxim Serebryakov, Palo Alto, CA;
Shawn Zhang, Pleasanton, CA;

**** CONTINUING DATA ***************************
This appln claims benefit of 63/185,345 05/06/2021

**** FOREIGN APPLICATIONS ***************************

**** IF REQUIRED, FOREIGN FILING LICENSE GRANTED **** ** SMALL ENTITY **
09/08/2021

Foreign Priority claimed ❑ Yes ☑ No 35 USC 119(a-d) conditions met ❑ Yes ☑ No Verified and Acknowledged /BRYAN S BLANKENAGEL/ Examiner's Signature	❑ Met after Allowance Initials	STATE OR COUNTRY	SHEETS DRAWINGS	TOTAL CLAIMS	INDEPENDENT CLAIMS
		CA	4	20	3

ADDRESS

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661
UNITED STATES

TITLE

Real-Time Accent Conversion Model

FILING FEE RECEIVED		
830	FEES: Authority has been given in Paper No._____ to charge/credit DEPOSIT ACCOUNT No._____ for following:	❑ All Fees ❑ 1.16 Fees (Filing) ❑ 1.17 Fees (Processing Ext. of time) ❑ 1.18 Fees (Issue) ❑ Other _____ ❑ Credit

Web Search History

date, time	web site	search string
10/20/2021 9:53:31 AM	Google Scholar	accent AND frame AND (monophone OR phone OR phoneme) AND (triphone OR tri-phone)
10/20/2021 10:16:42 AM	Google Scholar	accent AND conversion AND frame AND (monophone OR phone OR phoneme) AND (triphone OR tri-phone)
10/20/2021 12:59:45 PM	Google Scholar	accent AND conversion AND frame AND (monophone OR phone OR phoneme) AND (triphone OR tri-phone) [before:2021]
10/20/2021 1:00:05 PM	Google Scholar	[before:2021] accent AND conversion AND "mel spectrogram"
10/20/2021 1:00:38 PM	Google Scholar	[before:2021] accent AND conversion AND "machine learning"
10/20/2021 1:03:15 PM	Google Scholar	[before:2021] accent AND conversion AND "machine learning" and "real time"

IPR2026-00272
Krisp Technologies EX1002 Page 100

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 01/14/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
01/14/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 101

Applicant-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J. KENNEDY	Attorney of Record	

Date of Interview: 11 January 2022

Issues Discussed:

Proposed Amendment(s)

Applicant and Examiner discussed proposed amendments, and Examiner agreed that changes to the proposed amended claims would overcome the prior art rejection.

☑ Attachment

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 102

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

AGENDA FOR TELEPHONE INTERVIEW

Thank you for agreeing to interview the above-referenced application on Tuesday, January 11, 2022 at 11:30am ET. An agenda of topics is outlined below.

- Rejection of Claims 1, 10, and 19 under 35 U.S.C. § 102 in the Non-Final Office Action mailed October 25, 2021.

Without conceding the merits of the rejections, and in an attempt to advance prosecution, Applicant submits the following claim amendments and arguments for consideration.

Amendments for Consideration

1. (Proposed Amendment) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

IPR2026-00272
Krisp Technologies EX1002 Page 103

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Arguments for Consideration:

- Dirac does not teach at least "mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent" in combination with the other elements of Applicant's claims. 1
- Rather, Dirac is directed to an "accent translation model" that determines "differences in various audio characteristics between the sample sets, such as differences in pitch, tone, melody, stress, and other audio characteristics . . . audio characteristics associated with the first accent sample set are adjusted to more closely resemble those of the second accent audio sample set." Dirac at col. 2, ll. 35-44.
- At best, Dirac discloses determining that "speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." *Id.* at col. 10, ll. 7-12.
- However, this disclosure in Dirac does not amount to "mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent."

IPR2026-00272
Krisp Technologies EX1002 Page 104

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

RESPONSE TO NON-FINAL OFFICE ACTION
MAILED OCTOBER 25, 2021

In response to the Non-Final Office Action mailed October 25, 2021, Applicant submits the following amendments and remarks.

Specification Amendments begin at page 2.

Claim Amendments begin at page 3.

Remarks begin at page 12.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 105

SPECIFICATION AMENDMENTS

- *Please amend current paragraph [25] of the specification as indicated below.*

[25] The speech content may then be passed to the accent-conversion application 203 shown in FIG. 2. In some implementations, the accent-conversion application 203 may function as a virtual microphone that receives the captured speech content from the hardware microphone 202 of the computing device 200, performs accent conversion as discussed herein, and then routes the converted speech content to a digital communication application 204 (e.g., a digital communication application such as those using the trade names Zoom©, Skype©, Viber®, Telegram®, etc.) that would normally receive input speech content directly from the hardware microphone 202. Advantageously, the accent conversion may be accomplished locally on the computing device 200, which may tend to minimize the latency associated with other applications that may rely on cloud-based computing.

IPR2026-00272
Krisp Technologies EX1002 Page 106

CLAIM AMENDMENTS

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent; and

IPR2026-00272
Krisp Technologies EX1002 Page 107

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

2. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

3. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

4. (Original) The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 108

5. (Original) The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Currently Amended) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to ~~receive the indication of the first accent~~ receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a ~~first~~ second user input indicating a selection of the ~~first~~ second accent.

7. (Original) The computing device of claim 1, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 109

8. (Original) The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

IPR2026-00272
Krisp Technologies EX1002 Page 110

receive an indication of a first accent;

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

11. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived linguistic representation of the received speech content

IPR2026-00272
Krisp Technologies EX1002 Page 111

having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

12. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

13. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

14. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

15. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to ~~receive the indication of the first accent~~ receive the indication of the second

IPR2026-00272
Krisp Technologies EX1002 Page 112

accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a ~~first~~ second user input indicating a selection of the ~~first~~ second accent.

16. (Original) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

17. (Original) The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when

IPR2026-00272
Krisp Technologies EX1002 Page 113

executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

receiving an indication of a first accent;

receiving an indication of a second accent;

receiving, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the

IPR2026-00272
Krisp Technologies EX1002 Page 114

audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

20. (Original) The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 115

REMARKS

1. Summary of the Office Action

In the Non-Final Office action dated October 25, 2021, ("the Action") the Examiner (i) suggested amendments regarding the tradenames used in the specification; (ii) objected to claims 6 and 15 for informalities; (iii) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17 and 19 under 35 U.S.C. § 102 as allegedly anticipated by U.S. Pat. 10,163,451 ("Dirac"); (iv) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (v) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of U.S. Pat. 10,614,826 ("Huffman"); (vi) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (vii) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interview

A telephonic Examiner Interview took place on January 11, 2022. Participants included Examiner Bryan Blankenagel and Applicant's representative Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims, and the Examiner agreed that the suggested amendments would overcome the current rejections of record. Applicant thanks the Examiner for his time in conducting the interview.

IPR2026-00272
Krisp Technologies EX1002 Page 116

3. Amendment to the Specification

In the present response, Applicant has amended the specification as suggested by the Examiner to identify each of the trade names listed in paragraph [25] with a trademark registration symbol. No new matter has been introduced by way of these amendments.

4. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 6, 10, 15, and 19. Claims 1-20 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found generally throughout Applicant's specification. No new matter has been introduced by way of these amendments.

5. Response to Claim Objections

As noted above, the Examiner objected to claims 6 and 15 for informalities related to typographical errors. Applicant has amended claims 6 and 15 accordingly to correct these errors. No new matter has been introduced by way of these amendments.

6. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 102

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 102 as allegedly anticipated by Dirac. Applicant respectfully disagrees and submits that Applicant's claims has originally presented are not anticipated by Dirac. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. As agreed during the interview, Applicant submits that Dirac does not teach at least "mapping the set of phonemes associated with the first accent to an

IPR2026-00272
Krisp Technologies EX1002 Page 117

updated set of phonemes associated with the second accent" in combination with the other elements of Applicant's amended independent claims 1, 10, and 19.

Rather, Dirac is directed to an "accent translation model" that determines "differences in various audio characteristics between the sample sets, such as differences in pitch, tone, melody, stress, and other audio characteristics . . . audio characteristics associated with the first accent sample set are adjusted to more closely resemble those of the second accent audio sample set." Dirac at col. 2, ll. 35-44. At best, Dirac discloses determining that "speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." Id. at col. 10, ll. 7-12. However, this disclosure in Dirac does not amount to "mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent" as found in Applicant's claims.

Applicant has also reviewed the Zhao, Huffman, Sajjan, and Leff references cited by the Examiner in the Action and respectfully submits that they do not compensate for the shortcomings of Dirac.

Because Dirac does not teach each and every element of independent claims 1, 10, and 19, Dirac does not anticipate the independent claims. Consequently, Applicant requests withdrawal of the § 102 rejections of claims 1, 10, and 19, over Dirac, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 20 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

IPR2026-00272
Krisp Technologies EX1002 Page 118

7. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: January 24, 2022

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 119

Electronic Acknowledgement Receipt

EFS ID:	44822176
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	24-JAN-2022
Filing Date:	27-AUG-2021
Time Stamp:	21:53:58
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		Sanas-00003_Response_to_NFOA.pdf	152694 0c299f106955f58cc77f10b5360edb3ad868ab93	yes	15

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment/Req. Reconsideration-After Non-Final Reject	1	1
	Specification	2	2
	Claims	3	11
	Applicant Arguments/Remarks Made in an Amendment	12	15
Warnings:			
Information:			
	Total Files Size (in bytes):	152694	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 02/04/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
02/04/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 122

Office Action Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (FITF) Status** Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 01/24/2022.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☑ This action is **FINAL.** 2b) ☐ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-20 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-20 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Response to Arguments

2. Applicant's arguments with respect to claim(s) 1-20 have been considered but are moot because the new ground of rejection does not rely on any reference applied in the prior rejection of record for any teaching or matter specifically challenged in the argument.

Claim Rejections - 35 USC § 103

3. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

4. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

5. Claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, and 19 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac, in view of Gruber et al. (US 2014/0365216 A1), hereinafter referred to as Gruber.

Regarding claim 1, Dirac teaches:

A computing device comprising:

at least one microphone (Fig. 6 element 621, col. 8 lines 16-32, where a microphone is used);

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used);

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used); and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) such that the computing device is configured to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via the at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation

model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber

para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Regarding claim 2, Dirac in view of Gruber teaches:

The computing device of claim I, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac in view of Gruber teaches:

The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac in view of Gruber teaches:

The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 8, Dirac in view of Gruber teaches:

The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used), wherein the non-transitory computer- readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Regarding claim 11, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 17, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

receiving an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with the first accent (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Claim Rejections - 35 USC § 103

6. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention

and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

7. Claims 3 and 12 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and further in view of Zhao et al. (Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *INTERSPEECH* (pp. 2843-2847).), hereinafter referred to as Zhao.

Regarding claim 3, Dirac in view of Gruber teaches:

The computing device of claim 1,

Dirac in view of Gruber does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the mel-spectrograms of

Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

Regarding claim 12, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

8. Claims 5 and 14 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and further in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac in view of Gruber teaches:

The computing device of claim 1,

Dirac in view of Gruber does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

Regarding claim 14, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

9. Claims 7 and 16 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber. and further in view of Sajjan et al. (Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.), hereinafter referred to as Sajjan.

Regarding claim 7, Dirac in view of Gruber teaches:

The computing device of claim 1, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent) sound detected in the frame based on the learned linguistic representation for the first accent.

Dirac in view of Gruber does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

(ii) triphone (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used) sound detected in the frame

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

Regarding claim 16, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone based on the learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Dirac in view of Gruber does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

ii) triphone sound detected in the frame (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used)

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

10. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and further in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac in view of Gruber teaches:

The computing device of claim 1,

Dirac in view of Gruber does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac in view of Gruber teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a

synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac in view of Gruber teaches:

The method of claim 19,

Dirac in view of Gruber does not teach:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a

synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

11. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2010/0211376 A1 para [0098] teaches phoneme mapping instructions for mapping phonemes from one language to another.

12. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

13. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20140365216-A1	12-2014	GRUBER; Thomas R.	G10L13/04	704/235
*	B	US-20100211376-A1	08-2010	Chen; Ruxin	G10L15/187	704/250
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	Date	Examiner
G10L13/02	01/31/2022	BB
G10L15/02	01/31/2022	BB

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
Performed inventor search in PE2E, EAST	01/31/2022	BB
Performed text search	01/31/2022	BB
Performed text limited CPC search	01/31/2022	BB
Updated search	01/31/2022	BB

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 146

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022							
	1	✓	✓							
	2	✓	✓							
	3	✓	✓							
	4	✓	✓							
	5	✓	✓							
	6	✓	✓							
	7	✓	✓							
	8	✓	✓							
	9	✓	✓							
	10	✓	✓							
	11	✓	✓							
	12	✓	✓							
	13	✓	✓							
	14	✓	✓							
	15	✓	✓							
	16	✓	✓							
	17	✓	✓							
	18	✓	✓							
	19	✓	✓							
	20	✓	✓							

IPR2026-00272
Krisp Technologies EX1002 Page 147

Web Search History

date, time	web site	search string
1/31/2022 4:06:50 PM	Google Scholar	"accent conversion" AND phoneme
1/31/2022 4:07:10 PM	Google Scholar	"accent conversion" AND phoneme [before:2021]
1/31/2022 4:07:31 PM	Google Scholar	[before:2021] "accent conversion" AND phoneme AND synthesis
1/31/2022 4:13:39 PM	Google Scholar	[before:2021] "(voice OR speech) (conversion OR morphong)" AND phoneme AND synthesis
1/31/2022 4:13:43 PM	Google Scholar	[before:2021] "(voice OR speech) (conversion OR morphing)" AND phoneme AND synthesis
1/31/2022 4:13:51 PM	Google Scholar	[before:2021] "(voice OR speech) (conversion OR morphing)" AND (accent OR dialect) AND phoneme AND synthesis

IPR2026-00272
Krisp Technologies EX1002 Page 148

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L17	122	(accent pronunciation pronounc$4) SAME (conver$4 morph$3 chang$3) SAME (phoneme phone) near9 (map$4)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 11:25 AM
L18	49	(accent pronunciation pronounc$4) near7 (conver$4 morph$3 chang$3) SAME (phoneme phone) near9 (map$4)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 11:30 AM
L19	51	(voice speech) near7 (conver$4 morph$3 chang$3) SAME (phoneme phone) near9 (map$4) SAME (accent pronunciation pronounc$4 dialect)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 11:49 AM
L20	24	((voice speech) near7 (conver$4 morph$3 chang$3)).ab. AND (phoneme phone) near9 (map$4) SAME (accent pronunciation pronounc$4 dialect)	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 12:10 PM
L21	3	20 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2022/01/31 04:04 PM
L22	0	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/01/31 04:04 PM

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 03/23/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
03/23/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 150

Applicant-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J KENNEDY	Attorney of Record	

Date of Interview: 18 March 2022

Issues Discussed:

Proposed Amendment(s)

Applicant and Examiner discussed the proposed amendments, and Examiner agreed that they would overcome the prior art rejection, and would require further search and consideration.

☑ Attachment

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 151

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

AGENDA FOR TELEPHONE INTERVIEW

Thank you for agreeing to interview the above-referenced application on Friday, March 18, 2022 at 2:00pm ET. An agenda of topics is outlined below.

- Rejection of Claims 1, 10, and 19 under 35 U.S.C. § 103 in the Final Office Action mailed February 4, 2022.

Without conceding the merits of the rejections, and in an attempt to advance prosecution, Applicant submits the following proposed claim amendments and arguments for consideration.

Amendments for Consideration

1. (Proposed Amendment) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

IPR2026-00272
Krisp Technologies EX1002 Page 152

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

21. (Proposed New Claim) The computing device of claim 1, wherein the derived linguistic representation of the received speech content having the first accent does not include a textual representation of the received speech content.

Arguments for Consideration:

- Gruber does not teach at least "mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation" in combination with the other elements of Applicant's claims.
- Rather, Gruber states that "Both the representation 404 in the recognition alphabet and the representation 406 in the synthesis alphabet are based on the same pronunciation, and, therefore, the user's preferred pronunciation will both be accurately recognized by the STT processing module 330 and accurately synthesized by the speech synthesis module 265." Gruber at [0116].

IPR2026-00272
Krisp Technologies EX1002 Page 153

Doc Code: A.NE.AFCP
Document Description: After Final Consideration Pilot Program Request

PTO/SB/434 (05-13)

CERTIFICATION AND REQUEST FOR CONSIDERATION UNDER THE AFTER FINAL CONSIDERATION PILOT PROGRAM 2.0

Practitioner Docket No.:	Application No.:	Filing Date:
Sanas-00003	17/460,145	August 27, 2021
First Named Inventor:	**Title:**	
Maxim Serebryakov	Real-Time Accent Conversion Model	

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS CONSIDERATION UNDER THE AFTER FINAL CONSIDERATION PILOT PROGRAM 2.0 (AFCP 2.0) OF THE ACCOMPANYING RESPONSE UNDER 37 CFR 1.116.

1. The above-identified application is (i) an original utility, plant, or design nonprovisional application filed under 35 U.S.C. 111(a) [a continuing application (*e.g.*, a continuation or divisional application) is filed under 35 U.S.C. 111(a) and is eligible under (i)], or (ii) an international application that has entered the national stage in compliance with 35 U.S.C. 371(c).
2. The above-identified application contains an outstanding final rejection.
3. Submitted herewith is a response under 37 CFR 1.116 to the outstanding final rejection. The response includes an amendment to at least one independent claim, and the amendment does not broaden the scope of the independent claim in any aspect.
4. This certification and request for consideration under AFCP 2.0 is the only AFCP 2.0 certification and request filed in response to the outstanding final rejection.
5. Applicant is willing and available to participate in any interview requested by the examiner concerning the present response.
6. This certification and request is being filed electronically using the Office's electronic filing system (EFS-Web).
7. Any fees that would be necessary consistent with current practice concerning responses after final rejection under 37 CFR 1.116, *e.g.*, extension of time fees, are being concurrently filed herewith. [There is no additional fee required to request consideration under AFCP 2.0.]
8. By filing this certification and request, applicant acknowledges the following:
    - Reissue applications and reexamination proceedings are not eligible to participate in AFCP 2.0.
    - The examiner will verify that the AFCP 2.0 submission is compliant, *i.e.*, that the requirements of the program have been met (see items 1 to 7 above). For compliant submissions:
        - The examiner will review the response under 37 CFR 1.116 to determine if additional search and/or consideration (i) is necessitated by the amendment and (ii) could be completed within the time allotted under AFCP 2.0. If additional search and/or consideration is required but cannot be completed within the allotted time, the examiner will process the submission consistent with current practice concerning responses after final rejection under 37 CFR 1.116, *e.g.*, by mailing an advisory action.
        - If the examiner determines that the amendment does not necessitate additional search and/or consideration, or if the examiner determines that additional search and/or consideration is required and could be completed within the allotted time, then the examiner will consider whether the amendment places the application in condition for allowance (after completing the additional search and/or consideration, if required). If the examiner determines that the amendment does not place the application in condition for allowance, then the examiner will contact the applicant and request an interview.
            - The interview will be conducted by the examiner, and if the examiner does not have negotiation authority, a primary examiner and/or supervisory patent examiner will also participate.
            - If the applicant declines the interview, or if the interview cannot be scheduled within ten (10) calendar days from the date that the examiner first contacts the applicant, then the examiner will proceed consistent with current practice concerning responses after final rejection under 37 CFR 1.116.

Signature	Date
/Brandon J. Kennedy/	April 1, 2022
Name (Print/Typed) Brandon J. Kennedy	**Practitioner Registration No.** 67894

***Note**: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required, see below*.*

[✓] * Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of: Sanas.ai Serial No.: 17/460,145 Filing Date: August 27, 2021 Title: Real-Time Accent Conversion Model	Confirmation No.: 3314 Examiner: Blankenagel, Bryan S Group Art Unit: 2658

RESPONSE TO FINAL OFFICE ACTION
MAILED FEBRUARY 4, 2022

In response to the Final Office Action mailed February 4, 2022, Applicant submits the following amendments and remarks pursuant to 37 C.F.R. § 1.116 and the After Final Consideration Pilot Program 2.0.

Claim Amendments begin at page 2.

Remarks begin at page 11.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 156

CLAIM AMENDMENTS

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

IPR2026-00272
Krisp Technologies EX1002 Page 157

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

2. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

3. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

4. (Original) The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 158

5. (Original) The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Previously Presented) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent.

7. (Original) The computing device of claim 1, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 159

8. (Original) The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent;

IPR2026-00272
Krisp Technologies EX1002 Page 160

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

11. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived linguistic representation of the received speech content

IPR2026-00272
Krisp Technologies EX1002 Page 161

having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

12. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

13. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

14. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions

IPR2026-00272
Krisp Technologies EX1002 Page 162

that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent.

16. (Original) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the learned linguistic representation for the first accent.

17. (Original) The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when

IPR2026-00272
Krisp Technologies EX1002 Page 163

executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

receiving an indication of a first accent;

receiving an indication of a second accent;

receiving, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent;

based on the derived linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated

IPR2026-00272
Krisp Technologies EX1002 Page 164

set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

20. (Original) The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 165

REMARKS

1. Summary of the Office Action

In the Final Office action dated February 4, 2022, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17 and 19 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interview

A telephonic Examiner Interview took place on March 18, 2022. Participants included Examiner Bryan Blankenagel and Applicant's representative Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims, and the Examiner agreed that the suggested amendments would overcome the current rejections of record. Applicant thanks the Examiner for his time in conducting the interview.

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 10, and 19. Claims 1-20 are pending, of which claims 1, 10, and 19 are independent and the remainder

IPR2026-00272
Krisp Technologies EX1002 Page 166

are dependent. Support for the claim amendments can be found generally throughout Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac and Gruber. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. As agreed during the interview, Applicant submits that the combination of Dirac and Gruber does not teach at least "wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation . . ." in combination with the other elements of Applicant's amended independent claims 1, 10, and 19.

In the Action at p. 4, the Examiner admitted that Dirac does not teach "synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent." For similar reasons, Applicant submits that the model discussed in Dirac also does not teach ". . . an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation." As noted in Applicant's prior response, the model in Dirac determines "differences in various audio characteristics between the sample sets, such as differences in pitch, tone, melody, stress, and other audio characteristics . . . audio characteristics associated with the first accent sample set are adjusted to more closely resemble those of the second accent audio sample set." Dirac at col. 2,

IPR2026-00272
Krisp Technologies EX1002 Page 167

ll. 35-44. However, there is nothing in Dirac that accounts for differences in pronunciation between accents.

Applicant's specification at paragraph [5] specifically highlights this distinction. At best, Dirac discloses determining that "speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." *Id.* at col. 10, ll. 7-12. But this does not amount to "a second pronunciation of the speech content that is different from the first pronunciation" as found in Applicant's claims.

Gruber also does not teach these elements that are missing from Dirac. Rather, Gruber is directed to a system that allows "users to specify a correct or a preferred pronunciation of words, such as proper names, so that a digital assistant can both recognize and synthesize the word based on the specified pronunciation." Gruber at para. [0007]. Thus, Gruber is directed to synthesizing speech that has the same pronunciation specified by the user. Indeed, with reference to FIG. 4, Gruber states that "[b]oth the representation 404 in the recognition alphabet and the representation 406 in the synthesis alphabet are based on the same pronunciation, and, therefore, the user's preferred pronunciation will both be accurately recognized by the STT processing module 330 and accurately synthesized by the speech synthesis module 265." Gruber at [0116]. Accordingly, nowhere does Gruber teach or suggest "mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation"

IPR2026-00272
Krisp Technologies EX1002 Page 168

Applicant has also reviewed the Zhao, Huffman, Sajjan, and Leff references cited by the Examiner in the Action and respectfully submits that they do not compensate for the shortcomings of Dirac and Gruber.

Because the combination of Dirac and Gruber does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of claims 1, 10, and 19, over the combination of Dirac and Gruber, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 20 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

5. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: April 1, 2022

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 169

Electronic Acknowledgement Receipt

EFS ID:	45379475
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	01-APR-2022
Filing Date:	27-AUG-2021
Time Stamp:	22:13:13
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	After Final Consideration Program Request	Sanas-00003_AFCP_Request.pdf	226731 14dec3760729c979212ed2d407889c5fb72faec6	no	2

Warnings:

Information:					
2		Sanas-00003_Response_to_FOA.pdf	124943 8d10dc9ba044aa06282428b68a40f07b19cfefa0	yes	14

Multipart Description/PDF files in .zip description		
Document Description	**Start**	**End**
Response After Final Action	1	1
Claims	2	10
Applicant Arguments/Remarks Made in an Amendment	11	14

Warnings:	
Information:	
Total Files Size (in bytes):	351674

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	04/01/2022	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 20	Minus	** 20	= 0		x $50 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $240 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/MONICA D FRANCIS/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 04/11/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
04/11/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 173

Advisory Action Before the Filing of an Appeal Brief	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (FITF) Status** Yes

--The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

THE REPLY FILED 01 April 2022 FAILS TO PLACE THIS APPLICATION IN CONDITION FOR ALLOWANCE.

NO NOTICE OF APPEAL FILED

1. ☑ The reply was filed after a final rejection. No Notice of Appeal has been filed. To avoid abandonment of this application, applicant must timely file one of the following replies: (1) an amendment, affidavit, or other evidence, which places the application in condition for allowance; (2) a Notice of Appeal (with appeal fee) in compliance with 37 CFR 41.31; or (3) a Request for Continued Examination (RCE) in compliance with 37 CFR 1.114 if this is a utility or plant application. Note that RCEs are not permitted in design applications. The reply must be filed within one of the following time periods:
   a) ☐ The period for reply expires ____ months from the mailing date of the final rejection.
   b) ☑ The period for reply expires on: (1) the mailing date of this Advisory Action; or (2) the date set forth in the final rejection, whichever is later. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the mailing date of the final rejection.
   c) ☐ A prior Advisory Action was mailed more than 3 months after the mailing date of the final rejection in response to a first after-final reply filed within 2 months of the mailing date of the final rejection.The current period for reply expires ____ months from the mailing date of *the prior Advisory Action* or SIX MONTHS from the mailing date of the final rejection, whichever is earlier.
   Examiner Note: If box 1 is checked, check either box (a), (b) or (c). ONLY CHECK BOX (b) WHEN THIS ADVISORY ACTION IS THE FIRST RESPONSE TO APPLICANTS FIRST AFTER-FINAL REPLY WHICH WAS FILED WITHIN TWO MONTHS OF THE FINAL REJECTION. ONLY CHECK BOX (c) IN THE LIMITED SITUATION SET FORTH UNDER BOX (c). See MPEP 706.07(f).

Extensions of time may be obtained under 37 CFR 1.136(a). The date on which the petition under 37 CFR 1.136(a) and the appropriate extension fee have been filed is the date for purposes of determining the period of extension and the corresponding amount of the fee. The appropriate extension fee under 37 CFR 1.17(a) is calculated from: (1) the expiration date of the shortened statutory period for reply originally set in the final Office action; or (2) as set forth in (b) or (c) above, if checked. Any reply received by the Office later than three months after the mailing date of the final rejection, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

NOTICE OF APPEAL

2. ☐ The Notice of Appeal was filed on _____. A brief in compliance with 37 CFR 41.37 must be filed within two months of the date of filing the Notice of Appeal (37 CFR 41.37(a)), or any extension thereof (37 CFR 41.37(e)), to avoid dismissal of the appeal. Since a Notice of Appeal has been filed, any reply must be filed within the time period set forth in 37CFR 41.37(a).

AMENDMENTS

3. ☑ The proposed amendments filed after a final rejection, but prior to the date of filing a brief, will not be entered because
   a) ☑ They raise new issues that would require further consideration and/or search (see NOTE below);
   b) ☐ They raise the issue of new matter (see NOTE below);
   c) ☐ They are not deemed to place the application in better form for appeal by materially reducing or simplifying the issues for appeal; and/or
   d) ☐ They present additional claims without canceling a corresponding number of finally rejected claims.
   NOTE: See below (See 37 CFR 1.116 and 41.33(a)).
4. ☐ The amendments are not in compliance with 37 CFR 1.121. See attached Notice of Non-Compliant Amendment (PTOL-324).
5. ☐ Applicant's reply has overcome the following rejection(s): _____
6. ☐ Newly proposed or amended claim(s) _____ would be allowable if submitted in a separate, timely filed amendment canceling the non-allowable claim(s).
7. ☑ For purposes of appeal, the proposed amendment(s):(a) ☑ will not be entered, or (b) ☐ will be entered, and an explanation of how the new or amended claims would be rejected is provided below or appended.

AFFIDAVIT OR OTHER EVIDENCE

8. ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____
9. ☐ The affidavit or other evidence filed after final action, but before or on the date of filing a Notice of Appeal will not be entered because applicant failed to provide a showing of good and sufficient reasons why the affidavit or other evidence is necessary and was not earlier presented. See 37 CFR 1.116(e).
10. ☐ The affidavit or other evidence filed after the date of filing the Notice of Appeal, but prior to the date of filing a brief, will not be entered because the affidavit or other evidence failed to overcome all rejections under appeal and/or appellant fails to provide a showing of good and sufficient reasons why it is necessary and was not earlier presented. See 37 CFR 41.33(d)(1).
11. ☐ The affidavit or other evidence is entered. An explanation of the status of the claims after entry is below or attached.

REQUEST FOR RECONSIDERATION/OTHER

12. ☑ The request for reconsideration has been considered but does NOT place the application in condition for allowance because: See Continuation Sheet.
13. ☐ Note the attached Information *Disclosure Statement*(s). (PTO/SB/08) Paper No(s). _____
14. ☑ Other: See attached PTO-413, PTO-892, PTO-2323.

STATUS OF CLAIMS

15. The status of the claim(s) is (or will be) as follows:
    Claim(s) allowed:_____.
    Claim(s) objected to:_____.
    Claim(s) rejected:1-20.
    Claim(s) withdrawn from consideration:_____.

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 174

Continuation of REQUEST FOR RECONSIDERATION/OTHER 12. The request for reconsideration has been considered but does NOT place the application in condition for allowance because: The amended claims would require further search and consideration. A brief search conducted under the AFCP2.0 guidelines revealed the following references, which appear to be relevant to the amended claims: US 2014/0258462 A1 para [0087], [0089]; US 2008/0133241 A1 para [0037]; and Zheng, D. C. (2011). (Accent Conversion via Formant-based Spectral Mapping and Pitch Contour Modification.) pages 19-20.

Examiner-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J. KENNEDY	Attorney of Record	

Date of Interview: 07 April 2022

Issues Discussed:

Proposed Amendment(s)

Examiner informed Applicant of the results of the search conducted under the AFCP2.0 guidelines, including the following references: US 2014/0258462 A1 para [0087], [0089]; US 2008/0133241 A1 para [0037]; and Zheng, D. C. (2011). (Accent Conversion via Formant-based Spectral Mapping and Pitch Contour Modification.) pages 19-20.

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: It is not necessary for applicant to provide a separate record of the substance of interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 176

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20080133241-A1	06-2008	Baker; David Robert	G10L19/0018	704/260
*	B	US-20140258462-A1	09-2014	Hwang; Douglas	G06Q30/0621	709/219
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Zheng, D. C. (2011). Accent Conversion via Formant-based Spectral Mapping and Pitch Contour Modification.
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

IPR2026-00272
Krisp Technologies EX1002 Page 177

AFCP 2.0 Decision

Application No.	Applicant(s)	
17/460,145	Serebryakov et al.	
Examiner	**Art Unit**	**AIA (FITF) Status**
BRYAN S BLANKENAGEL	2658	Yes

This is in response to the After Final Consideration Pilot request filed 01 April 2022.

1. **Improper Request** – The AFCP 2.0 request is improper for the following reason(s) and the after final amendment submitted with the request will be treated under pre-pilot procedure.

   ☐ An AFCP 2.0 request form PTO/SB/434 (or equivalent document) was not submitted.

   ☐ A non-broadening amendment to at least one independent claim was not submitted.

   ☐ The request is not the first proper AFCP 2.0 request submitted in response to the most recent final rejection.

   ☐ Other: _____

2. **Proper Request**

   A. After final amendment submitted with the request will not be treated under AFCP 2.0.
   The after final amendment cannot be reviewed and a search conducted within the guidelines of the pilot program.

   ☐ The after final amendment will be treated under pre-pilot procedure.

   B. Updated search and/or completed additional consideration.
   The examiner performed an updated search and/or completed additional consideration of the after final amendment within the time authorized for the pilot program. The result(s) of the updated search and/or completed additional consideration are:

   ☐ 1. All of the rejections in the most recent final Office action are overcome and a Notice of Allowance is issued herewith.

   ☐ 2. The after final amendment would not overcome all of the rejections in the most recent final Office action . See attached interview summary for further details.

   ☐ 3. The after final amendment was reviewed, and it raises a new issue(s). See attached interview summary for further details.

   ☑ 4. The after final amendment raises new issues, but would overcome all of the rejections in the most recent final Office action. A decision on determining allowability could not be made within the guidelines of the pilot. See attached interview summary for further details, including any newly discovered prior art.

   ☐ 5. Other: _____

   Examiner Note: Please attach an interview summary when necessary as described above.

IPR2026-00272
Krisp Technologies EX1002 Page 178

Do Not Enter /B.B./ 04/05/2022

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

RESPONSE TO FINAL OFFICE ACTION
MAILED FEBRUARY 4, 2022

In response to the Final Office Action mailed February 4, 2022, Applicant submits the following amendments and remarks pursuant to 37 C.F.R. § 1.116 and the After Final Consideration Pilot Program 2.0.

Claim Amendments begin at page 2.

Remarks begin at page 11.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 179

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	05/04/2022	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 20	Minus	** 20	= 0		x $50 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $240 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/ALLYSON PURNELL/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc code: RCEX
Doc description: Request for Continued Examination (RCE)

PTO/SB/30EFS (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

REQUEST FOR CONTINUED EXAMINATION(RCE)TRANSMITTAL
(Submitted Only via EFS-Web)

Application Number	17/460,145	Filing Date	2021-08-27	Docket Number (if applicable)	Sanas-00003	Art Unit	2658
First Named Inventor	Maxim Serebryakov			Examiner Name	Bryan S. Blankenagel		

This is a Request for Continued Examination (RCE) under 37 CFR 1.114 of the above-identified application.
Request for Continued Examination (RCE) practice under 37 CFR 1.114 does not apply to any utility or plant application filed prior to June 8, 1995, or to any design application. The Instruction Sheet for this form is located at WWW.USPTO.GOV

SUBMISSION REQUIRED UNDER 37 CFR 1.114

Note: If the RCE is proper, any previously filed unentered amendments and amendments enclosed with the RCE will be entered in the order in which they were filed unless applicant instructs otherwise. If applicant does not wish to have any previously filed unentered amendment(s) entered, applicant must request non-entry of such amendment(s).

☐ Previously submitted. If a final Office action is outstanding, any amendments filed after the final Office action may be considered as a submission even if this box is not checked.

- ☐ Consider the arguments in the Appeal Brief or Reply Brief previously filed on _____
- ☐ Other _____

☒ Enclosed

- ☒ Amendment/Reply
- ☐ Information Disclosure Statement (IDS)
- ☐ Affidavit(s)/ Declaration(s)
- ☐ Other _____

MISCELLANEOUS

☐ Suspension of action on the above-identified application is requested under 37 CFR 1.103(c) for a period of months _____
(Period of suspension shall not exceed 3 months; Fee under 37 CFR 1.17(i) required)

☐ Other _____

FEES

☒ **The RCE fee under 37 CFR 1.17(e) is required by 37 CFR 1.114 when the RCE is filed.**
The Director is hereby authorized to charge any underpayment of fees, or credit any overpayments, to Deposit Account No 506632

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT REQUIRED

☒ Patent Practitioner Signature

Applicant Signature

PTO/SB/30EFS (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Signature of Registered U.S. Patent Practitioner			
Signature	/Brandon J. Kennedy/	Date (YYYY-MM-DD)	2022-05-04
Name	Brandon J. Kennedy	Registration Number	67894

This collection of information is required by 37 CFR 1.114. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IPR2026-00272
Krisp Technologies EX1002 Page 183

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

RESPONSE TO FINAL OFFICE ACTION
MAILED FEBRUARY 4, 2022

In response to the Final Office Action mailed February 4, 2022, Applicant submits the following amendments and remarks pursuant to 37 C.F.R. § 1.116.

Claim Amendments begin at page 2.

Remarks begin at page 11.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 184

CLAIM AMENDMENTS

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent;

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

IPR2026-00272
Krisp Technologies EX1002 Page 185

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

2. (Currently Amended) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

3. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

4. (Original) The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 186

5. (Original) The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Previously Presented) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent.

7. (Currently Amended) The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the non-text linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 187

8. (Original) The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent;

IPR2026-00272
Krisp Technologies EX1002 Page 188

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with [[the]] a first ~~accent~~ pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent;

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

11. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent

IPR2026-00272
Krisp Technologies EX1002 Page 189

comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent.

12. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

13. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

14. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

IPR2026-00272
Krisp Technologies EX1002 Page 190

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent.

16. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the non-text linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

17. (Original) The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device

IPR2026-00272
Krisp Technologies EX1002 Page 191

to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

receiving an indication of a first accent;

receiving an indication of a second accent;

receiving, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with {{the}} a first ~~accent~~ pronunciation of the speech content;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent;

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio

IPR2026-00272
Krisp Technologies EX1002 Page 192

data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with [[the]] a second ~~accent~~ pronunciation of the speech content that is different from the first pronunciation; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

20. (Original) The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 193

REMARKS

1. Summary of the Office Action

In the Final Office action dated February 4, 2022, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17 and 19 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interviews

A telephonic Examiner Interview took place on March 18, 2022. Participants included Examiner Bryan Blankenagel and Applicant's representative Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims, and the Examiner agreed that the suggested amendments would overcome the current rejections of record.

A follow up interview took place on April 7, 2022 with the same participants, in which Examiner Blankenagel informed Applicant's representative of the results of the search conducted under the AFCP 2.0 guidelines. Applicant thanks the Examiner for his time in conducting the interviews.

IPR2026-00272
Krisp Technologies EX1002 Page 194

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 2, 7, 10, 11, 16, and 19. Claims 1-20 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found generally throughout Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac and Gruber. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. As agreed during the interview, Applicant submits that the combination of Dirac and Gruber does not teach at least "wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation . . ." in combination with the other elements of Applicant's amended independent claims 1, 10, and 19.

In the Action at p. 4, the Examiner admitted that Dirac does not teach "synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent." For similar reasons, Applicant submits that the model discussed in Dirac also does not teach ". . . an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation." As

IPR2026-00272
Krisp Technologies EX1002 Page 195

noted in Applicant's prior response, the model in Dirac determines "differences in various audio characteristics between the sample sets, such as differences in pitch, tone, melody, stress, and other audio characteristics . . . audio characteristics associated with the first accent sample set are adjusted to more closely resemble those of the second accent audio sample set." Dirac at col. 2, ll. 35-44. However, there is nothing in Dirac that accounts for differences in pronunciation between accents.

Applicant's specification at paragraph [5] specifically highlights this distinction. At best, Dirac discloses determining that "speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." *Id.* at col. 10, ll. 7-12. But this does not amount to "a second pronunciation of the speech content that is different from the first pronunciation" as found in Applicant's claims.

Gruber also does not teach these elements that are missing from Dirac. Rather, Gruber is directed to a system that allows "users to specify a correct or a preferred pronunciation of words, such as proper names, so that a digital assistant can both recognize and synthesize the word based on the specified pronunciation." Gruber at para. [0007]. Thus, Gruber is directed to synthesizing speech that has the same pronunciation specified by the user. Indeed, with reference to FIG. 4, Gruber states that "[b]oth the representation 404 in the recognition alphabet and the representation 406 in the synthesis alphabet are based on the same pronunciation, and, therefore, the user's preferred pronunciation will both be accurately recognized by the STT processing module 330 and accurately synthesized by the speech synthesis module 265." Gruber at [0116]. Accordingly, nowhere does Gruber teach or suggest "mapping the set of phonemes associated with the first

IPR2026-00272
Krisp Technologies EX1002 Page 196

accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation"

Applicant has also reviewed the Zhao, Huffman, Sajjan, and Leff references cited by the Examiner in the Action and respectfully submits that they do not compensate for the shortcomings of Dirac and Gruber.

Because the combination of Dirac and Gruber does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of claims 1, 10, and 19, over the combination of Dirac and Gruber, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 20 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

5. Response to Advisory Action

In the Advisory Action, AFCP 2.0 Decision, and Interview Summary mailed on April 11, 2022, the Examiner indicated that although the amendments presented in Applicant's after-final response filed on April 1, 2022 would overcome the rejections in the Final Office Action, further search and consideration would be required. Following a brief search, the Examiner indicated that US Pub. 2014/0258462 ("Hwang"), US Pub. 2008/0133241 ("Baker"), and Zheng, D.C. (2011) Accent Conversion via Formant-based Spectral Mapping and Pitch Contour Modification ("Zheng") appeared to be relevant to the amended claims. Applicant thanks the Examiner for his diligence under the AFCP 2.0 program and for the identification of potentially relevant references in the Advisory Action.

IPR2026-00272
Krisp Technologies EX1002 Page 197

Without conceding the relevance of the identified references, Applicant has further amended the claims as noted above in an attempt to expedite prosecution following the present request for continued examination. In particular, the claims have been amended to include "derive, using a first machine-learning algorithm . . . a non-text linguistic representation of the received speech content having the first accent" and "based on the derived non-text linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm . . . audio data representative of the received speech content having the second accent." Support for the additional amendments can be found in at least paragraphs [37] and [44] of Applicant's originally-filed specification.

Applicant respectfully submits that neither the references cited in the Final Office Action, the references cited in the Advisory Action, nor their combination teaches a "non-text" linguistic representation when considered in combination with the other elements of Applicant's claims. For at least this reason, Applicant submits that the claims are in condition for allowance.

6. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: May 4, 2022

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 198

Electronic Patent Application Fee Transmittal

Application Number:	17460145
Filing Date:	27-Aug-2021
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Filer:	Brandon Jacob Kennedy/Holly Mueller
Attorney Docket Number:	Sanas-00003

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
RCE- 1ST REQUEST	2801	1	680	680
		Total in USD ($)		**680**

Electronic Acknowledgement Receipt	
EFS ID:	45635552
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	04-MAY-2022
Filing Date:	27-AUG-2021
Time Stamp:	21:33:58
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$680
RAM confirmation Number	E202254L34201611
Deposit Account	506632
Authorized User	Brandon Kennedy
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: 37 CFR 1.16 (National application filing, search, and examination fees) 37 CFR 1.17 (Patent application and reexamination processing fees)	

37 CFR 1.19 (Document supply fees)

37 CFR 1.20 (Post Issuance fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Request for Continued Examination (RCE)	Sanas-00003_RCE.pdf	1667140 c88387433154ecf978064d0ecca102d2324f71da	no	3
Warnings:					
Information:					
2		Sanas-00003_Response_to_AdvisoryAction.pdf	128410 f51fec38288d97c2b571b7eb0c4f8496b73aa679	yes	15

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Response After Final Action	1	1
	Claims	2	10
	Applicant Arguments/Remarks Made in an Amendment	11	15

Warnings:					
Information:					
3	Fee Worksheet (SB06)	fee-info.pdf	37933 8d302afd58ad1c699f31db52a6e35b85fba4f930	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	1833483		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 08/10/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
08/10/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 204

Office Action Summary	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 05/04/2022.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-20 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-20 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Continued Examination Under 37 CFR 1.114

2. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on 05/04/2022 has been entered.

Response to Arguments

3. Applicant's arguments with respect to claim(s) 1-20 have been considered but are moot because the new ground of rejection does not rely on any reference applied in the prior rejection of record for any teaching or matter specifically challenged in the argument.

4. Applicant's arguments filed 05/04/2022 have been fully considered but they are not persuasive. Regarding arguments on page 15 of the Remarks, Examiner notes that the linguistic representations as taught in Dirac col. 5 lines 9-27 are accent audio characteristics of the accent sample sets, and are not text.

Claim Rejections - 35 USC § 103

5. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

6. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

7. Claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, and 19 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac, in view of Gruber et al. (US 2014/0365216 A1), hereinafter referred to as Gruber, and further in view of Lurie et al. (US 2022/0180762 A1), hereinafter referred to as Lurie.

Regarding claim 1, Dirac teaches:

A computing device comprising:

at least one microphone (Fig. 6 element 621, col. 8 lines 16-32, where a microphone is used);

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used);

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used); and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) such that the computing device is configured to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via the at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125]) to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Regarding claim 2, Dirac in view of Gruber and Lurie teaches:

The computing device of claim I, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first

accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 8, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used), wherein the non-transitory computer- readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125])

to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Regarding claim 11, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the audio data comprising the first accent and the audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 17, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

receiving an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the received speech content having the first accent (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the received speech content having the first accent, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning); and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125]) to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Claim Rejections - 35 USC § 103

8. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

9. Claims 3 and 12 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Zhao et al. (Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *INTERSPEECH* (pp. 2843-2847).), hereinafter referred to as Zhao.

Regarding claim 3, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms

representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

Regarding claim 12, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

10. Claims 5 and 14 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

Regarding claim 14, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

11. Claims 7 and 16 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Sajjan et al. (Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.), hereinafter referred to as Sajjan.

Regarding claim 7, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the non-text linguistic representation of the received speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent) sound detected in the frame based on the non-text learned linguistic representation for the first accent.

Dirac in view of Gruber and Lurie does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

(ii) triphone (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used) sound detected in the frame

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber and Lurie (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

Regarding claim 16, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the non-text linguistic representation of the received speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Dirac in view of Gruber and Lurie does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

ii) triphone sound detected in the frame (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used)

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber and Lurie (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

12. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a

synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac in view of Gruber and Lurie teaches:

The method of claim 19,

Dirac in view of Gruber and Lurie does not teach:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

13. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2003/0018473 A1 para [0082-83] teaches certain accents having strongly pronounced phonemes, while the abstract teaches automatically changing an accent.

14. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20220180762-A1	06-2022	Lurie; Lisa Beth	G10L15/22	1/1
*	B	US-20030018473-A1	01-2003	OHNISHI, HIROKI	G10L13/10	704/258
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	**Date**	**Examiner**
G10L13/02	08/04/2022	BB
G10L15/02	08/04/2022	BB

CPC Combination Sets - Searched*		
Symbol	**Date**	**Examiner**

US Classification - Searched*			
Class	**Subclass**	**Date**	**Examiner**

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	**Date**	**Examiner**
Performed inventor search in PE2E, EAST	08/04/2022	BB
Performed text search	08/04/2022	BB
Performed text limited CPC search	08/04/2022	BB
Updated search	08/04/2022	BB

Interference Search			
US Class/CPC Symbol	**US Subclass/CPC Group**	**Date**	**Examiner**

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 230

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022	08/04/2022						
	1	✓	✓	✓						
	2	✓	✓	✓						
	3	✓	✓	✓						
	4	✓	✓	✓						
	5	✓	✓	✓						
	6	✓	✓	✓						
	7	✓	✓	✓						
	8	✓	✓	✓						
	9	✓	✓	✓						
	10	✓	✓	✓						
	11	✓	✓	✓						
	12	✓	✓	✓						
	13	✓	✓	✓						
	14	✓	✓	✓						
	15	✓	✓	✓						
	16	✓	✓	✓						
	17	✓	✓	✓						
	18	✓	✓	✓						
	19	✓	✓	✓						
	20	✓	✓	✓						

IPR2026-00272
Krisp Technologies EX1002 Page 231

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L28	234	accent near9 phoneme near9 (pronunciation pronounc$3)	(US-PGPUB; USPAT)	OR	ON	ON	2022/08/04 12:51 PM
L29	136	accent near9 phoneme near9 (pronunciation pronounc$3) SAME (conver$4 chang$3 shift$3)	(US-PGPUB; USPAT)	OR	ON	ON	2022/08/04 12:52 PM
L30	19	(conver$4 chang$3 shift$3) near5 accent AND accent near9 phoneme near9 (pronunciation pronounc$3)	(US-PGPUB; USPAT)	OR	ON	ON	2022/08/04 03:52 PM
L31	4	30 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2022/08/04 03:55 PM
L32	0	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2022/08/04 03:55 PM

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

Web Search History

date, time	web site	search string
8/4/2022 3:55:37 PM	Google Scholar	"accent conversion" AND phoneme
8/4/2022 3:55:46 PM	Google Scholar	"accent conversion" AND phoneme [before:2021]
8/4/2022 3:55:56 PM	Google Scholar	[before:2021] "accent conversion" AND phoneme AND (pronunciation OR pronounce)
8/4/2022 3:56:03 PM	Google Scholar	[before:2021] "accent conversion" AND phoneme AND (pronunciation OR pronounce) and mapping
8/4/2022 4:12:42 PM	Google Scholar	[before:2021] accent AND conversion AND phoneme AND (pronunciation OR pronounce) and mapping
8/4/2022 4:12:56 PM	Google Scholar	[before:2021] accent AND conversion AND phoneme AND (pronunciation OR pronounce) and mapping and synthesize

IPR2026-00272
Krisp Technologies EX1002 Page 233

Doc code: IDS
Doc description: Information Disclosure Statement (IDS) Filed

PTO/SB/08a (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (**Not for submission under 37 CFR 1.99)**		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	Bryan S. Blankenagel
	Attorney Docket Number	Sanas-00003

U.S.PATENTS Remove

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Patent citation information please click the Add button. Add

U.S.PATENT APPLICATION PUBLICATIONS Remove

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button. Add

FOREIGN PATENT DOCUMENTS Remove

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2] i	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]
	1	111462769	CN	A	2020-07-28	SPEECHX LTD		☓
	2	112382267	CN	A	2021-02-19	BEIJING YOUZHUJU NETWORK TECH CO LTD		☒
	3	112382270	CN	A	2021-02-19	BEIJING YOUZHUJU NETWORK TECH CO LTD		☒

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	Bryan S. Blankenagel
	Attorney Docket Number	Sanas-00003

	4	2005234418	JP	A	2005-09-02	ATR ADVANCED TELECOMM RES INST		☒

If you wish to add additional Foreign Patent Document citation information please click the Add button **Add**

NON-PATENT LITERATURE DOCUMENTS **Remove**

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
	1	INTERNATIONAL SEARCHING AUTHORITY, PCT International Search Report and Written Opinion, PCT International Application No. PCT/US2022/028156, dated 17 August 2022, 8 pages.	

If you wish to add additional non-patent literature document citation information please click the Add button **Add**

EXAMINER SIGNATURE

Examiner Signature		Date Considered	

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	Bryan S. Blankenagel
	Attorney Docket Number	Sanas-00003

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☒ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

See attached certification statement.

The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Brandon J. Kennedy/	Date (YYYY-MM-DD)	2022-09-20
Name/Print	Brandon J. Kennedy	Registration Number	67,894

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IPR2026-00272
Krisp Technologies EX1002 Page 237

Espacenet

CN111462769A End-to-end accent conversion method

Applicants: SPEECHX LTD

Inventors: LIU SONGXIANG,WANG DISONG,CAO YUEWEN,SUN LIFA,WU XIXIN,KANG SHIYIN,WU ZHIYONG,LIU XUNYING,MENG MEILING

Classifications:

IPC ***G10L13/02; G10L13/04; G10L15/02; G10L15/06; G10L19/00; G10L19/16; G10L21/013; G10L25/24; G10L25/30;***

CPC **G10L13/02 (CN); G10L15/02 (CN); G10L15/063 (CN); G10L19/0018 (CN); G10L19/16 (CN); G10L21/013 (CN); G10L25/24 (CN); G10L25/30 (CN);** G10L2021/0135 (CN);

Priorities: CN202010239586A 2020-03-30

Application: CN202010239586A·2020-03-30

Publication: CN111462769A·2020-07-28

Published as: CN111462769A

End-to-end accent conversion method

Abstract

The invention discloses an end-to-end accent conversion method, and belongs to the technical field of voice processing. The method converts a non-standard accent into a standard accent, and can be used for converting the voice of a patient with pronunciation disorder into standard voice. An accent conversion system for realizing the accent conversion method comprises a voice recognition module, aspeaker encoder, a voice synthesis module and a neural network vocoder, wherein the voice recognition module is used for adjusting the acoustic characteristics of input non-standard accent into the signal parameters of a standard accent, wherein the signal parameters are only related to the speaking content of the non-standard accent; and inputting the signal parameters of the non-standard accentand the speaker vector into the voice synthesis module, and synthesizing the standard accent of the specific speaker through the neural network vocoder after the voice is processed by the voice synthesis module. The method has the advantages that in the conversion process, the non-standard accent can be converted into the standard accent without any guidance of standard accent reference audios, and the original tone of a speaker is kept.

IPR2026-00272
Krisp Technologies EX1002 Page 238

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

DESCRIPTION CN111462769A

An end-to-end approach to accent conversion

[0001]

technical field

[0002]

The invention relates to the technical field of speech processing, in particular to an end-to-end accent conversion method.

[0003]

Background technique

[0004]

The application of speech recognition technology is becoming more and more extensive. The current speech recognition library is basically based on the speech of the standard Mandarin. The accuracy rate of converting the standard Mandarin of the speaker into text is relatively high.

However, in real life, most people's pronunciation is not standard, and they carry some local accents more or less. In order to communicate with them better, it is necessary to turn non-native accents into authentic ones.

In addition, there are many patients with dysphonia (such as Chinese patients) who cannot communicate with others normally. Converting the patients' non-standard pronunciation into standard pronunciation is particularly important for patients' daily communication.

The traditional speech conversion method is to convert the source speaker identity of the native accent to the speaker identity of the non-native accent, that is, only the timbre is changed, while the basic content and pronunciation remain unchanged. The need to use authentic accents during the conversion phase hinders their practical application. Therefore, this is an urgent problem to be solved at present.

[0005]

SUMMARY OF THE INVENTION

[0006]

IPR2026-00272
Krisp Technologies EX1002 Page 239

The purpose of the present invention is to provide an end-to-end accent conversion method, so as to solve the troublesome problem that the existing speech recognition library needs to use the authentic accent to convert the non-native accent to the authentic accent.

[0007]

In order to solve the above technical problems, the technical solution of the present invention is: an end-to-end accent conversion method, comprising an accent conversion system for implementing the accent conversion method, the accent conversion system comprising a speech recognition module, a speaker encoder, a speech A synthesis module and a neural network vocoder, the speech recognition module is used to adjust the acoustic characteristics of the input non-native accent into signal parameters of the authentic accent, and the signal parameters are only related to the speech content of the non-native accent; The signal parameters of the non-native accent and the speaker vector are input to the speech synthesis module, and the speech processed by the speech synthesis module can finally synthesize the native accent of a specific speaker through a neural network vocoder.

[0008]

As a preferred solution of the present invention, the speaker encoder is a neural network framework that is scalable and can verify the speaker, and generates an acoustic frame from any length of input speech and then converts it into a fixed-dimensional speaker embedding vector. The speaker embedding vector is only related to the speaker.

[0009]

As a preferred solution of the present invention, the speech synthesis module uses the speaker embedding vector to train the speech synthesis model using mean square error loss (LTTS), so that it only generates native accents embedded in the corresponding timbre by the speaker.

[0010]

As a preferred solution of the present invention, the speech recognition module is used to adjust the acoustic characteristics of the non-native accent to the signal parameters of the native accent.

[0011]

As a preferred solution of the present invention, the neural network vocoder is WaveRNN network, LPCNet or WaveNet.

[0012]

As a preferred solution of the present invention, it comprises the following steps: a, collecting the voice information of a plurality of speakers and corresponding text information, for training the speaker encoder and the speech synthesis module; b, screening out the speakers of non-native accents as the target Speaker; c. The text samples and speech samples of the target speaker are used for the training of the speech recognition module.

[0013]

The beneficial effects of adopting the above technical solutions are: the end-to-end accent conversion method proposed by the present invention does not require any guidance from the reference audio of the native accent in the conversion process, that is, the non-native accent can be converted into a native accent, and the speaker can be maintained. the original sound.

[0014]

Description of drawings

[0015]

Fig. 1 is the model schematic diagram of the training stage of the present invention;

IPR2026-00272
Krisp Technologies EX1002 Page 240

[0016]

Fig. 2 is the model schematic diagram of the conversion stage of the present invention;

[0017]

3 is a schematic diagram of the average opinion score results of the present invention in the experimental stage;

[0018]

FIG. 4 is a schematic diagram of the test results of the present invention in the experimental stage.

[0019]

Detailed ways

[0020]

The specific embodiments of the present invention will be further described below with reference to the accompanying drawings.

It should be noted here that the descriptions of these embodiments are used to help the understanding of the present invention, but do not constitute a limitation of the present invention.

In addition, the technical features involved in the various embodiments of the present invention described below can be combined with each other as long as they do not conflict with each other.

[0021]

This embodiment proposes an end-to-end accent conversion method, including an accent conversion system that implements the accent conversion method. The accent conversion system includes a speech recognition module, a speaker encoder, a speech synthesis module, a neural network vocoder, and a speech recognition module. The module is used to adjust the input acoustic features of the non-native accent to the signal parameters of the native accent, and the signal parameters are only related to the speech content of the non-native accent; the signal parameters of the non-native accent and the speaker vector are input to the speech synthesis module, and the voice The speech processed by the synthesis module can finally synthesize the native accent of a specific speaker through a neural network vocoder.

The speech recognition module is used to adjust the acoustic features of the non-native accent to the signal parameters of the native accent.

[0022]

The accent conversion method of the present invention mainly comprises the following steps: collecting the speech information and corresponding text information of a plurality of speakers for training the speaker encoder and the speech synthesis module; b. screening out the non-native accent speakers as the target speakers ; c. The text samples and speech samples of the target speaker are used for the training of the speech recognition module.

[0023]

The speaker encoder is a scalable and speaker-verifiable neural network framework that generates variable-length acoustic features from input speech of arbitrary length and converts them into fixed-dimensional speaker embedding vectors.

The speaker encoder model used in this paper is a scalable and very accurate neural network framework for speaker verification that generates a fixed-dimensional speaker embedding vector from a series of acoustic frames computed from speech sentences of arbitrary length.

The method uses the speaker embedding vector to adjust the TTS model of the target speaker's reference speech signal, so that the generated speech has the speaker identity of the target speaker.

IPR2026-00272
Krisp Technologies EX1002 Page 241

The generalized end-to-end speaker verification loss-loss is optimized by training a speaker encoder such that utterance embeddings from the same speaker have high cosine similarity, while utterance embeddings from different speakers are far apart in space.

We expect the speaker encoder to learn a representation relevant to speech synthesis that captures the characteristics of non-native accent speakers that were not visible during training.

[0024]

The speech synthesis module uses the speaker embedding vector to train the speech synthesis model with the mean square error loss LTTS, so that it only generates native accents that are embedded by the speaker in the corresponding timbre.

The model is an attention-based encoder-decoder model to support multiple speakers.

As shown in Figure 1, the speaker embedding vectors computed by the speaker encoder for the desired target speaker are concatenated with the TTS decoder outputs at each time step, and the attention TTS decoder takes these as additional inputs to generate a mel spectrogram.

The TTS model is trained with a mean squared error (MSE) loss LTTS to generate only authentic accents whose identity/timbre is determined by the speaker embeddings.

We map text transcripts into phoneme sequences as input to the TTS model, as it has been shown that using phoneme sequences can speed up convergence and improve pronunciation of rare words and proper nouns.

[0025]

The multi-task ASR model is also connected with a fully connected transformation layer to compute a connector temporal classification loss LCTC.

We use an accent ASR model to learn accent-independent language representations from acoustic features.

The ASR model applies an end-to-end attention-based encoder-decoder framework.

Given a pair of audio and its phoneme transcriptions, phonetic features are computed from the audio and linguistic representations of the phoneme sequence, respectively, by a TTS decoder.

We add a fully connected (FC) conversion layer to the ASR encoder and compute a connector temporal classification (CTC) loss LCTC to stabilize the training process.

Since the training data of the ASR model contains accents, we concatenate the accent embeddings and acoustic features of each frame as the input of the ASR model, and add an accent classifier on top of the ASR encoder to make it more robust to accent recognition .

We assume that different accents are associated with different speakers.

This paper obtains the speaker's accent embedding by averaging all the speaker's embeddings.

The output of the accent classifier is used to compute the cross-entropy loss LACC.

An attention-based ASR decoder is used to predict phoneme labels and language representations in both streams.

The cross-entropy loss LCE is used for phoneme token prediction, and the MSE loss LTTSE is used to measure the linguistic difference between the TTS decoder output H1 and the ASR decoder output H1.

[0026]

The neural network vocoder is a neural network vocoder such as WaveRNN network, LPCNet or WaveNet.

In this paper, the WaveRNN network is preferably used as the neural network vocoder, and the open source Pytorch is used for training. Since the mel-spectrogram captures all relevant details required for high-quality speech synthesis, we only use mel-spectrograms from multiple speakers to train WaveRNN without adding any speaker embeddings.

[0027]

IPR2026-00272
Krisp Technologies EX1002 Page 242

In the training phase, as shown in Figure 3, we first train the speaker-encoder model.

Then, a multi-speaker TTS model is trained using only English native speech data in the case of loss LTTS. After that, the speech conversion (AC) technology ASR model is first pre-trained using speech data from multiple native speakers and one non-native speaker. The ASR model is then fine-tuned using only speech data from target speakers with non-native accents. In both stages, we use the lossy lidar in Equation 2 and train a WaveRNN using only the speech data of native accent users. In the accent conversion stage, as shown in Figure 4, acoustic features are first computed from non-native accents. The speaker encoder then receives the acoustic features and outputs a speaker embedding vector representing the identity of the non-native accent speaker. The accent embedding is the average embedding of speakers with non-native accents. We concatenate the accent embedding with the acoustic features of each frame, then feed it into the ASR model to generate the language representation H--l, and then the attention-based TTS decoder combines the language representation and speaker embedding to generate native accent sounds academic characteristics. Finally, we use the WaveRNN model to convert the acoustic features into time-domain waveforms, hoping to render the accent more naturally.

[0028]

The accent conversion method includes the following steps: a. collecting speech information of multiple speakers; b. screening out speakers with Hindi accents as target speakers; c. using a part of the speech samples of the target speakers for multiple speakers The TTS model and the neural network vocoder are trained, and another part of the target speaker's speech samples is used for the training of the multi-task ASR model.

[0029]

During the experimental phase, the speech samples we collected consisted of 109 speakers with 44 hours of clear speech.

In this study, we choose speaker p248 with a Hindi accent as the target speaker.

We consider someone without a Hindi accent to be a native English speaker, and then resample the audio to 22.05kHz. For the training of the TTS model and the WaveRNN model, we only use the speech data of 105 native speakers. The mel spectrogram has 80 channels and is calculated using a 50ms window width and 12.5ms frame offset. 1000 samples were randomly selected as the validation set, and the rest were used for training. In the training of the accent ASR model, speech data of native accents and p248 were used. A 40-channel mel spectrogram and its delta and delta features computed with a 25 ms window width and 10 ms displacement were used as acoustic features. The number of utterances used for training, validation, and testing in p248 are 326, 25, and 25, respectively. The SI-ASR model used to extract ppg was trained using the TIMIT dataset.

[0030]

As shown in Figure 1, it is the training phase of the method.

hs is the language representation, hl is the language representation of the TTS decoder, and Hl is the language representation of the ASR decoder.

[0031]

Figure 2. Transformation stages of research methods.

hs and Hl denote speaker and language representations, respectively. LTTSE is expressed as follows:

[0033]

where N is the number of training samples.

The ASR model is trained with multi-task losses: LASR=λ1LCE+λ2LTTSE+λ3LCTC+λ4LACC, where λ is a hyperparameter, weighting four losses.

[0034]

The LTTS speaker encoder is a 3-layer LSTM with 256 hidden nodes followed by a projection layer of 256 units.

IPR2026-00272
Krisp Technologies EX1002 Page 243

The output is the second-level normalized hidden state of the last layer, which is a 256-element vector.

The TTS model adopts the same architecture as in [. The ASR encoder is a 5-layer bidirectional LSTM (BLSTM) with 320 units in each direction. 300-dimensional position-aware attention is used in the attention layer. The ASR decoder is a single-layer LSTM with 320 units. The accent classifier is a 2-layer 1D convolutional network with 128 channels and 3 kernel sizes, followed by an average pooling layer and a final FC output layer. In Equation 2, $\lambda 1 = 0.5$, $\lambda 2 = 0.1$, $\lambda 3 = 0.5$ and $\lambda 4 = 0.1$, heuristically makes the four loss terms at similar numerical scales. Transform 1 and Transform 2 in the baseline method are 2-layer and 4-layer BLSTMs, respectively, with 128 units in each direction.

[0035]

The speaker encoder model was trained for 1000k steps using the Adam optimizer with a batch size of 640 and a learning rate of 0.0001.

The TTS model was trained for 100k steps using the Adam optimizer with a batch size of 16 and a learning rate of 0.001. The accent ASR model is first pre-trained for 160k steps using the Adadelta optimizer with a batch size of 16 and a learning rate on speech data obtained from native speakers and p248. Then, while keeping the batch size and learning rate constant, fine-tuning is performed for another 5k steps on speech data from p248 only.

[0036]

As shown in Figure 3, (a) is the mean opinion score result with 95% confidence interval.

(b) is the test result of speaker phasic preference. AB-BL, P-BL, and P-AB represent the comparison of the ablation method v.s baseline method, the proposed method v.s baseline method, and the proposed method v.s ablation method, respectively.

[0037]

As shown in Figure 4, it is the result of the accent preference test.

"AB-BL", "P-BL", "P-AB", "P-L2" and "P-L1" denote "ablation method v.s baseline method", "proposed method v.s baseline method", "proposed method", respectively v.s ablation method", "Proposed method v.s non-native accent recordings" and "Proposed method v.s native accent recordings" comparison.

[0038]

Three perceptual listening tests were used to evaluate the transfer performance of the Baseline (BL), Proposed (P), and Ablation (AB) systems: Mean Opinion Score (MOS) test for audio naturalness, speaker similarity XAB test, and accent AB test.

We randomly sample 20 utterances from the test utterances of p248 speakers for evaluation.

[0039]

Audio naturalness.

In the MOS test, the naturalness of the audio is evaluated on a five-point scale (1 point difference, 2 points difference, 3 points difference, 4 points good, 5 points excellent). Audio produced by the three systems as well as reference recordings of non-native accents ("L2-Ref") were randomly sequenced prior to presentation to listeners. Each set of audio corresponds to the same text content. Listeners can listen to the audio repeatedly as needed. This method yields 4.0 MOS (4.4 for L2-Refreceives-MOS), which is statistically significantly higher than the baseline method. The maximum likelihood of this method is slightly higher than that of the ablation system, which illustrates that adding accent embeddings and classifiers helps to extract accent-independent linguistic content from accents that is beneficial for speech synthesis. Since we used seq2seq-based ASR and TTS models, the transformed speech has a more native accent-like pronunciation pattern, such as in duration and speech rate, which is very different from the source accent speech. We recommend that readers listen to the audio samples.

[0040]

Speaker similarity.

IPR2026-00272
Krisp Technologies EX1002 Page 244

We compare the differences in speaker similarity between the converted speech and non-native accents among the three systems. In the XAB test, X represents a non-native accent reference sample. We are given pairs of speech samples (A and B) with the same textual content as the reference, and ask the listener to determine which speech sample has a timbre closer to the reference. Listeners can also replay audio samples, and can choose "No Preference (NP)" if the difference cannot be distinguished. To avoid the influence of potential content, the audio is played in reverse. We can see that the baseline system has better similarity performance than the proposed system. Since the speech synthesis model in this method never found speech data from non-native accents, the results are reasonable. We expect the synthetic model to infer the timbre of the speaker from the speaker embeddings generated by the speaker encoder with only one utterance (ie, a speech clone). Access to speech data from more speakers helps train more general speaker encoders. Very good speech cloning performance can be obtained by training the speaker encoder with speech data from 18K speakers, however, we do not have access to such a large corpus of data. We found no statistically significant difference between the two ablation systems (p value 0.36). In the accent AB test, we first asked participants to listen to reference audio in native and non-native accents. Pairs of speech samples (A and B) with the same text content are then given, and the listener is asked to choose the sample that is more like the native accent speech. The results are shown in Figure 4. According to the preference test between "P-BL" and "P-AB", the listeners were very confident that the proposed method was able to produce more native accents than the baseline and ablation methods ($P<0.001$), even with ablation training Better accent performance than baseline treatment ($p<0.001$). The DTW process in the baseline method may introduce alignment errors, and the mapping from L2-PPGs to L1-PPGs using neural networks may be ineffective. According to the results of "P-L2" and "P-L1", we can draw the conclusion that this method can remove non-native accents from the speech of the second language, making the converted speech similar to the native accent in accent . The p-values were 2.3 × e-8 and 0.06, respectively.

[0041]

This paper presents an end-to-end accent conversion method, the first model capable of converting non-native accents to native accents without any guidance from native accent reference audio during the conversion process.

It consists of four independently trained neural networks: speaker encoder, multi-speaker TTS model, multi-task ASR model, neural network vocoder. The experimental results show that this method can convert non-native accent English speech into English speech with high naturalness and no accent, which is indistinguishable from native accent in accent. We expect the synthetic model to generate the desired target speaker timbre from the speaker embeddings obtained from the speaker encoder. This article takes English accent conversion as an example, the same framework can be used for accent conversion in any other language.

[0042]

In addition, an end-to-end accent conversion method proposed in this paper can also be used for pronunciation correction in patients with dysphonia due to brain injury, stroke, Parkinson's disease or amyotrophic lateral sclerosis.

Accent correction is especially important for patients' daily communication. The implementation method is the same as the above steps, only the non-native accent data needs to be replaced with the patient's non-standard accent data, and other training methods and conversion methods are the same as the above non-native accent conversion methods.

[0043]

The embodiments of the present invention have been described in detail above with reference to the accompanying drawings, but the present invention is not limited to the described embodiments.

For those skilled in the art, without departing from the principle and spirit of the present invention, various changes, modifications, substitutions and alterations to these embodiments still fall within the protection scope of the present invention.

IPR2026-00272
Krisp Technologies EX1002 Page 245

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

CLAIMS CN111462769A

1.

An end-to-end accent conversion method, characterized by comprising an accent conversion system for implementing the accent conversion method, the accent conversion system comprising a speech recognition module, a speaker encoder, a speech synthesis module, and a neural network vocoder, The speech recognition module is used to adjust the acoustic characteristics of the input non-native accent to the signal parameters of the native accent, and the signal parameters are only related to the speech content of the non-native accent; the signal parameters of the non-native accent and speaking The human vector is input to the speech synthesis module, and the speech processed by the speech synthesis module can finally be synthesized into the native accent of a specific speaker through a neural network vocoder.

2.

The end-to-end accent conversion method according to claim 1, wherein the speaker encoder is a neural network framework that is scalable and can verify the speaker, and generates an acoustic frame from any length of input speech and then converts it into A fixed-dimensional speaker embedding vector that is only relevant to the speaker.

3.

The end-to-end accent conversion method according to claim 2, wherein the speech synthesis module uses the speaker embedding vector to train the speech synthesis model by using mean square error loss (LTTS), so that it only generates data generated by the speaker. Embed the native accent of the corresponding tone.

4.

The end-to-end accent conversion method according to claim 1, wherein the speech recognition module is used to adjust the acoustic features of the non-native accent to the signal parameters of the authentic accent, and in the training process, the signal parameters of the authentic accent The extraction comes from the non-native accents of multiple speakers in the speech synthesis module, and the signal parameters of the native accent are only related to the speech content and have nothing to do with the speakers.

5.

The end-to-end accent conversion method according to claim 1, wherein the neural network vocoder is WaveRNN network, LPCNet or WaveNet.

6.

IPR2026-00272
Krisp Technologies EX1002 Page 246

The end-to-end accent conversion method according to claim 1, is characterized in that, comprises the following steps: A, collect the speech information of multiple speakers and corresponding text information, be used for training speaker encoder and speech synthesis module; B , Screen out non-native accent speakers as target speakers; c. Text samples and speech samples of target speakers are used for the training of the speech recognition module.

IPR2026-00272
Krisp Technologies EX1002 Page 247

(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111462769 A
(43)申请公布日 2020.07.28

(21)申请号 202010239586.2

(22)申请日 2020.03.30

(71)申请人 深圳市声希科技有限公司
地址 518101 广东省深圳市宝安区新安街道兴东社区留仙大道2号汇聚创新园2栋2710

(72)发明人 刘颂湘 王迪松 曹悦雯 孙立发 吴锡欣 康世胤 吴志勇 刘循英 蒙美玲

(74)专利代理机构 北京联瑞联丰知识产权代理事务所(普通合伙) 11411
代理人 郭蔚

(51)Int.Cl.
G10L 21/013(2013.01)
G10L 19/00(2013.01)
G10L 19/16(2013.01)
G10L 25/24(2013.01)
G10L 25/30(2013.01)
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 13/02(2013.01)
G10L 13/04(2013.01)

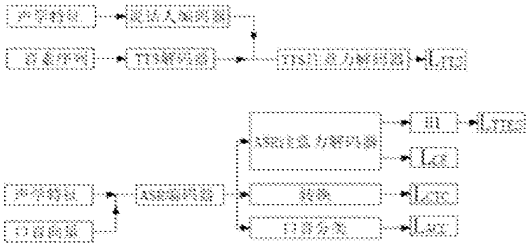
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种端到端的口音转换方法

(57)摘要

本发明公开了一种端到端的口音转换方法，将非地道口音转换为地道口音，属于语音处理技术领域，也可用于将发音障碍的患者的语音转换为标准语音，包括实现口音转换方法的口音转换系统，口音转换系统包括语音识别模块、说话人编码器、语音合成模块、神经网络声码器，语音识别模块用于将输入的非地道口音的声学特征调整为地道口音的信号参数，信号参数仅与非地道口音的说话内容相关；非地道口音的信号参数和说话人向量输入至语音合成模块，经语音合成模块处理过的语音最后通过神经网络声码器即可合成特定说话人的地道口音；有益效果是：在转换过程中无需地道口音参考音频的任何指导，即能够将非地道口音转换为地道口音，并且保持说话人的原本音色。



CN 111462769 A

IPR2026-00272
Krisp Technologies EX1002 Page 248

1.一种端到端的口音转换方法，其特征在于，包括实现所述口音转换方法的口音转换系统，所述口音转换系统包括语音识别模块、说话人编码器、语音合成模块、神经网络声码器，所述语音识别模块用于将输入的非地道口音的声学特征调整为地道口音的信号参数，所述信号参数仅与所述非地道口音的说话内容相关；所述非地道口音的信号参数和说话人向量输入至所述语音合成模块，经所述语音合成模块处理过的语音最后通过神经网络声码器即可合成特定说话人的地道口音。

2.根据权利要求1所述的端到端的口音转换方法，其特征在于，所述说话人编码器是一个可扩展且能验证说话者的神经网络框架，将任意长度的输入语音生成声学帧再转换成固定维的说话者嵌入向量，所述说话者嵌入向量仅与说话人相关。

3.根据权利要求2所述的端到端的口音转换方法，其特征在于，所述语音合成模块将所述说话者嵌入向量利用均方误差损失L_{TTS}对语音合成模型进行训练，使其只生成由说话者嵌入对应音色的地道口音。

4.根据权利要求1所述的端到端的口音转换方法，其特征在于，所述语音识别模块用于将非地道口音的声学特征调整为地道口音的信号参数，在训练过程中，地道口音的信号参数的提取来自于语音合成模块中的多个说话人的非地道口音，地道口音的信号参数仅仅和语音内容相关，与说话人无关。

5.根据权利要求1所述的端到端的口音转换方法，其特征在于，所述神经网络声码器为WaveRNN网络、LPCNet或WaveNet。

6.根据权利要求1所述的端到端的口音转换方法，其特征在于，包括以下步骤：a、采集多位说话者的语音信息以及对应文本信息，用于训练说话人编码器和语音合成模块；b、筛选出非地道口音的说话者作为目标说话人；c、目标说话人的文本样本和语音样本用于语音识别模块的训练。

IPR2026-00272
Krisp Technologies EX1002 Page 249

一种端到端的口音转换方法

技术领域

[0001] 本发明涉及语音处理技术领域，具体涉及一种端到端的口音转换方法。

背景技术

[0002] 语音识别技术应用越来越广泛，现在的语音识别库基本都是基于标准国语的语音而言，将说话人的标准国语转换为文字，准确率比较高。但是，现实生活中大部分人的发音都不标准，或多或少都携带一些地方口音，为了与他们更好的交流，需要将非地道口音变成地道口音。另外，目前有很多的发音障碍患者(比如中文患者)无法与其他人正常交流，通过将患者的不标准的发音转换为标准发音，对于患者的日常交流尤为重要。传统的语音转换方法是将地道口音的源说话人身份转换为非地道口音的说话人身份，也就是仅仅改变了音色，而基本内容和发音保持不变。由于在转换阶段需要使用到地道口音，这就阻碍了它们在实际生活中的应用。因此，这是目前亟待解决的问题。

发明内容

[0003] 本发明的目的在于提供一种端到端的口音转换方法，以解决现有的语音识别库将非地道口音转换为地道口音需要使用到地道口音的麻烦问题。

[0004] 为了解决上述技术问题，本发明的技术方案为：一种端到端的口音转换方法，包括实现所述口音转换方法的口音转换系统，所述口音转换系统包括语音识别模块、说话人编码器、语音合成模块、神经网络声码器，所述语音识别模块用于将输入的非地道口音的声学特征调整为地道口音的信号参数，所述信号参数仅与所述非地道口音的说话内容相关；所述非地道口音的信号参数和说话人向量输入至所述语音合成模块，经所述语音合成模块处理过的语音最后通过神经网络声码器即可合成特定说话人的地道口音。

[0005] 作为本发明的优选方案，所述说话人编码器是一个可扩展且能验证说话者的神经网络框架，将任意长度的输入语音生成声学帧再转换成固定维的说话者嵌入向量，所述说话者嵌入向量仅与说话人相关。

[0006] 作为本发明的优选方案，所述语音合成模块将所述说话者嵌入向量利用均方误差损失L_{TTS}对语音合成模型进行训练，使其只生成由说话者嵌入对应音色的地道口音。

[0007] 作为本发明的优选方案，所述语音识别模块用于将非地道口音的声学特征调整为地道口音的信号参数。

[0008] 作为本发明的优选方案，所述神经网络声码器为WaveRNN网络、LPCNet或WaveNet。

[0009] 作为本发明的优选方案，包括以下步骤：a、采集多位说话者的语音信息以及对应文本信息，用于训练说话人编码器和语音合成模块；b、筛选出非地道口音的说话者作为目标说话人；c、目标说话人的文本样本和语音样本用于语音识别模块的训练。

[0010] 采用上述技术方案的有益效果是：本发明提出的一种端到端的口音转换方法，在转换过程中无需地道口音参考音频的任何指导，即能够将非地道口音转换为地道口音，并且保持说话人的原本音色。

IPR2026-00272
Krisp Technologies EX1002 Page 250

IPR2026-00272
Krisp Technologies EX1002 Page 251

附图说明

[0011] 图1为本发明的训练阶段的模型示意图；

[0012] 图2为本发明的转换阶段的模型示意图；

[0013] 图3为本发明在实验阶段的平均意见得分结果示意图；

[0014] 图4为本发明在实验阶段的测试结果示意图。

具体实施方式

[0015] 下面结合附图对本发明的具体实施方式作进一步说明。在此需要说明的是，对于这些实施方式的说明用于帮助理解本发明，但并不构成对本发明的限定。此外，下面所描述的本发明各个实施方式中所涉及的技术特征只要彼此之间未构成冲突就可以相互组合。

[0016] 本实施例提出一种端到端的口音转换方法，包括实现所述口音转换方法的口音转换系统，口音转换系统包括语音识别模块、说话人编码器、语音合成模块、神经网络声码器，语音识别模块用于将输入的非地道口音的声学特征调整为地道口音的信号参数，信号参数仅与非地道口音的说话内容相关；非地道口音的信号参数和说话人向量输入至语音合成模块，经语音合成模块处理过的语音最后通过神经网络声码器即可合成特定说话人的地道口音。语音识别模块用于将非地道口音的声学特征调整为地道口音的信号参数。

[0017] 本发明的口音转换方法主要包括以下步骤：采集多位说话者的语音信息以及对应文本信息，用于训练说话人编码器和语音合成模块；b、筛选出非地道口音的说话者作为目标说话人； c、目标说话人的文本样本和语音样本用于语音识别模块的训练。

[0018] 说话人编码器是一个可扩展且能验证说话者的神经网络框架，将任意长度的输入语音生成可变长度的声学特征，再转换成固定维的说话者嵌入向量。本文使用的说话人编码器模型是一个可扩展且非常精确的说话者验证神经网络框架，可从由任意长度的语音语句计算出的一系列声学帧中生成一个固定维的说话者嵌入向量。该方法利用说话者嵌入向量对目标说话者的参考语音信号进行TTS模型调整，使生成的语音具有目标说话者的说话人身份。通过训练说话者编码器来优化广义的端到端说话者验证损-失，使得来自同一说话者的话语嵌入具有很高的共正弦相似度，而来自不同说话者的话语嵌入空间相差很远。我们期望说话者编码器学习与语音合成相关的表示，该表示法可捕捉在训练过程中不可见的非地道口音者的特征。

[0019] 语音合成模块将说话者嵌入向量利用均方误差损失L_{TTS}对语音合成模型进行训练，使其只生成由说话者嵌入对应音色的地道口音。该模型是一个基于注意力的编码器-解码器模型，以可支持多个说话人。如图1所示，由所需目标说话者的说话者编码器计算的说话者嵌入向量在每个时间步处与TTS解码器输出相连接，带注意力的TTS解码器将这些作为额外的输入来生成梅尔频谱图。利用均方误差(MSE)损失L_{TTS}对TTS模型进行训练，使其只生成由说话者嵌入决定身份/音色的地道口音。我们将文本转录本映射到音素序列中，作为TTS模型的输入，因为已经表明：使用音素序列可以加快收敛速度，改善稀有词和专有名词的发音效果。

[0020] 多任务ASR模型还连接有一个全连接转换层，用于计算一个连接器时间分类损失L_{CTC}。我们使用口音ASR模型从声学特征中学习与口音无关的语言表示。ASR模型应用了基于端到端的注意力编码器-解码器框架。在给定一对音频及其音素转录的情况下，通过TTS解

码器分别从音素序列的音频和语言表示中计算语音特征。我们在ASR编码器上增加了一个全连接(FC) 转换层，并计算了一个连接器时间分类(CTC)损失L_{CTC}，以稳定训练过程。由于ASR模型的训练数据包含口音，因此我们将每个帧的口音嵌入和声学特征连接起来作为ASR模型的输入，并在ASR编码器的顶部添加口音分类器，使其对于口音的识别更为稳健。我们假设不同的口音与不同的说话者有关。本文通过对说话者的所有嵌入进行平均，得到了说话者的口音嵌入。口音分类器的输出用于计算交叉熵损失L_{ACC}。采用基于注意力的ASR译码器对两个流中的音素标签和语言表示进行预判。交叉熵损失L_{CE}用于音素标记预测，MSE损失L_{TTSE}用于测量TTS 解码器输出H1和ASR解码器输出H1之间的语言差异。

[0021] 神经网络声码器为WaveRNN网络、LPCNet或WaveNet等神经网络声码器。本文优选采用WaveRNN网络作为神经网络声码器，使用开源Pytorch实现训练。由于梅尔频谱图捕获了高质量语音合成所需的所有相关细节，因此我们只使用来自多个说话者的梅尔频谱图来训练 WaveRNN，而不添加任何说话者嵌入。

[0022] 在训练阶段，如图3所示，我们首先训练说话者编码器模型。然后，在损失L_{TTS}的情况下，仅使用英语母语语音数据训练多说话者TTS模型。之后，语音转换(AC)技术ASR模型首先使用来自多个母语使用者和一个非母语使用者的语音数据进行预训练。然后，ASR模型仅使用来自非母语口音目标说话者的语音数据进行微调。在这两个阶段中，我们使用公式2中的损耗激光雷达，且只使用地道口音使用者的语音数据训练一个WaveRNN。在口音转换阶段，如图4所示，声学特征首先从非地道口音计算。然后，说话者编码器接收声学特征并输出表示非地道口音说话人身份的说话人嵌入向量。口音嵌入是非地道口音的说话人的平均嵌入。我们将口音嵌入与每帧的声学特征连接起来，然后将其输入到ASR模型中生成语言表示H--1，然后基于注意力的TTS解码器将语言表示和说话人嵌入结合起来生成地道口音声学特征。最后，我们使用WaveRNN模型将声学特征转换为时域波形，希望能更自然地呈现出口音。

[0023] 该口音转换方法包括以下步骤：a、采集多位说话者的语音信息；b、筛选出印地语口音的说话者作为目标说话人；c、将一部分目标说话人的语音样本用于多说话者TTS模型和神经网络声码器训练，另一部分目标说话人的语音样本用于多任务ASR模型的训练。

[0024] 在实验阶段，我们采集的语音样本是包含了109位说话者长达44小时的清晰讲话。在本研究中，我们选择印地语口音的说话者p248作为目标说话人。我们认为没有印地语口音的人即是以英语为母语的人，然后将音频重新采样至22.05kHz。对于TTS模型和WaveRNN模型的训练，我们只使用105位母语者的语音数据。梅尔频谱图有80个通道，使用50ms窗口宽度和12.5ms帧偏移计算。随机抽取1000个样本作为验证集，其余样本用于训练。在口音ASR 模型的训练中，使用了地道口音者和p248的语音数据。用25毫秒窗宽和10毫秒位移计算的 40道梅尔频谱图及其delta和delta特征作为声学特征。p248中用于训练、验证及测试的话语数分别为326、25和25。用于提取ppg的SI-ASR模型使用TIMIT数据集进行训练。

[0025] 如图1所示，为该方法的训练阶段。hs为语言表示，h1是TTS解码器的语言表示，H1是ASR解码器的语言表示。

[0026] 图2研究方法的转换阶段。hs和H1分别表示说话者和语言表征。L_{TTSE}表示如下：

IPR2026-00272
Krisp Technologies EX1002 Page 252

[0027] $$L_{TTSE}=\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{H}_i-\widehat{\mathbf{H}}_i\|_2,$$

[0028] 其中N是训练样本数。ASR模型是在多任务损失的情况下进行的训练：$L_{ASR}=\lambda_1 L_{CE}+\lambda_2 L_{TTSE}+\lambda_3 L_{CTC}+\lambda_4 L_{ACC}$，其中，λ是超参数，加权四个损失。

[0029] L_{TTS}扬声器编码器是一个3层LSTM，有256个隐藏节点，后面是256个单元的投影层。输出是最后一层的二级规范化隐藏状态，它是256个元素的向量。TTS模型采用了与[中相同的架构。ASR编码器是一个5层双向LSTM(BLSTM)，每个方向320个单元。注意层中使用300 维位置感知注意。ASR解码器是单层LSTM，有320个单元。口音分类器是一个2层的1维卷积网络，有128个通道和3个内核大小，然后是一个平均池化层和一个最终的FC输出层。在式2中，λ1＝0.5，λ2＝0.1，λ3＝0.5和λ4＝0.1，启发式地使四个损失项处于相似的数值尺度。基线方法中的变换1和变换2分别是2层和4层BLSTMs，每个方向有128个单位。

[0030] 使用Adam优化器对说话者编码器模型进行了1000k步的训练，批处理大小为640，学习率为0.0001。使用Adam优化器对TTS模型进行100k步的训练，批处理大小为16，学习率为 0.001。口音ASR模型首先使用Adadelta优化器对160k个步骤进行预训练，该优化器使用16 的批处理大小和从本族语者和p248获得的语音数据的学习率。然后，在保持批量大小和学习率不变的情况下，对仅来自p248的语音数据再进行5k步的微调。

[0031] 如图3所示，(a)为95％置信区间的平均意见得分结果。(b)为说话者相度性偏好测试结果，AB-BL、P-BL、P-AB分别代表消融方法v.s基线方法、提议方法v.s基线方法、提议方法v.s消融方法的比较。

[0032] 如图4所示，为口音偏好测试结果。“AB-BL”、“P-BL”、“P-AB”、“P-L2”及“P-L1”分别表示“消融方法v.s基线方法”、“提议方法v.s基线方法”、“提议方法 v.s消融方法”、“提议方法v.s非地道口音录音”和“提议方法v.s地道口音录音”的比较。

[0033] 采用三种感知听力测试来评价基线方法(BL)、提议方法(P)及消融方法(AB)系统的转换性能：音频自然度的平均意见得分(MOS)测试、说话人相似性XAB测试及口音AB测试。我们从p248说话人的测试话语中随机抽取20个话语进行评价。

[0034] 音频自然度。在MOS测试中，音频自然度按五分制(1分差、2分差、3分差、4分好、5分优)评定。三个系统产生的音频以及非地道口音的参考录音(“L2-Ref”)在呈现给听众之前被随机排序。每组音频对应相同的文本内容。听者可以根据需要重复听音频。该方法得到4.0个MOS(L2-Refreceives-MOS为4.4)，统计上显著高于基线方法。该方法的最大似然比消融术系统略高，这说明了添加口音嵌入和分类器有助于从口音中提取有利于语音合成的口音无关语言内容。由于我们使用了基于seq2seq的ASR和TTS模型，转换后的语音具有更像地道口音的发音模式，比如在持续时间和语速方面，这与源口音的语音非常不同。我们建议读者收听音频样本。

[0035] 说话者相似度。我们比较了这三种系统在转换语音和非地道口音的说话者相似度方面的差异。在XAB测试中，X表示非地道口音参考样本。我们给出了与参考语具有相同文本内容的成对语音样本(A和B)，并要求听者确定哪个语音样本的音色更接近参考语。监听器也可以重播音频样本，如果无法区分差异，则可以选择“无首选项(NP)”。为了避免潜在内容的影响，音频是反向播放的。我们可以看到，基线系统比所提出的系统具有更好的相似度性

IPR2026-00272
Krisp Technologies EX1002 Page 253

能。由于该方法中的语音合成模型从未发现来自非地道口音者的语音数据，因此结果是合理的。我们期望合成模型能从只有一条语音(即语音克隆)的说话者编码器生成的说话人嵌入中推断出说话人的音色。来自更多说话人的语音数据访问有助于训练更通用的说话人编码器。通过使用来自18K说话者的语音数据训练说话者编码器，能够获得非常好的语音克隆性能，但是，我们无法访问如此大的数据语料库。我们发现两种消融系统之间在统计学上无差异(p值为0.36)。在口音AB测试中，我们首先让参与者听地道口音和非地道口音的参考音频。然后给出具有相同文本内容的成对语音样本(A和B)，并要求听者选择更像地道口音语音的样本。结果如图4所示。根据"P-BL"和"P-AB"之间的偏好测试，听者非常确信所提出的方法能够产生比基线和消融方法更多的地道口音 ($P<0.001$)，即使消融训练也能获得比基线治疗更好的口音表现($p<0.001$)。基线方法中的DTW过程可能引入对准误差，使用神经网络从L2-PPGs到L1-PPGs的映射可能无效。根据"P-L2"和"P-L1"的结果，我们可以得出这样的结论：该方法可以从第二语言的语音中去除非地道口音，使转换后的语音在口音上与地道口音相似。p值分别为$2.3\times e-8$和0.06。

[0036] 本文提出了一种端到端的口音转换方法，这是第一个在转换过程中无需地道口音参考音频的任何指导，即能够将非地道口音转换为地道口音的模型。它由四个独立训练的神经网络组成：说话人编码器、多说话者TTS模型、多任务ASR模型、神经网络声码器。实验结果表明：该方法能将非地道口音的英语语音转换成语音自然度高、无口音的英语语音，在口音上与地道口音难以区分。我们期望综合模型能从从说话者编码器得到的说话人嵌入中生成所需的目标说话者的音色。本文以英语口音转换为例，相同的框架可以用于其他任何语种的口音转换。

[0037] 此外，本文提出的一种端到端的口音转换方法，也可用于发音障碍患者的发音修正，造成发音障碍的原因包括脑部受伤，中风，帕金森氏病或肌萎缩性侧索硬化等。口音修正对于患者的日常交流尤为重要。其实施方法与上述步骤一致，只需要将非地道口音数据替换为患者的非标准口音数据，其他训练方法和转换方法与上述非地道口音的转换方法相同。

[0038] 以上结合附图对本发明的实施方式作了详细说明，但本发明不限于所描述的实施方式。对于本领域的技术人员而言，在不脱离本发明原理和精神的情况下，对这些实施方式进行多种变化、修改、替换和变型，仍落入本发明的保护范围内。

IPR2026-00272
Krisp Technologies EX1002 Page 254

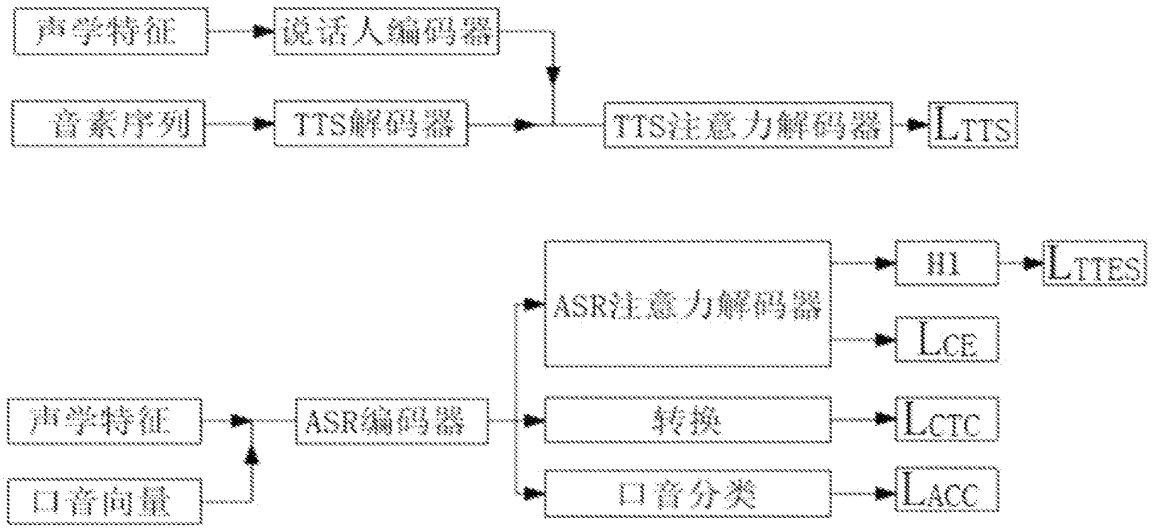


图1

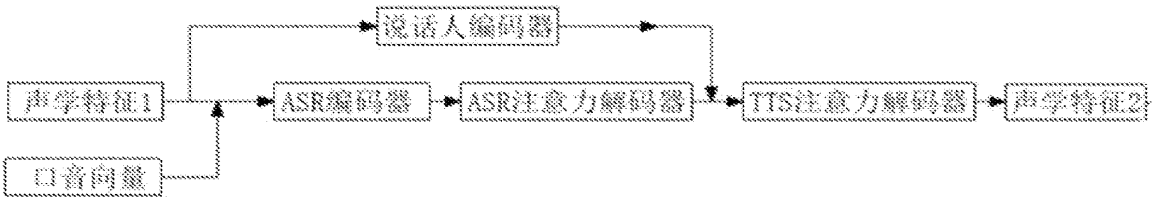


图2

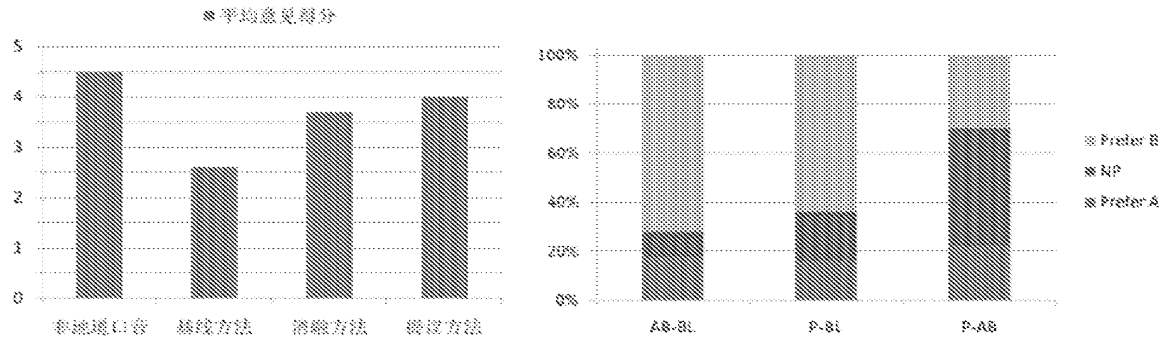


图3

IPR2026-00272
Krisp Technologies EX1002 Page 255

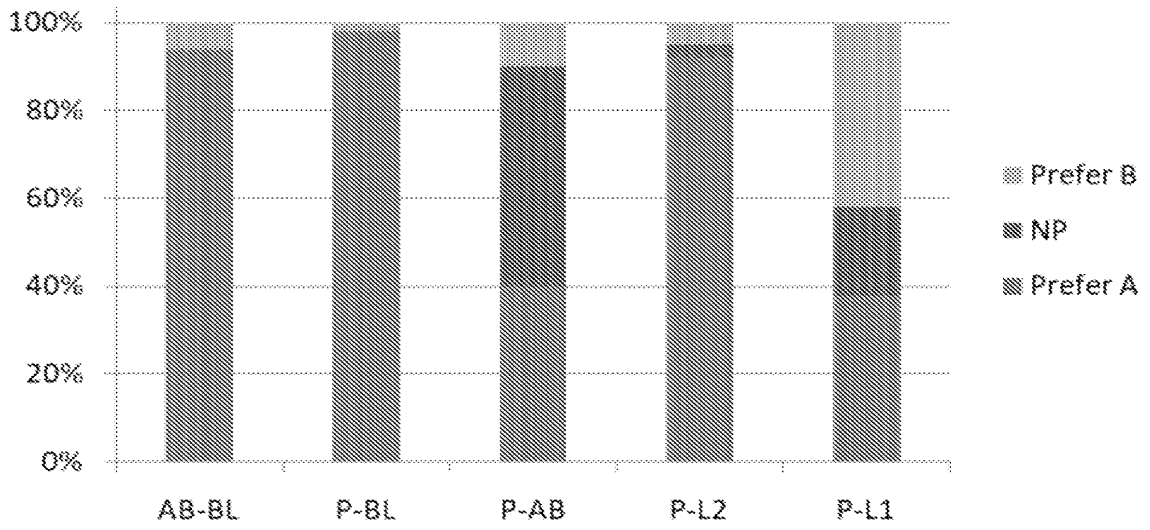


图4

IPR2026-00272
Krisp Technologies EX1002 Page 256

Espacenet

CN112382267A Method, device and equipment for converting accent and storage medium

Applicants: BEIJING YOUZHUJU NETWORK TECH CO LTD

Inventors: TANG BENLAI,LI ZHONGHAO

Classifications:
IPC ***G10L13/02; G10L13/04;***

CPC **G10L13/02 (CN); G10L13/04 (CN);**

Priorities: CN202011270379A 2020-11-13

Application: CN202011270379A·2020-11-13

Publication: CN112382267A·2021-02-19

Published as: CN112382267A

Method, device and equipment for converting accent and storage medium

Abstract

The invention discloses a method, a device and equipment for converting accent and a storage medium, and relates to the fields of speech synthesis, natural language processing, computer technology, artificial intelligence and deep learning. According to the specific implementation scheme, the method comprises the steps of acquiring a target speaking text; determining first acoustic features for indicating a target accent based on the target speaking text; determining linguistic features in the first acoustic features; and based on the linguistic features and the target speaking text, determining an audio with the target accent corresponding to the target speaking text, and outputting the audio. According to the implementation mode, the acquired speaking text data of a user is subjected tovoice synthesis, recognition and conversion, so that any text input by the user can be accurately and quickly converted into an audio, with the target accent, of the user.

IPR2026-00272
Krisp Technologies EX1002 Page 257

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

DESCRIPTION CN112382267A

Method, apparatus, device and storage medium for converting accents

[0001]

technical field

[0002]

The present application relates to the field of computer technology, in particular to the fields of speech synthesis, natural language processing, computer technology, artificial intelligence, and deep learning, and in particular, to a method, apparatus, device, and storage medium for converting accents.

[0003]

Background technique

[0004]

In recent years, due to the rapid development of online education and online learning, accent conversion technology has received extensive research and attention. Accent conversion aims to convert the target user's accent into a target accent.

Likewise, accent conversion technology has great application prospects in entertainment.

Accent conversion using existing accent conversion technology is not only slow but also the results of accent conversion are often inaccurate.

[0005]

SUMMARY OF THE INVENTION

[0006]

The present disclosure provides a method, apparatus, device, and storage medium for converting accents.

[0007]

According to an aspect of the present disclosure, there is provided a method for converting an accent, comprising: obtaining a target speaking text; determining a first acoustic feature indicating the target accent based on the target speaking text; Based on

IPR2026-00272
Krisp Technologies EX1002 Page 258

the linguistic features and the target speaking text, determine the audio with the target accent corresponding to the target speaking text, and output the audio.

[0008]

According to another aspect of the present disclosure, there is provided an apparatus for converting an accent, comprising: an acquisition unit configured to acquire a target utterance text; and a first acoustic feature determination unit configured to determine a target utterance text based on the target utterance text a first acoustic feature indicative of a target accent; a linguistic feature determination unit configured to determine a linguistic feature in the first acoustic feature; a conversion unit configured to determine the target utterance based on the linguistic feature and the target utterance text Audio with the target accent corresponding to the text, and output the audio.

[0009]

According to yet another aspect of the present disclosure, there is provided an electronic device for converting an accent, comprising: at least one processor; and a memory communicatively connected with the at least one processor; wherein the memory stores a memory executable executable by the at least one processor. The instructions are executed by the at least one processor to enable the at least one processor to perform the method for converting an accent as described above.

[0010]

According to yet another aspect of the present disclosure, there is provided a non-transitory computer-readable storage medium storing computer instructions for causing a computer to perform the method for converting an accent as described above.

[0011]

The technology according to the present application solves the problem that accent conversion cannot be performed accurately and quickly. By synthesizing, recognizing, and converting the acquired user's speech text data through speech synthesis, it is possible to realize that any text input by the user can be accurately and quickly Convert to audio with the target accent for this user.

[0012]

It should be understood that what is described in this section is not intended to identify key or critical features of embodiments of the disclosure, nor is it intended to limit the scope of the disclosure.

Other features of the present disclosure will become readily understood from the following description.

[0013]

Description of drawings

[0014]

The accompanying drawings are used for better understanding of the present solution, and do not constitute a limitation to the present application.

in:

[0015]

FIG. 1 is an exemplary system architecture diagram to which an embodiment of the present application may be applied;

[0016]

Figure 2 is a flowchart of one embodiment of a method for converting accents according to the present application;

IPR2026-00272
Krisp Technologies EX1002 Page 259

[0017]

3 is a schematic diagram of an application scenario of the method for converting accents according to the present application;

[0018]

Figure 4 is a flowchart of another embodiment of a method for converting accents according to the present application;

[0019]

5 is a schematic structural diagram of an embodiment of an apparatus for converting accents according to the present application;

[0020]

FIG. 6 is a block diagram of an electronic device used to implement the method for converting an accent according to an embodiment of the present application.

[0021]

Detailed ways

[0022]

Exemplary embodiments of the present application are described below with reference to the accompanying drawings, which include various details of the embodiments of the present application to facilitate understanding, and should be considered as exemplary only.

Accordingly, those of ordinary skill in the art will recognize that various changes and modifications of the embodiments described herein can be made without departing from the scope and spirit of the present application.

Also, descriptions of well-known functions and constructions are omitted from the following description for clarity and conciseness.

[0023]

It should be noted that the embodiments in the present application and the features of the embodiments may be combined with each other in the case of no conflict.

The present application will be described in detail below with reference to the accompanying drawings and in conjunction with the embodiments.

[0024]

FIG. 1 illustrates an exemplary system architecture 100 to which embodiments of the method for converting an accent or an apparatus for converting an accent of the present application may be applied.

[0025]

As shown in FIG. 1 , the system architecture 100 may include terminal devices 101 , 102 , and 103 , a network 104 and a server 105 .

The network 104 is a medium used to provide a communication link between the terminal devices 101 , 102 , 103 and the server 105 .

The network 104 may include various connection types, such as wired, wireless communication links, or fiber optic cables, among others.

IPR2026-00272
Krisp Technologies EX1002 Page 260

[0026]

The user can use the terminal devices 101, 102, 103 to interact with the server 105 through the network 104 to receive or send messages and the like.

Various communication client applications, such as speech synthesis applications, may be installed on the terminal devices 101 , 102 and 103 .

[0027]

The terminal devices 101, 102, and 103 may be hardware or software.

When the terminal devices 101, 102, and 103 are hardware, they can be various electronic devices, including but not limited to smart phones, tablet computers, in-vehicle computers, laptop computers, desktop computers, and the like.

When the terminal devices 101, 102, and 103 are software, they can be installed in the electronic devices listed above.

It can be implemented as a plurality of software or software modules, and can also be implemented as a single software or software module.

There is no specific limitation here.

[0028]

The server 105 may be a server that provides various services, such as a background server that processes the target speech text collected by the terminal devices 101 , 102 , and 103 .

The background server may acquire the target speech text, and based on the target speech text, determine a first acoustic feature for indicating the target accent; determine the linguistic feature in the first acoustic feature; determine the target based on the linguistic feature and the target speech text Audio with the target accent corresponding to the spoken text, and output the audio.

[0029]

It should be noted that the server 105 may be hardware or software.

When the server 105 is hardware, it can be implemented as a distributed server cluster composed of multiple servers, or can be implemented as a single server.

When the server 105 is software, it may be implemented as a plurality of software or software modules, or may be implemented as a single software or software module.

There is no specific limitation here.

[0030]

It should be noted that, the method for converting accents provided by the embodiments of the present application is generally executed by the server 105 .

Accordingly, means for converting accents are generally provided in the server 105 .

[0031]

It should be understood that the numbers of terminal devices, networks and servers in FIG. 1 are merely illustrative.

There can be any number of terminal devices, networks and servers according to implementation needs.

[0032]

With continued reference to FIG. 2, a flow 200 of one embodiment of a method for converting accents according to the present application is shown.

IPR2026-00272
Krisp Technologies EX1002 Page 261

The method for converting accents of the present embodiment includes the following steps:

[0033]

Step 201, acquiring the target speech text.

[0034]

In this embodiment, the execution body of the method for converting accents (for example, the server 105 in FIG. 1) can obtain the target speech text collected by the terminal device through recording or scanning through wired connection or wireless connection.

Specifically, the target spoken text may be the spoken text converted from the speech uttered by any person with a definite identity who wants to convert the accent.

Spoken text, which can be text.

For example, if Xiao Ming said "I love the motherland" in Mandarin, the target speech text can be expressed as "I love the motherland" in written form.

[0035]

Step 202 , based on the target spoken text, determine a first acoustic feature for indicating the target accent.

[0036]

After acquiring the target speech text, the execution subject may determine the first acoustic feature for indicating the target accent based on the target speech text.

Specifically, the first acoustic feature may be a speech feature parameter used to characterize the target accent, such as a Mel spectrum.

In this embodiment, the execution subject may pre-emphasize, divide and window the audio based on the target speaking text and the corresponding audio, and then perform short-time Fourier transform on the audio signal of each frame (STFT, short-time Fourier transform, or short-term Fourier transform), and obtain the short-term amplitude spectrum; the short-term amplitude spectrum obtains the Mel spectrum through the Mel filter bank.

The present application does not specifically limit the acquisition method of the Mel spectrum (ie, the first acoustic feature).

[0037]

Step 203, determining linguistic features in the first acoustic features.

[0038]

After determining the first acoustic feature, the executing subject may determine the linguistic feature in the first acoustic feature.

Specifically, linguistic features may include prosodic features, syntax, discourse structure, information structure, and the like.

Among them, the prosodic feature can also be a supersonic feature or a supersegmental feature, which is a phonological structure of a language.

The prosodic features can be divided into three main aspects: intonation, temporal distribution and stress, which are achieved by suprasegmental features.

Suprasegmental features include pitch, intensity, and temporal properties, carried by phonemes or groups of phonemes.

Prosody is a typical feature of human natural language, and has many common features across languages, such as pitch dip, stress, pause, etc., which are commonly found in different languages.

Prosodic features are one of the important forms of language and emotional expression.

IPR2026-00272
Krisp Technologies EX1002 Page 262

The execution subject may compare the existing known linguistic features with each feature in the first acoustic feature, and determine a feature in the first acoustic feature whose similarity with the known linguistic feature is greater than a threshold value as the first acoustic feature. Linguistic features in acoustic features.

Specifically, the execution body may convert each feature in the first acoustic feature into a first acoustic feature vector, and convert an existing known linguistic feature into a linguistic feature vector.

Calculate the cosine similarity between each first acoustic feature vector and each linguistic feature vector. The closer the cosine similarity is to 1, it indicates that the first acoustic feature vector and the corresponding linguistic feature vector participate in the calculation of the cosine similarity. The more similar, the first acoustic feature vector may be determined as a linguistic feature vector, and accordingly, the first acoustic feature corresponding to the first acoustic feature vector is determined as the linguistic feature corresponding to the linguistic feature vector.

In the same way, traverse each first acoustic feature and each linguistic feature until the absolute value of the difference between the cosine similarity and 1 of the corresponding vectors of each first acoustic feature and each linguistic feature is found to be less than The first acoustic feature corresponding to the first acoustic feature vector of the threshold value is determined, and the first acoustic feature is determined as a linguistic feature, according to which all linguistic features in the first acoustic feature can be obtained.

[0039]

Step 204 , based on the linguistic feature and the target spoken text, determine the audio with the target accent corresponding to the target spoken text, and output the audio.

[0040]

After determining the linguistic feature, the executing subject may determine the audio with the target accent corresponding to the target speaking text based on the linguistic feature and the target speaking text, and output the audio.

Specifically, the execution subject may input the linguistic features and the target spoken text into the pre-trained speech model to generate audio with the target accent corresponding to the target spoken text, and output the audio.

The pre-trained speech model is used to represent linguistic features and the correspondence between the target speech text and the audio with the target accent, and may be a pre-trained neural network model. When training the speech model, an initial neural network model can be obtained first; a training sample set can be obtained, wherein the training samples in the training sample set include linguistic features, target speech text corresponding to the linguistic features, and marked and linguistic features. The audio with the target accent corresponding to the target speaking text corresponding to the linguistic features; the linguistic features of the training samples in the training sample set and the target speaking text corresponding to the linguistic features are used as the input of the initial neural network model, and the input The linguistic feature and the audio with the target accent corresponding to the target speech text corresponding to the linguistic feature are used as the expected output to train the initial neural network model; the trained initial neural network model is determined as the above speech model. For example, the target speech text may be the text "I love the motherland" in Mandarin, and the audio corresponding to the target speech text with the target accent may be the audio of "I love the motherland" in the Northeast accent, and the audio may be in the form of MP3 or it is in the form of MP4, and this application does not specifically limit the form of the output audio.

[0041]

Continue to refer to FIG. 3 , which shows a schematic diagram of an application scenario of the method for converting accents according to the present application.

In the application scenario of FIG. 3 , the server 303 obtains the target speech text 301 through the network 302 . Based on the target spoken text 301, the server 303 determines a first acoustic feature 304 indicative of the target accent. The server 303 determines the linguistic features 305 in the first acoustic features 304 . Based on the linguistic feature 305 and the target spoken text 301 , the server 303 determines the audio 306 with the target accent corresponding to the target spoken text 301 , and outputs the audio 306 . This application is actually the conversion of spoken text data to audio with a specific accent. The text content does not change.

[0042]

IPR2026-00272
Krisp Technologies EX1002 Page 263

In this embodiment, by synthesizing, recognizing, and converting the acquired speech text data of the user, any text input by the user can be accurately and quickly converted into audio with the target accent of the user.

[0043]

With continued reference to FIG. 4, a flow 400 of another embodiment of a method for converting accents according to the present application is shown.

As shown in FIG. 4 , the method for converting accents in this embodiment may include the following steps:

[0044]

Step 401, acquiring the target speech text.

[0045]

Step 402 , based on the target spoken text, determine a first acoustic feature for indicating the target accent.

[0046]

The principles of steps 401 to 402 are similar to those of steps 201 to 202, and are not repeated here.

[0047]

Specifically, step 402 can be implemented through step 4021:

[0048]

Step 4021, according to the target speech text and the pre-trained speech synthesis model, determine the first acoustic feature corresponding to the target speech text and used to indicate the target accent.

[0049]

In this embodiment, the pre-trained speech synthesis model is used to represent the correspondence between the spoken text and the first acoustic feature.

After obtaining the target speech text, the execution subject can determine the first acoustic feature corresponding to the target speech text and used to indicate the target accent according to the target speech text and the pre-trained speech synthesis model.

Specifically, the first acoustic feature may be a Mel spectrum used to indicate a target accent, such as a northeastern accent.

The first acoustic feature corresponding to the target speech text obtained by the speech synthesis model and used to indicate the target accent cannot be used for the output of the target accent at this time, and further feature recognition and feature conversion are required to obtain the target speech text with more target speech information features, and then further speech synthesis is performed according to the obtained features that carry more target speech text information.

In this embodiment, specifically, the execution subject may input the target speech text into a pre-trained speech synthesis model, and the pre-trained speech synthesis model extracts the first acoustic feature of the target speech text to obtain the corresponding A first acoustic feature for indicating the target accent.

[0050]

In this embodiment, the audio mel spectrum with the target accent can be obtained according to the target speech text and the pre-trained speech synthesis model, so that the speech text of any user can be accurately analyzed based on the obtained audio mel spectrum with the target accent. Conversion of the target accent, resulting in audio with the target accent.

[0051]

IPR2026-00272
Krisp Technologies EX1002 Page 264

Step 403, determining linguistic features in the first acoustic features.

[0052]

The principle of step 403 is similar to that of step 203, and details are not repeated here.

[0053]

Specifically, step 403 can be implemented through step 4031:

[0054]

Step 4031, using the pre-trained recognition model to extract linguistic features in the first acoustic features.

[0055]

The pre-trained recognition model is used to represent the correspondence between the first acoustic feature and the linguistic feature.

After obtaining the first acoustic feature, the executing subject may extract the linguistic feature in the first acoustic feature by using the pre-trained recognition model.

Specifically, the linguistic features in the first acoustic feature may include prosody features such as intonation, temporal distribution, stress, pitch, stress, pause, and the like.

Linguistic features are one of the important forms of language and emotional expression.

[0056]

In some optional implementations of this embodiment, the execution body may also determine a category identifier corresponding to the first acoustic feature according to the first acoustic feature and a pre-trained identification model, where the pre-trained identification model is used for The correspondence between the first acoustic feature and the class identifier is characterized.

The obtained category identifier may be an identifier used to characterize the category of each phoneme in the first acoustic feature, for example, each phoneme in the first acoustic feature may be a tone phoneme, a time-domain distribution phoneme, an accent phoneme, a pitch phoneme, Stressed phonemes and paused phonemes, for example, can be represented by signs 1, 2, 3, 4, 5, 6, and 7, respectively.

Then, the execution main body may determine, according to each phoneme in the first acoustic feature corresponding to each obtained logo and the preset corresponding relationship between the logo, the phoneme and the second acoustic feature, the method for generating the audio with the target accent. The second acoustic signature.

The second acoustic feature may be a Mel spectrum corresponding to each phoneme required for generating the target accent.

The executing subject may determine audio with the target accent corresponding to the target spoken text according to the second acoustic feature, and output the audio.

This implementation can enrich the mel spectrum required to generate the audio of the target accent, and improve the accuracy of generating the audio of the target accent.

[0057]

In this embodiment, by using the pre-trained recognition model to extract the linguistic features in the first acoustic features, the relevant features of the corresponding language and emotional expression in the target speech text can be extracted, so as to improve the audio features used to generate the target accent , which improves the accuracy of generating audio with the target accent.

[0058]

IPR2026-00272
Krisp Technologies EX1002 Page 265

Step 404 , based on the linguistic feature and the target speaking text, determine the audio with the target accent corresponding to the target speaking text, and output the audio.

[0059]

The principle of step 404 is similar to that of step 204, and details are not repeated here.

[0060]

Specifically, step 404 can be implemented through steps 4041 to 4042:

[0061]

Step 4041 , according to the linguistic features, the target spoken text and the pre-trained conversion model, determine the second acoustic feature corresponding to the target spoken text and used to generate the audio with the target accent.

[0062]

In this embodiment, the transformation model is used to characterize the linguistic feature and the correspondence between the target spoken text and the second acoustic feature used to generate the audio with the target accent.

Specifically, the second acoustic feature may include the Mel spectrum corresponding to the above-mentioned linguistic feature.

Specifically, the second acoustic feature may be a Mel spectrum corresponding to each phoneme required for generating the target accent.

The execution body may input the linguistic feature and the target speech text into the pre-trained conversion model, and output the second acoustic feature corresponding to the target speech text for generating audio with the target accent.

[0063]

Step 4042, based on the second acoustic feature, determine the audio with the target accent.

[0064]

After obtaining the second acoustic feature, the executing subject may determine the audio with the target accent based on the second acoustic feature.

Specifically, the execution subject may input the obtained second acoustic feature into the vocoder, and obtain audio with the target accent through the conversion of the vocoder.

The vocoder encodes and encrypts the second acoustic feature at its transmitting end to obtain a match with the channel, transmits it to the receiving end of the vocoder through the information channel, analyzes the received feature in the frequency domain, and identifies the clear signal. For voiced sound, the fundamental frequency of the voiced sound is measured, and then the unvoiced-voiced judgment, the fundamental frequency of the voiced sound and the spectral envelope are selected as characteristic parameters to transmit.

Of course, the analysis can also be performed in the time domain, and some second acoustic features are periodically extracted using it for linear prediction, and audio with the target accent corresponding to the second acoustic features is generated.

Specifically, the above-mentioned vocoder may include a channel vocoder, a formant vocoder, a pattern vocoder, a linear prediction vocoder, a correlation vocoder and an orthogonal function vocoder. The type is not specifically limited.

[0065]

In this embodiment, by determining the second acoustic feature corresponding to the target speaking text according to the linguistic feature, the target speaking text and the pre-trained conversion model, and determining the audio with the target accent based on the second acoustic feature, it is possible to make the audio with the target accent for any user. For any spoken text, more accurate audio of the target accent can be obtained.

IPR2026-00272
Krisp Technologies EX1002 Page 266

[0066]

In some optional implementations of this embodiment, the method for converting accents further includes the following model training steps not shown in FIG. 4 : acquiring an initial neural network model; acquiring a training sample set, wherein the training sample set is The training samples of the The target speech text corresponding to the features and linguistic features is used as the input of the initial neural network model, and the second acoustic feature corresponding to the target speech text corresponding to the input linguistic features and linguistic features is used as the expected output. Train; determine the initial neural network model after training as the transformation model.

[0067]

In this embodiment, the execution body may acquire the initial neural network model through wired connection or wireless connection.

The above-mentioned initial neural network model may include various artificial neural networks (Artificial Neural Network, ANN) including hidden layers.

In this embodiment, the above-mentioned execution body may also acquire a pre-stored initial model locally, or may acquire the above-mentioned initial model from a communication-connected electronic device, which is not limited herein.

[0068]

In this embodiment, the execution body may acquire the training sample set in various ways.

Specifically, the training samples in the training sample set may include linguistic features, target speaking text corresponding to the linguistic features, and annotated second acoustic features corresponding to the linguistic features and the target speaking text corresponding to the linguistic features.

The second acoustic feature marked in the training sample and corresponding to the linguistic feature and the target speech text corresponding to the linguistic feature may be obtained from a local or communicatively connected electronic device through a wired or wireless connection, or it may be it can be manually marked in real time, or it can be obtained after automatic marking and manual supplementation and modification to correct marking errors, which is not specifically limited in this application.

Linguistic features in training samples can be extracted in real time, or acquired from local or communicatively connected electronic devices through wired or wireless connections.

The target speech texts corresponding to linguistic features in the training samples may be collected in real time from different users, or may be obtained from local or communication-connected electronic devices through wired or wireless connections. Make specific restrictions.

[0069]

In this embodiment, by acquiring the training sample set to train the initial neural network model, a conversion model capable of accurately generating audio with a target accent according to linguistic features and speaking text can be obtained, so that the acquired target speaking text can be translated into The text is more accurately converted into audio with the target accent, and can improve the quality of the resulting audio with the target accent.

[0070]

In some optional implementations of this embodiment, the audio with the target accent includes at least one of the following: speaking audio with the target accent, and singing audio with the target accent.

[0071]

In some optional implementations of this embodiment, the second acoustic feature includes a Mel spectrum corresponding to the target accent; and determining the audio with the target accent based on the second acoustic feature includes: according to the Mel spectrum and a preset A neural network vocoder that synthesizes audio with the target accent.

[0072]

IPR2026-00272
Krisp Technologies EX1002 Page 267

Specifically, after obtaining the Mel spectrum corresponding to the target accent, the executing subject can automatically input the Mel spectrum into a preset neural network vocoder, so as to synthesize the mel spectrum based on the neural network vocoder and the Mel spectrum. Audio for the target accent.

[0073]

In this implementation manner, the synthesized audio with the target accent can be made more accurate by using the Mel spectrum corresponding to the target accent and the preset neural network vocoder.

[0074]

Referring further to FIG. 5 , as an implementation of the methods shown in the above figures, the present application provides an embodiment of an apparatus for converting accents. The apparatus embodiment corresponds to the method embodiment shown in FIG. 2 . The device can be specifically applied to various electronic devices.

[0075]

As shown in FIG. 5 , the apparatus 500 for converting accents in this embodiment includes: an acquiring unit 501 , a first acoustic feature determining unit 502 , a linguistic feature determining unit 503 , and a converting unit 504 .

[0076]

The obtaining unit 501 is configured to obtain the target speech text.

[0077]

The first acoustic feature determination unit 502 is configured to determine a first acoustic feature indicative of the target accent based on the target spoken text.

[0078]

The linguistic feature determining unit 503 is configured to determine linguistic features in the first acoustic features.

[0079]

The converting unit 504 is configured to determine the audio with the target accent corresponding to the target speaking text based on the linguistic feature and the target speaking text, and output the audio.

[0080]

In some optional implementations of this embodiment, the first acoustic feature determining unit 502 is further configured to: determine, according to the target speaking text and the pre-trained speech synthesis model, a signal corresponding to the target speaking text and used to indicate the target accent The first acoustic feature, wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0081]

In some optional implementations of this embodiment, the linguistic feature determining unit 503 is further configured to: extract linguistic features in the first acoustic feature by using a pre-trained recognition model, wherein the pre-trained recognition model uses it is used to characterize the correspondence between the first acoustic feature and the linguistic feature.

[0082]

IPR2026-00272
Krisp Technologies EX1002 Page 268

In some optional implementations of this embodiment, the conversion unit 504 is further configured to: according to the linguistic feature, the target speech text and the pre-trained conversion model, determine the corresponding target speech text for generating the audio with the target accent The second acoustic feature of , wherein the transformation model is used to characterize the linguistic feature and the correspondence between the target spoken text and the second acoustic feature used to generate the audio with the target accent; based on the second acoustic feature, determine the audio with the target accent .

[0083]

In some optional implementations of this embodiment, the apparatus further includes a training unit not shown in FIG. 5, configured to: acquire an initial neural network model; acquire a training sample set, wherein the training samples in the training sample set including linguistic features, the target speaking text corresponding to the linguistic features, and the marked second acoustic feature corresponding to the target speaking text corresponding to the linguistic features and the linguistic features; the linguistic features and language of the training samples in the training sample set The target speech text corresponding to the linguistic feature is used as the input of the initial neural network model, and the second acoustic feature corresponding to the input linguistic feature and the target speech text corresponding to the linguistic feature is used as the expected output to train the initial neural network model; The initial neural network model after training is determined as the transformation model.

[0084]

In some optional implementations of this embodiment, the audio with the target accent includes at least one of the following: speaking audio with the target accent, and singing audio with the target accent.

[0085]

In some optional implementations of this embodiment, the second acoustic feature includes a mel spectrum corresponding to the target accent; and the converting unit 504 is further configured to: according to the mel spectrum and a preset neural network vocoder, Synthesize audio with the target accent.

[0086]

It should be understood that the units 501 to 504 recorded in the apparatus 500 for converting accents correspond to respective steps in the method described with reference to FIG. 2 .

Therefore, the operations and features described above with respect to the method for converting accents are also applicable to the apparatus 500 and the units contained therein, and will not be repeated here.

[0087]

According to the embodiments of the present application, the present application further provides an electronic device and a readable storage medium.

[0088]

As shown in FIG. 6 , it is a block diagram of an electronic device for a method for converting accents according to an embodiment of the present application.

Electronic devices are intended to represent various forms of digital computers, such as laptops, desktops, workstations, personal digital assistants, servers, blade servers, mainframe computers, and other suitable computers.

Electronic devices may also represent various forms of mobile devices, such as personal digital processors, cellular phones, smart phones, wearable devices, and other similar computing devices.

The components shown herein, their connections and relationships, and their functions are by way of example only, and are not intended to limit implementations of the application described and/or claimed herein.

[0089]

IPR2026-00272
Krisp Technologies EX1002 Page 269

As shown in FIG. 6, the electronic device includes: one or more processors 601, a memory 602, and interfaces for connecting various components, including a high-speed interface and a low-speed interface.

The various components are interconnected using different buses 605 and may be mounted on a common motherboard or otherwise as desired.

The processor may process instructions executed within the electronic device, including instructions stored in or on memory to display graphical information of the GUI on an external input/output device, such as a display device coupled to the interface.

In other embodiments, multiple processors and/or multiple buses 605 may be used with multiple memories and multiple memories, if desired.

Likewise, multiple electronic devices may be connected, each providing some of the necessary operations (eg, as a server array, a group of blade servers, or a multiprocessor system).

A processor 601 is taken as an example in FIG. 6 .

[0090]

The memory 602 is the non-transitory computer-readable storage medium provided by the present application.

Wherein, the memory stores instructions executable by at least one processor, so that the at least one processor executes the method for converting an accent provided by the present application.

The non-transitory computer-readable storage medium of the present application stores computer instructions for causing a computer to perform the method for converting an accent provided by the present application.

[0091]

As a non-transitory computer-readable storage medium, the memory 602 can be used to store non-transitory software programs, non-transitory computer-executable programs, and units, such as program instructions/units (For example, the acquisition unit 501, the first acoustic feature determination unit 502, the linguistic feature determination unit 503 and the conversion unit 504 shown in FIG. 5).

The processor 601 executes various functional applications and data processing of the server by running the non-transitory software programs, instructions and modules stored in the memory 602, that is, implementing the method for converting accents in the above method embodiments.

[0092]

The memory 602 may include a storage program area and a storage data area, wherein the storage program area may store an operating system, an application program required for at least one function; the storage data area may store data created according to the use of the electronic device for converting accents Wait.

Additionally, memory 602 may include high-speed random access memory, and may also include non-transitory memory, such as at least one magnetic disk storage device, flash memory device, or other non-transitory solid state storage device.

In some embodiments, the memory 602 may optionally include memory located remotely from the processor 601, and these remote memories may be connected via a network to the electronic device for converting accents.

Examples of such networks include, but are not limited to, the Internet, an intranet, a local area network, a mobile communication network, and combinations thereof.

[0093]

The electronic device for the method of converting accents may further include: an input device 603 and an output device 604 .

The processor 601, the memory 602, the input device 603, and the output device 604 may be connected through a bus 605 or other means, and the connection through the bus 605 is taken as an example in FIG. 6 .

[0094]

IPR2026-00272
Krisp Technologies EX1002 Page 270

The input device 603 can receive input numerical or character information, and generate key signal input related to user settings and function control of electronic equipment for converting accents, such as touch screen, keypad, mouse, trackpad, touchpad, pointing stick , one or more mouse buttons, trackballs, joysticks and other input devices.

Output devices 604 may include display devices, auxiliary lighting devices (eg, LEDs), haptic feedback devices (eg, vibration motors), and the like.

The display device may include, but is not limited to, a liquid crystal display (LCD), a light emitting diode (LED) display, and a plasma display.

In some implementations, the display device may be a touch screen.

[0095]

Various implementations of the systems and techniques described herein can be implemented in digital electronic circuitry, integrated circuit systems, application specific ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof.

These various embodiments may include being implemented in one or more computer programs executable and/or interpretable on a programmable system including at least one programmable processor that The processor, which may be a special purpose or general-purpose programmable processor, may receive data and instructions from a storage system, at least one input device, and at least one output device, and transmit data and instructions to the storage system, the at least one input device, and the at least one output device an output device.

[0096]

These computational programs (also referred to as programs, software, software applications, or codes) include machine instructions for programmable processors, and may be implemented using high-level procedural and/or object-oriented programming languages, and/or assembly/machine languages calculation program.

As used herein, the terms "machine-readable medium" and "computer-readable medium" refer to any computer program product, apparatus, and/or apparatus for providing machine instructions and/or data to a programmable processor (For example, magnetic disks, optical disks, memories, programmable logic devices (PLDs), including machine-readable media that receive machine instructions as machine-readable signals.

The term "machine-readable signal" refers to any signal used to provide machine instructions and/or data to a programmable processor.

[0097]

To provide interaction with a user, the systems and techniques described herein may be implemented on a computer having a display device (eg, a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user); and a keyboard and pointing device (eg, a mouse or trackball) through which a user can provide input to the computer.

Other kinds of devices can also be used to provide interaction with the user; for example, the feedback provided to the user can be any form of sensory feedback (eg, visual feedback, auditory feedback, or tactile feedback); and can be in any form (including acoustic input, voice input, or tactile input) to receive input from the user.

[0098]

The systems and techniques described herein may be implemented on a computing system that includes back-end components (eg, as a data server), or a computing system that includes middleware components (eg, an application server), or a computing system that includes front-end components (eg, a user's computer having a graphical user interface or web browser through which a user may interact with implementations of the systems and techniques described herein), or including such backend components, middleware components, Or any combination of front-end components in a computing system.

The components of the system may be interconnected by any form or medium of digital data communication (eg, a communication network).

Examples of communication networks include: Local Area Networks (LANs), Wide Area Networks (WANs), and the Internet.

IPR2026-00272
Krisp Technologies EX1002 Page 271

[0099]

A computer system can include clients and servers.

Clients and servers are generally remote from each other and usually interact through a communication network.

The relationship of client and server arises by computer programs running on the respective computers and having a client-server relationship to each other.

[0100]

According to the technical solutions of the embodiments of the present application, by synthesizing, recognizing, and converting the acquired speech text data of the user, any text input by the user can be accurately and quickly converted into the user's speech with the target accent. audio.

[0101]

According to one or more embodiments of the present disclosure, the present disclosure provides a method for converting an accent, comprising: obtaining a target speaking text; determining a first acoustic feature for indicating the target accent based on the target speaking text; determining Linguistic features in the first acoustic features; based on the linguistic features and the target speaking text, determine the audio with the target accent corresponding to the target speaking text, and output the audio.

[0102]

According to one or more embodiments of the present disclosure, determining the first acoustic feature used to indicate the target accent based on the target speech text includes: determining, according to the target speech text and a pre-trained speech synthesis model, an application corresponding to the target speech text. is used to indicate the first acoustic feature of the target accent, wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0103]

According to one or more embodiments of the present disclosure, determining the linguistic feature in the first acoustic feature includes: extracting the linguistic feature in the first acoustic feature by using a pretrained recognition model, wherein the pretrained recognition model Used to characterize the correspondence between the first acoustic feature and the linguistic feature.

[0104]

According to one or more embodiments of the present disclosure, determining the audio with the target accent corresponding to the target speaking text based on the linguistic features and the target speaking text includes: determining, according to the linguistic features, the target speaking text, and a pre-trained conversion model, determining a second acoustic feature corresponding to the target spoken text for generating audio with the target accent, wherein the transformation model is used to characterize the linguistic feature and the correspondence between the target speaking text and the second acoustic feature for generating the audio with the target accent ; determining audio with the target accent based on the second acoustic feature.

[0105]

According to one or more embodiments of the present disclosure, the method for converting accents further includes: acquiring an initial neural network model; acquiring a training sample set, wherein the training samples in the training sample set include linguistic features, The target speech text and the marked second acoustic feature corresponding to the target speech text corresponding to the linguistic feature and the linguistic feature; the linguistic feature of the training sample in the training sample set and the target speech text corresponding to the linguistic feature are used as the initial neural The input of the network model takes the second acoustic feature corresponding to the input linguistic feature and the target speech text corresponding to the linguistic feature as the expected output, and trains the initial neural network model; the trained initial neural network model is determined as the conversion Model.

IPR2026-00272
Krisp Technologies EX1002 Page 272

[0106]

According to one or more embodiments of the present disclosure, the audio with the target accent includes at least one of the following: speaking audio with the target accent, singing audio with the target accent.

[0107]

According to one or more embodiments of the present disclosure, the second acoustic feature includes a Mel spectrum corresponding to the target accent; and based on the second acoustic feature, determining the audio with the target accent includes: according to the Mel spectrum and a preset neural A network vocoder that synthesizes audio with targeted accents.

[0108]

According to one or more embodiments of the present disclosure, the present disclosure provides an apparatus for converting an accent, comprising: an acquisition unit configured to acquire a target spoken text; a first acoustic feature determination unit configured to be based on the target a spoken text, determining a first acoustic feature for indicating a target accent; a linguistic feature determining unit, configured to determine a linguistic feature among the first acoustic features; a conversion unit, configured to speak based on the linguistic feature and the target text, determine the audio with the target accent corresponding to the target spoken text, and output the audio.

[0109]

According to one or more embodiments of the present disclosure, the first acoustic feature determination unit is further configured to: determine, according to the target speech text and the pre-trained speech synthesis model, a first sound corresponding to the target speech text and used to indicate the target accent acoustic features, wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0110]

According to one or more embodiments of the present disclosure, the linguistic feature determination unit is further configured to: extract linguistic features in the first acoustic feature using a pretrained recognition model, wherein the pretrained recognition model is used to characterize the first acoustic feature. A correspondence between acoustic features and linguistic features.

[0111]

According to one or more embodiments of the present disclosure, the conversion unit is further configured to: determine, according to the linguistic feature, the target utterance text, and the pretrained conversion model, a second method for generating the audio with the target accent corresponding to the target utterance text Acoustic features, wherein the transformation model is used to characterize the linguistic features and the correspondence between the target spoken text and the second acoustic features used to generate the audio with the target accent; based on the second acoustic features, the audio with the target accent is determined.

[0112]

According to one or more embodiments of the present disclosure, the apparatus for converting accents further includes a training unit not shown in FIG. 5, configured to: obtain an initial neural network model; obtain a training sample set, wherein the training sample set is The training samples of the The target speech text corresponding to the features and linguistic features is used as the input of the initial neural network model, and the second acoustic feature corresponding to the target speech text corresponding to the input linguistic features and linguistic features is used as the expected output. Train; determine the initial neural network model after training as the transformation model.

[0113]

According to one or more embodiments of the present disclosure, the audio with the target accent includes at least one of the following: speaking audio with the target accent, singing audio with the target accent.

IPR2026-00272
Krisp Technologies EX1002 Page 273

[0114]

According to one or more embodiments of the present disclosure, the second acoustic feature includes a mel spectrum corresponding to the target accent; and the converting unit is further configured to: synthesize the target accent according to the mel spectrum and the preset neural network vocoder Accent audio.

[0115]

It should be understood that the above embodiments are only exemplary embodiments, but not limited thereto, and also include other methods known in the art that can be implemented for converting accents.

Steps can be reordered, added, or removed using the various forms of flow shown above.

For example, the steps described in the present application can be executed in parallel, sequentially or in different orders, as long as the desired results of the technical solutions disclosed in the present application can be achieved, no limitation is imposed herein.

[0116]

The above-mentioned specific embodiments do not constitute a limitation on the protection scope of the present application.

It should be understood by those skilled in the art that various modifications, combinations, sub-combinations and substitutions may occur depending on design requirements and other factors.

Any modifications, equivalent replacements and improvements made within the spirit and principles of this application shall be included within the protection scope of this application.

IPR2026-00272
Krisp Technologies EX1002 Page 274

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

CLAIMS CN112382267A

1.

A method for converting accents, including:

Get the target speech text;

determining a first acoustic feature indicative of a target accent based on the target spoken text;

determining linguistic features in the first acoustic features;

Based on the linguistic feature and the target spoken text, audio with a target accent corresponding to the target spoken text is determined, and the audio is output.

2.

The method of claim 1, wherein the determining, based on the target spoken text, a first acoustic feature indicative of a target accent comprises:

According to the target speech text and a pre-trained speech synthesis model, a first acoustic feature corresponding to the target speech text and used to indicate the target accent is determined, wherein the pre-trained speech synthesis model is used to characterize the speech text and Correspondence of the first acoustic feature.

3.

The method of claim 1, wherein said determining a linguistic feature in the first acoustic feature comprises:

The linguistic features in the first acoustic features are extracted by using a pre-trained recognition model, wherein the pre-trained recognition model is used to represent the correspondence between the first acoustic features and the linguistic features.

4.

The method according to claim 1, wherein the determining the audio with the target accent corresponding to the target speaking text based on the linguistic feature and the target speaking text comprises:

A second acoustic feature corresponding to the target spoken text for generating audio with the target accent is determined according to the linguistic feature, the target spoken text and the pre-trained conversion model, wherein the transformation model is used to represent the linguistic a correspondence between the feature and the target spoken text and the second acoustic feature used to generate the audio with the target accent;

Based on the second acoustic feature, audio having the target accent is determined.

IPR2026-00272
Krisp Technologies EX1002 Page 275

5.

The method of claim 4, wherein the method further comprises:

Get the initial neural network model;

Obtain a training sample set, wherein the training samples in the training sample set include linguistic features, target speech texts corresponding to the linguistic features, and marked target speech texts corresponding to the linguistic features and the linguistic features The second acoustic feature of ;

Taking the linguistic features of the training samples in the training sample set and the target speaking text corresponding to the linguistic features as the input of the initial neural network model, and corresponding to the target speaking text corresponding to the input linguistic features and linguistic features The second acoustic feature of is used as the expected output, and the initial neural network model is trained;

The trained initial neural network model is determined as the conversion model.

6.

The method according to any one of claims 1 to 5, wherein the audio with the target accent comprises at least one of the following: speaking audio with the target accent, singing audio with the target accent.

7.

5. The method of claim 4, wherein the second acoustic feature comprises a Mel spectrum corresponding to the target accent; and

The determining the audio with the target accent based on the second acoustic feature includes:

Based on the mel spectrum and a preset neural network vocoder, audio with the target accent is synthesized.

8.

A device for converting accents, comprising:

an acquisition unit, configured to acquire the target speech text;

a first acoustic feature determination unit configured to determine, based on the target spoken text, a first acoustic feature indicative of a target accent;

a linguistic feature determination unit configured to determine a linguistic feature in the first acoustic feature;

A conversion unit, configured to determine audio with a target accent corresponding to the target spoken text based on the linguistic feature and the target spoken text, and output the audio.

9.

An electronic device for converting accents, comprising:

at least one processor; and

a memory communicatively coupled to the at least one processor; wherein,

The memory stores instructions executable by the at least one processor, the instructions being executed by the at least one processor to enable the at least one processor to perform the execution of any of claims 1-7 Methods.

10.

A non-transitory computer-readable storage medium storing computer instructions, wherein the computer instructions are used to cause the computer to perform the method of any one of claims 1-7.

IPR2026-00272
Krisp Technologies EX1002 Page 276

IPR2026-00272
Krisp Technologies EX1002 Page 277

(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112382267 A
(43)申请公布日 2021.02.19

(21)申请号 202011270379.X

(22)申请日 2020.11.13

(71)申请人 北京有竹居网络技术有限公司
地址 101299 北京市大兴区林荫北街13号信息大厦802室

(72)发明人 汤本来 李忠豪

(74)专利代理机构 北京海智友知识产权代理事务所(普通合伙) 11455
代理人 吴京顺

(51)Int.Cl.
G10L 13/02(2013.01)
G10L 13/04(2013.01)

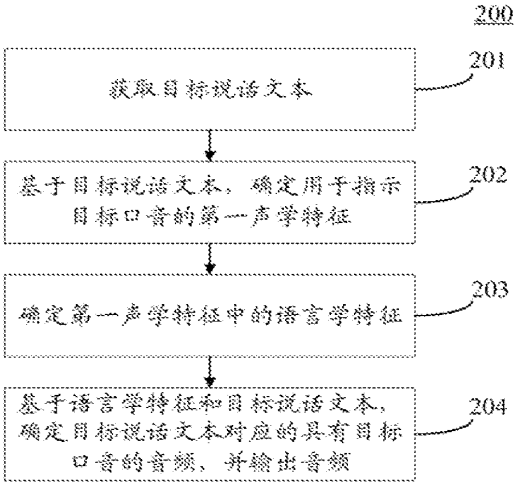
权利要求书2页 说明书11页 附图5页

(54)发明名称

用于转换口音的方法、装置、设备以及存储介质

(57)摘要

本申请公开了一种用于转换口音的方法、装置、设备以及存储介质，涉及语音合成、自然语言处理、计算机技术、人工智能、深度学习领域。具体实现方案为：获取目标说话文本；基于目标说话文本，确定用于指示目标口音的第一声学特征；确定第一声学特征中的语言学特征；基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。本实现方式通过将获取的用户的说话文本数据通过语音合成、识别、转换，可以实现对于用户输入的任意一个文本，能够准确、快速地转换成该用户的具有目标口音的音频。



CN 112382267 A

IPR2026-00272
Krisp Technologies EX1002 Page 278

1.一种用于转换口音的方法，包括：

获取目标说话文本；

基于所述目标说话文本，确定用于指示目标口音的第一声学特征；

确定所述第一声学特征中的语言学特征；

基于所述语言学特征和所述目标说话文本，确定所述目标说话文本对应的具有目标口音的音频，并输出所述音频。

2.根据权利要求1所述的方法，其中，所述基于所述目标说话文本，确定用于指示目标口音的第一声学特征，包括：

根据所述目标说话文本和预训练的语音合成模型，确定所述目标说话文本对应的用于指示目标口音的第一声学特征，其中，所述预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

3.根据权利要求1所述的方法，其中，所述确定所述第一声学特征中的语言学特征，包括：

利用预训练的识别模型提取所述第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

4.根据权利要求1所述的方法，其中，所述基于所述语言学特征和所述目标说话文本，确定所述目标说话文本对应的具有目标口音的音频，包括：

根据所述语言学特征、目标说话文本和预训练的转换模型，确定所述目标说话文本对应的用于生成具有目标口音的音频的第二声学特征，其中，所述转换模型用于表征语言学特征和目标说话文本与用于生成具有目标口音的音频的第二声学特征的对应关系；

基于所述第二声学特征，确定具有目标口音的音频。

5.根据权利要求4所述的方法，其中，所述方法还包括：

获取初始神经网络模型；

获取训练样本集合，其中，所述训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与所述语言学特征和所述语言学特征对应的目标说话文本对应的第二声学特征；

将所述训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为所述初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的第二声学特征作为期望输出，对所述初始神经网络模型进行训练；

将训练后的所述初始神经网络模型确定为所述转换模型。

6.根据权利要求1～5任一项所述的方法，其中，所述具有目标口音的音频，包括以下至少一项：具有目标口音的说话音频、具有目标口音的歌唱音频。

7.根据权利要求4所述的方法，其中，所述第二声学特征包括对应于所述目标口音的梅尔频谱；以及

所述基于所述第二声学特征，确定具有目标口音的音频，包括：

根据所述梅尔频谱和预设的神经网络声码器，合成具有目标口音的音频。

8.一种用于转换口音的装置，包括：

获取单元，被配置成获取目标说话文本；

第一声学特征确定单元，被配置成基于所述目标说话文本，确定用于指示目标口音的

IPR2026-00272
Krisp Technologies EX1002 Page 279

第一声学特征；

语言学特征确定单元，被配置成确定所述第一声学特征中的语言学特征；

转换单元，被配置成基于所述语言学特征和所述目标说话文本，确定所述目标说话文本对应的具有目标口音的音频，并输出所述音频。

9.一种用于转换口音的电子设备，其特征在于，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-7中任一项所述的方法。

10.一种存储有计算机指令的非瞬时计算机可读存储介质，其特征在于，所述计算机指令用于使所述计算机执行权利要求1-7中任一项所述的方法。

IPR2026-00272
Krisp Technologies EX1002 Page 280

用于转换口音的方法、装置、设备以及存储介质

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及语音合成、自然语言处理、计算机技术、人工智能、深度学习领域，尤其涉及用于转换口音的方法、装置、设备以及存储介质。

背景技术

[0002] 近年来，由于在线教育和在线学习的迅猛发展，口音转换技术得到了广泛的研究与关注，口音转换旨在于转换目标用户口音为目标口音。同样，口音转换技术在娱乐方面也有很大的应用前景。利用现有的口音转换技术进行口音转换不仅速度慢而且口音转换的结果往往不准确。

发明内容

[0003] 本公开提供了一种用于转换口音的方法、装置、设备以及存储介质。

[0004] 根据本公开的一方面，提供了一种用于转换口音的方法，包括：获取目标说话文本；基于目标说话文本，确定用于指示目标口音的第一声学特征；确定第一声学特征中的语言学特征；基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0005] 根据本公开的另一方面，提供了一种用于转换口音的装置，包括：获取单元，被配置成获取目标说话文本；第一声学特征确定单元，被配置成基于目标说话文本，确定用于指示目标口音的第一声学特征；语言学特征确定单元，被配置成确定第一声学特征中的语言学特征；转换单元，被配置成基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0006] 根据本公开的再一方面，提供了一种用于转换口音的电子设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行如上述用于转换口音的方法。

[0007] 根据本公开的又一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，上述计算机指令用于使计算机执行如上述用于转换口音的方法。

[0008] 根据本申请的技术解决了无法准确、快速地进行口音转换的问题，通过将获取的用户的说话文本数据通过语音合成、识别、转换，可以实现对于用户输入的任意一个文本，能够准确、快速地转换成该用户的具有目标口音的音频。

[0009] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0010] 附图用于更好地理解本方案，不构成对本申请的限定。其中：

[0011] 图1是本申请的一个实施例可以应用于其中的示例性系统架构图；

IPR2026-00272
Krisp Technologies EX1002 Page 281

[0012] 图2是根据本申请的用于转换口音的方法的一个实施例的流程图；

[0013] 图3是根据本申请的用于转换口音的方法的一个应用场景的示意图；

[0014] 图4是根据本申请的用于转换口音的方法的另一个实施例的流程图；

[0015] 图5是根据本申请的用于转换口音的装置的一个实施例的结构示意图；

[0016] 图6是用来实现本申请实施例的用于转换口音的方法的电子设备的框图。

具体实施方式

[0017] 以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0018] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0019] 图1示出了可以应用本申请的用于转换口音的方法或用于转换口音的装置的实施例的示例性系统架构100。

[0020] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0021] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如语音合成类应用等。

[0022] 终端设备101、102、103可以是硬件，也可以是软件。当终端设备101、102、103为硬件时，可以是各种电子设备，包括但不限于智能手机、平板电脑、车载电脑、膝上型便携计算机和台式计算机等等。当终端设备101、102、103为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

[0023] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103采集的目标说话文本进行处理的后台服务器。后台服务器可以获取目标说话文本，并基于目标说话文本，确定用于指示目标口音的第一声学特征；确定第一声学特征中的语言学特征；基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出该音频。

[0024] 需要说明的是，服务器105可以是硬件，也可以是软件。当服务器105为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器105为软件时，可以实现成多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

[0025] 需要说明的是，本申请实施例所提供的用于转换口音的方法一般由服务器105执行。相应地，用于转换口音的装置一般设置于服务器105中。

[0026] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0027] 继续参考图2，示出了根据本申请的用于转换口音的方法的一个实施例的流程

IPR2026-00272
Krisp Technologies EX1002 Page 282

200。本实施例的用于转换口音的方法，包括以下步骤：

[0028] 步骤201，获取目标说话文本。

[0029] 本实施例中，用于转换口音的方法的执行主体(例如图1中的服务器105)可以通过有线连接或无线连接的方式获取终端设备通过录音或扫描的方式采集的目标说话文本。具体地，目标说话文本可以是具有确定身份的任意一个想要转换口音的人所发出的语音转换成的说话文本。说话文本，可以是文字。比如，小明说了一句普通话“我爱祖国”，则目标说话文本就可以是以文字形式表示的“我爱祖国”。

[0030] 步骤202，基于目标说话文本，确定用于指示目标口音的第一声学特征。

[0031] 执行主体在获取目标说话文本后，可以基于目标说话文本，确定用于指示目标口音的第一声学特征。具体地，第一声学特征可以是用于表征目标口音的语音特征参数，例如可以是梅尔频谱。本实施例中，执行主体可以基于目标说话文本，根据目标说话文本和其对应的音频，对音频进行预加重、分帧和加窗，然后可以对每帧音频的信号进行短时傅里叶变换(STFT，short-time Fourier transform，或short-term Fourier transform)，得到短时幅度谱；短时幅度谱通过梅尔滤波器组得到梅尔频谱。本申请对梅尔频谱(即第一声学特征)的获取方式不做具体限定。

[0032] 步骤203，确定第一声学特征中的语言学特征。

[0033] 执行主体在确定第一声学特征后，可以确定第一声学特征中的语言学特征。具体地，语言学特征可以包括韵律特征、句法、语篇结构、信息结构等。其中，韵律特征又可以是超音质特征或超音段特征，是语言的一种音系结构。韵律特征可以分为三个主要方面：语调、时域分布和重音，通过超音段特征实现。超音段特征包括音高、强度以及时间特性，由音位或音位群负载。韵律是人类自然语言的一个典型特征，具有许多跨语言的共同特点，比如：音高下倾、重读、停顿等都普遍存在于不同的语言之中。韵律特征是语言和情绪表达的重要形式之一。执行主体可以用现有的已知的语言学特征跟第一声学特征中的各特征对比，将第一声学特征中与已知的语言学特征的相似度大于阈值的特征确定为第一声学特征中的语言学特征。具体地，执行主体可以分别将第一声学特征中的各特征转化为第一声学特征向量，将现有的已知的语言学特征转化为语言学特征向量。将各第一声学特征向量与各语言学特征向量进行余弦相似度计算，余弦相似度越趋近于1，表明参与计算该余弦相似度的第一声学特征向量和对应的语言学特征向量越相似，可以将该第一声学特征向量确定为语言学特征向量，相应地，将该第一声学特征向量对应的第一声学特征确定为该语言学特征向量对应的语言学特征。同理，遍历每个第一声学特征和每个语言学特征，直至找到每个第一声学特征和每个语言学特征各自对应的向量的余弦相似度与1的差值的绝对值小于阈值的第一声学特征向量对应的第一声学特征，并将该第一声学特征确定为语言学特征，据此可以得到第一声学特征中的所有语言学特征。

[0034] 步骤204，基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0035] 执行主体在确定出语言学特征后，可以基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。具体地，执行主体可以将语言学特征和目标说话文本输入预训练的语音模型以生成与目标说话文本对应的具有目标口音的音频，并输出该音频。其中，该预训练的语音模型用于表征语言学特征和目标说话文本与具有

IPR2026-00272
Krisp Technologies EX1002 Page 283

目标口音的音频的对应关系，可以是预训练的神经网络模型。在对语音模型进行训练时，可以首先获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的具有目标口音的音频；将训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的具有目标口音的音频作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为上述语音模型。示例的，目标说话文本可以为普通话的文本“我爱祖国”，目标说话文本对应的具有目标口音的音频，可以是东北口音的“我爱祖国”的音频，该音频可以是MP3形式或者也可以是MP4形式，本申请对输出音频的形式不做具体限定。

[0036] 继续参考图3，其示出了根据本申请的用于转换口音的方法的一个应用场景的示意图。在图3的应用场景中，服务器303通过网络302获取目标说话文本301。服务器303基于目标说话文本301，确定用于指示目标口音的第一声学特征304。服务器303确定第一声学特征304中的语言学特征305。服务器303基于语言学特征305和目标说话文本301，确定目标说话文本301对应的具有目标口音的音频306，并输出音频306。本申请实际上是说话文本数据到具有特定口音的音频的转换。文本内容不变。

[0037] 本实施例通过将获取的用户的说话文本数据通过语音合成、识别、转换，可以实现对于用户输入的任意一个文本，能够准确、快速地转换成该用户的具有目标口音的音频。

[0038] 继续参考图4，其示出了根据本申请的用于转换口音的方法的另一个实施例的流程400。如图4所示，本实施例的用于转换口音的方法，可以包括以下步骤：

[0039] 步骤401，获取目标说话文本。

[0040] 步骤402，基于目标说话文本，确定用于指示目标口音的第一声学特征。

[0041] 步骤401～步骤402的原理与步骤201～步骤202的原理类似，此处不再赘述。

[0042] 具体地，步骤402可以通过步骤4021实现：

[0043] 步骤4021，根据目标说话文本和预训练的语音合成模型，确定目标说话文本对应的用于指示目标口音的第一声学特征。

[0044] 本实施例中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。执行主体在得到目标说话文本后，可以根据目标说话文本和预训练的语音合成模型，确定目标说话文本对应的用于指示目标口音的第一声学特征。具体地，第一声学特征可以是用于指示目标口音(比如东北口音)的梅尔频谱。经过语音合成模型得到的目标说话文本对应的用于指示目标口音的第一声学特征此时还不能用于目标口音的输出，还需要进一步的特征识别和特征转换得到携带有较多目标说话文本信息的特征，然后根据得到的这些携带有较多目标说话文本信息的特征进行进一步的语音合成。本实施例，具体地，执行主体可以将目标说话文本输入到预训练的语音合成模型中，由预训练的语音合成模型对目标说话文本进行第一声学特征的提取，得到目标说话文本对应的用于指示目标口音的第一声学特征。

[0045] 本实施例通过根据目标说话文本和预训练的语音合成模型，可以得到具有目标口音的音频梅尔频谱，从而可以基于得到的具有目标口音的音频梅尔频谱准确地对任意一个用户的说话文本进行目标口音的转换，得到具有目标口音的音频。

IPR2026-00272
Krisp Technologies EX1002 Page 284

[0046] 步骤403，确定第一声学特征中的语言学特征。

[0047] 步骤403的原理与步骤203的原理类似，此处不再赘述。

[0048] 具体地，步骤403可以通过步骤4031实现：

[0049] 步骤4031，利用预训练的识别模型提取第一声学特征中的语言学特征。

[0050] 其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。执行主体在得到第一声学特征后，可以利用预训练的识别模型提取第一声学特征中的语言学特征。具体地，第一声学特征中的语言学特征可以包括语调、时域分布、重音、音高、重读、停顿等韵律特征。语言学特征是语言和情绪表达的重要形式之一。

[0051] 在本实施例的一些可选的实现方式中，执行主体还可以根据第一声学特征和预训练的识别模型，确定第一声学特征对应的类别标识，其中，预训练的识别模型用于表征第一声学特征和类别标识的对应关系。得到的类别标识可以是用于表征第一声学特征中各音素的类别的标识，例如，第一声学特征中的各音素可以是语调音素、时域分布音素、重音音素、音高音素、重读音素、停顿音素，例如可以分别用标识1、2、3、4、5、6、7来表示。然后，然后执行主体可以根据第一声学特征中与得到的各标识对应的各音素以及预设的标识、音素与第二声学特征之间的对应关系，确定用于生成具有目标口音的音频的第二声学特征。第二声学特征可以是用于生成目标口音所需的各音素对应的梅尔频谱。执行主体可以根据该第二声学特征，确定与目标说话文本对应的具有目标口音的音频，并输出音频。本实现方式可以丰富用于生成目标口音的音频所需的梅尔频谱，并提高生成目标口音的音频的准确性。

[0052] 本实施例通过利用预训练的识别模型提取第一声学特征中的语言学特征，可以将目标说话文本中对应语言和情绪表达的相关特征提取出来，以完善用于生成目标口音的音频的特征，提高生成具有目标口音的音频的准确性。

[0053] 步骤404，基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0054] 步骤404的原理与步骤204的原理类似，此处不再赘述。

[0055] 具体地，步骤404可以通过步骤4041～步骤4042实现：

[0056] 步骤4041，根据语言学特征、目标说话文本和预训练的转换模型，确定目标说话文本对应的用于生成具有目标口音的音频的第二声学特征。

[0057] 本实施例中，转换模型用于表征语言学特征和目标说话文本与用于生成具有目标口音的音频的第二声学特征的对应关系。具体地，第二声学特征可以包括上述语言学特征对应的梅尔频谱。具体地，第二声学特征可以是用于生成目标口音所需的各音素对应的梅尔频谱。执行主体可以将语言学特征和目标说话文本输入预训练的转换模型，输出目标说话文本对应的用于生成具有目标口音的音频的第二声学特征。

[0058] 步骤4042，基于第二声学特征，确定具有目标口音的音频。

[0059] 执行主体在得到第二声学特征后，可以基于第二声学特征，确定具有目标口音的音频。具体地，执行主体可以将得到的第二声学特征输入到声码器，经过声码器的转换得到具有目标口音的音频。声码器在其发送端对第二声学特征进行编码和加密，以取得和信道的匹配，经信息通道传递到声码器的接受端，在频域中对接收到的特征进行分析，鉴别清浊音，测定浊音基频，进而选取清-浊判断、浊音基频和频谱包络作为特征参量加以传送。当然，该分析也可以在时域中进行，利用其周期性提取一些第二声学特征进行线性预测，生成

IPR2026-00272
Krisp Technologies EX1002 Page 285

与第二声学特征对应的具有目标口音的音频。具体地，上述声码器可以包括通道式声码器、共振峰声码器、图案声码器、线性预测声码器、相关声码器和正交函数声码器，本申请对声码器的类型不做具体限定。

[0060] 本实施例通过根据语言学特征、目标说话文本和预训练的转换模型，确定目标说话文本对应的第二声学特征，并基于第二声学特征确定具有目标口音的音频，可以使得针对于任意用户的任意一个说话文本，均可以得到比较准确的目标口音的音频。

[0061] 在本实施例的一些可选的实现方式中，用于转换口音的方法还包括图4中未示出的以下模型训练步骤：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征；将训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0062] 本实施例中，执行主体可以通过有线连接方式或者无线连接方式获取初始神经网络模型。上述初始神经网络模型可以包括各种包含隐藏层的人工神经网络(Artificial Neural Network，ANN)。在本实施例中，上述执行主体还可以从本地获取预先存储的初始模型，也可以从通信连接的电子设备获取上述初始模型，在此不作限定。

[0063] 本实施例中，执行主体可以通过各种方式获取训练样本集合。具体地，训练样本集合中的训练样本可以包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征。其中，训练样本中标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的、也可以是人工实时标注的、也可以是先通过自动标注后，又人工补充修改纠正标注错误后得到的，本申请对此不做具体限定。训练样本中的语言学特征可以是实时提取的，也可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的。训练样本中的与语言学特征对应的目标说话文本可以是对不同的用户实时采集的，也可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的，本申请对此不做具体限定。

[0064] 本实施例通过获取训练样本集对初始神经网络模型进行训练，可以得到能够根据语言学特征和说话文本准确地生成具有目标口音的音频的能力的转换模型，从而可以实现将获取的目标说话文本上的文字更准确地转换成具有目标口音的音频，并可以提高生成的具有目标口音的音频的质量。

[0065] 在本实施例的一些可选的实现方式中，具有目标口音的音频，包括以下至少一项：具有目标口音的说话音频、具有目标口音的歌唱音频。

[0066] 在本实施例的一些可选的实现方式中，第二声学特征包括对应于目标口音的梅尔频谱；以及基于第二声学特征，确定具有目标口音的音频，包括：根据梅尔频谱和预设的神经网络声码器，合成具有目标口音的音频。

[0067] 具体地，执行主体在得到对应于目标口音的梅尔频谱后，可以将该梅尔频谱自动输入预设的神经网络声码器中，以基于该神经网络声码器和该梅尔频谱合成具有目标口音的音频。

IPR2026-00272
Krisp Technologies EX1002 Page 286

[0068] 本实现方式通过根据对应于目标口音的梅尔频谱和预设的神经网络声码器，可以使得合成的具有目标口音的音频更精确。

[0069] 进一步参考图5，作为对上述各图所示方法的实现，本申请提供了一种用于转换口音的装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0070] 如图5所示，本实施例的用于转换口音的装置500包括：获取单元501、第一声学特征确定单元502、语言学特征确定单元503和转换单元504。

[0071] 获取单元501，被配置成获取目标说话文本。

[0072] 第一声学特征确定单元502，被配置成基于目标说话文本，确定用于指示目标口音的第一声学特征。

[0073] 语言学特征确定单元503，被配置成确定第一声学特征中的语言学特征。

[0074] 转换单元504，被配置成基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0075] 在本实施例的一些可选的实现方式中，第一声学特征确定单元502进一步被配置成：根据目标说话文本和预训练的语音合成模型，确定目标说话文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0076] 在本实施例的一些可选的实现方式中，语言学特征确定单元503进一步被配置成：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0077] 在本实施例的一些可选的实现方式中，转换单元504进一步被配置成：根据语言学特征、目标说话文本和预训练的转换模型，确定目标说话文本对应的用于生成具有目标口音的音频的第二声学特征，其中，转换模型用于表征语言学特征和目标说话文本与用于生成具有目标口音的音频的第二声学特征的对应关系；基于第二声学特征，确定具有目标口音的音频。

[0078] 在本实施例的一些可选的实现方式中，装置还包括图5中未示出的训练单元，被配置成：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征；将训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0079] 在本实施例的一些可选的实现方式中，具有目标口音的音频，包括以下至少一项：具有目标口音的说话音频、具有目标口音的歌唱音频。

[0080] 在本实施例的一些可选的实现方式中，第二声学特征包括对应于目标口音的梅尔频谱；以及转换单元504进一步被配置成：根据梅尔频谱和预设的神经网络声码器，合成具有目标口音的音频。

[0081] 应当理解，用于转换口音的装置500中记载的单元501至单元504分别与参考图2中描述的方法中的各个步骤相对应。由此，上文针对用于转换口音的方法描述的操作和特征

IPR2026-00272
Krisp Technologies EX1002 Page 287

同样适用于装置500及其中包含的单元，在此不再赘述。

[0082] 根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

[0083] 如图6所示，是根据本申请实施例的用于转换口音的方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0084] 如图6所示，该电子设备包括：一个或多个处理器601、存储器602，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线605互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置（诸如，耦合至接口的显示设备）上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线605与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作（例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统）。图6中以一个处理器601为例。

[0085] 存储器602即为本申请所提供的非瞬时计算机可读存储介质。其中，存储器存储有可由至少一个处理器执行的指令，以使至少一个处理器执行本申请所提供的用于转换口音的方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的用于转换口音的方法。

[0086] 存储器602作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及单元，如本申请实施例中的用于转换口音的方法对应的程序指令/单元（例如，附图5所示的获取单元501、第一声学特征确定单元502、语言学特征确定单元503和转换单元504）。处理器601通过运行存储在存储器602中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的用于转换口音的方法。

[0087] 存储器602可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据用于转换口音的电子设备的使用所创建的数据等。此外，存储器602可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器602可选包括相对于处理器601远程设置的存储器，这些远程存储器可以通过网络连接至用于转换口音的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0088] 用于转换口音的方法的电子设备还可以包括：输入装置603和输出装置604。处理器601、存储器602、输入装置603和输出装置604可以通过总线605或者其他方式连接，图6中以通过总线605连接为例。

[0089] 输入装置603可接收输入的数字或字符信息，以及产生与用于转换口音的电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置604可以包括显

IPR2026-00272
Krisp Technologies EX1002 Page 288

示设备、辅助照明装置(例如,LED)和触觉反馈装置(例如,振动电机)等。该显示设备可以包括但不限于,液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中,显示设备可以是触摸屏。

[0090] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0091] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语"机器可读介质"和"计算机可读介质"指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语"机器可读信号"指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0092] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0093] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0094] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0095] 根据本申请实施例的技术方案,通过将获取的用户的说话文本数据通过语音合成、识别、转换,可以实现对于用户输入的任意一个文本,能够准确、快速地转换成该用户的具有目标口音的音频。

[0096] 根据本公开的一个或多个实施例,本公开提供了一种用于转换口音的方法,包括:获取目标说话文本;基于目标说话文本,确定用于指示目标口音的第一声学特征;确定第一声学特征中的语言学特征;基于语言学特征和目标说话文本,确定目标说话文本对应的具

IPR2026-00272
Krisp Technologies EX1002 Page 289

有目标口音的音频，并输出音频。

[0097] 根据本公开的一个或多个实施例，基于目标说话文本，确定用于指示目标口音的第一声学特征，包括：根据目标说话文本和预训练的语音合成模型，确定目标说话文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0098] 根据本公开的一个或多个实施例，确定第一声学特征中的语言学特征，包括：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0099] 根据本公开的一个或多个实施例，基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，包括：根据语言学特征、目标说话文本和预训练的转换模型，确定目标说话文本对应的用于生成具有目标口音的音频的第二声学特征，其中，转换模型用于表征语言学特征和目标说话文本与用于生成具有目标口音的音频的第二声学特征的对应关系；基于第二声学特征，确定具有目标口音的音频。

[0100] 根据本公开的一个或多个实施例，用于转换口音的方法还包括：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征；将训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0101] 根据本公开的一个或多个实施例，具有目标口音的音频，包括以下至少一项：具有目标口音的说话音频、具有目标口音的歌唱音频。

[0102] 根据本公开的一个或多个实施例，第二声学特征包括对应于目标口音的梅尔频谱；以及基于第二声学特征，确定具有目标口音的音频，包括：根据梅尔频谱和预设的神经网络声码器，合成具有目标口音的音频。

[0103] 根据本公开的一个或多个实施例，本公开提供了一种用于转换口音的装置，包括：获取单元，被配置成获取目标说话文本；第一声学特征确定单元，被配置成基于目标说话文本，确定用于指示目标口音的第一声学特征；语言学特征确定单元，被配置成确定第一声学特征中的语言学特征；转换单元，被配置成基于语言学特征和目标说话文本，确定目标说话文本对应的具有目标口音的音频，并输出音频。

[0104] 根据本公开的一个或多个实施例，第一声学特征确定单元进一步被配置成：根据目标说话文本和预训练的语音合成模型，确定目标说话文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0105] 根据本公开的一个或多个实施例，语言学特征确定单元进一步被配置成：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0106] 根据本公开的一个或多个实施例，转换单元进一步被配置成：根据语言学特征、目标说话文本和预训练的转换模型，确定目标说话文本对应的用于生成具有目标口音的音频的第二声学特征，其中，转换模型用于表征语言学特征和目标说话文本与用于生成具有目

IPR2026-00272
Krisp Technologies EX1002 Page 290

标口音的音频的第二声学特征的对应关系；基于第二声学特征，确定具有目标口音的音频。

[0107] 根据本公开的一个或多个实施例，用于转换口音的装置还包括图5中未示出的训练单元，被配置成：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括语言学特征、语言学特征对应的目标说话文本以及标注的与语言学特征和语言学特征对应的目标说话文本对应的第二声学特征；将训练样本集合中的训练样本的语言学特征和语言学特征对应的目标说话文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标说话文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0108] 根据本公开的一个或多个实施例，具有目标口音的音频，包括以下至少一项：具有目标口音的说话音频、具有目标口音的歌唱音频。

[0109] 根据本公开的一个或多个实施例，第二声学特征包括对应于目标口音的梅尔频谱；以及转换单元进一步被配置成：根据梅尔频谱和预设的神经网络声码器，合成具有目标口音的音频。

[0110] 应该理解，以上实施例仅是示例性实施例，但不限于此，还包括本领域已知的其他可以实现用于转换口音的方法。可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

[0111] 上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

IPR2026-00272
Krisp Technologies EX1002 Page 291

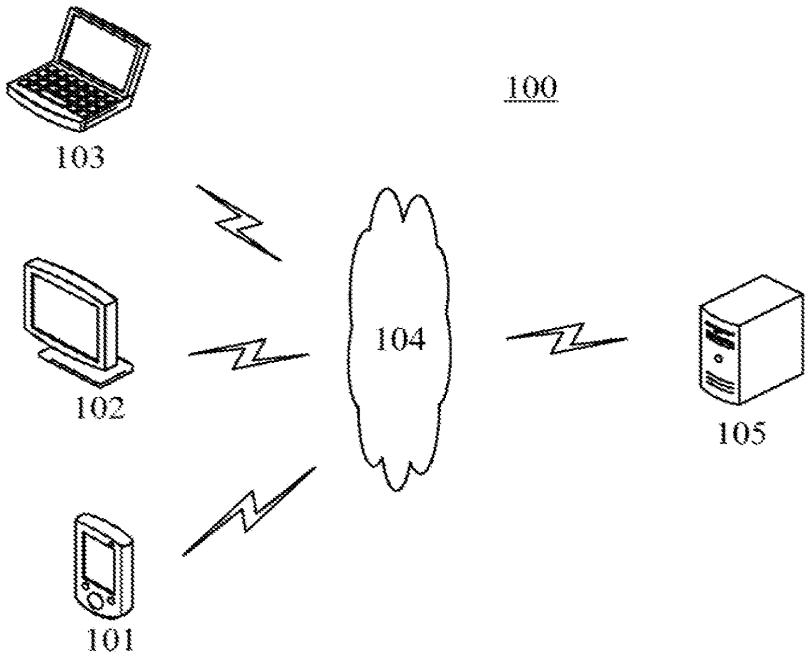


图1

IPR2026-00272
Krisp Technologies EX1002 Page 292

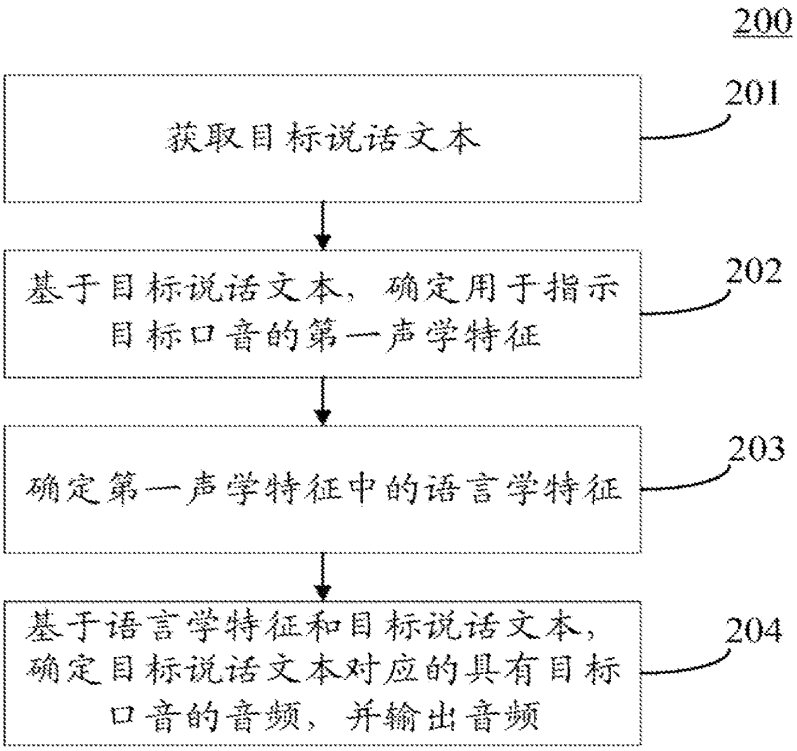


图2

IPR2026-00272
Krisp Technologies EX1002 Page 293

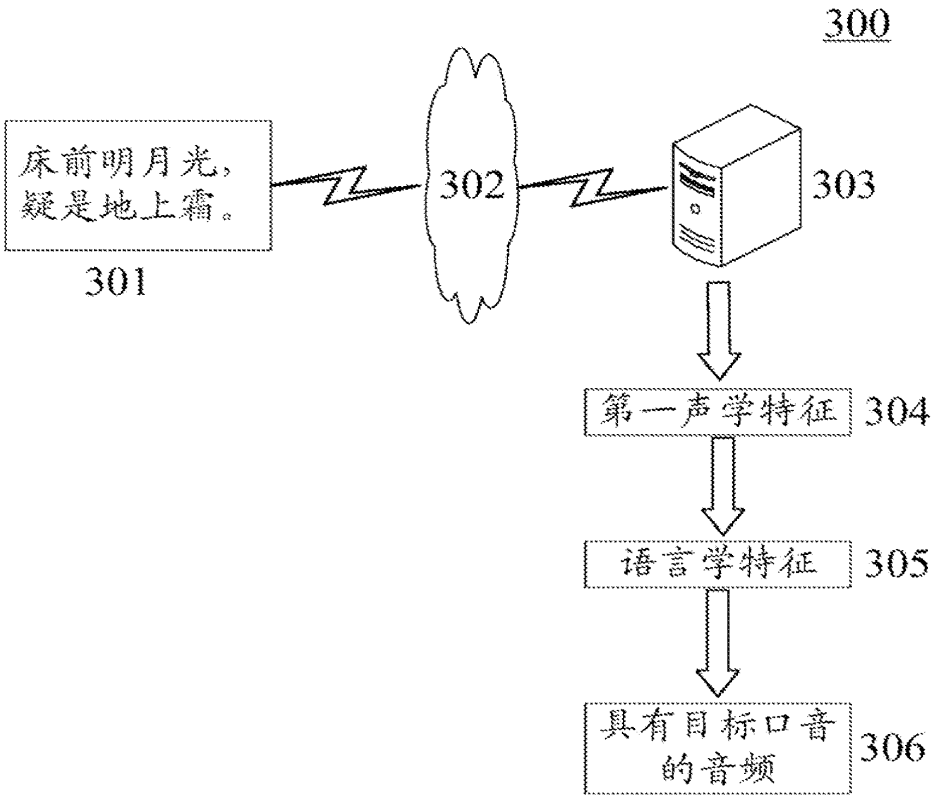


图3

IPR2026-00272
Krisp Technologies EX1002 Page 294

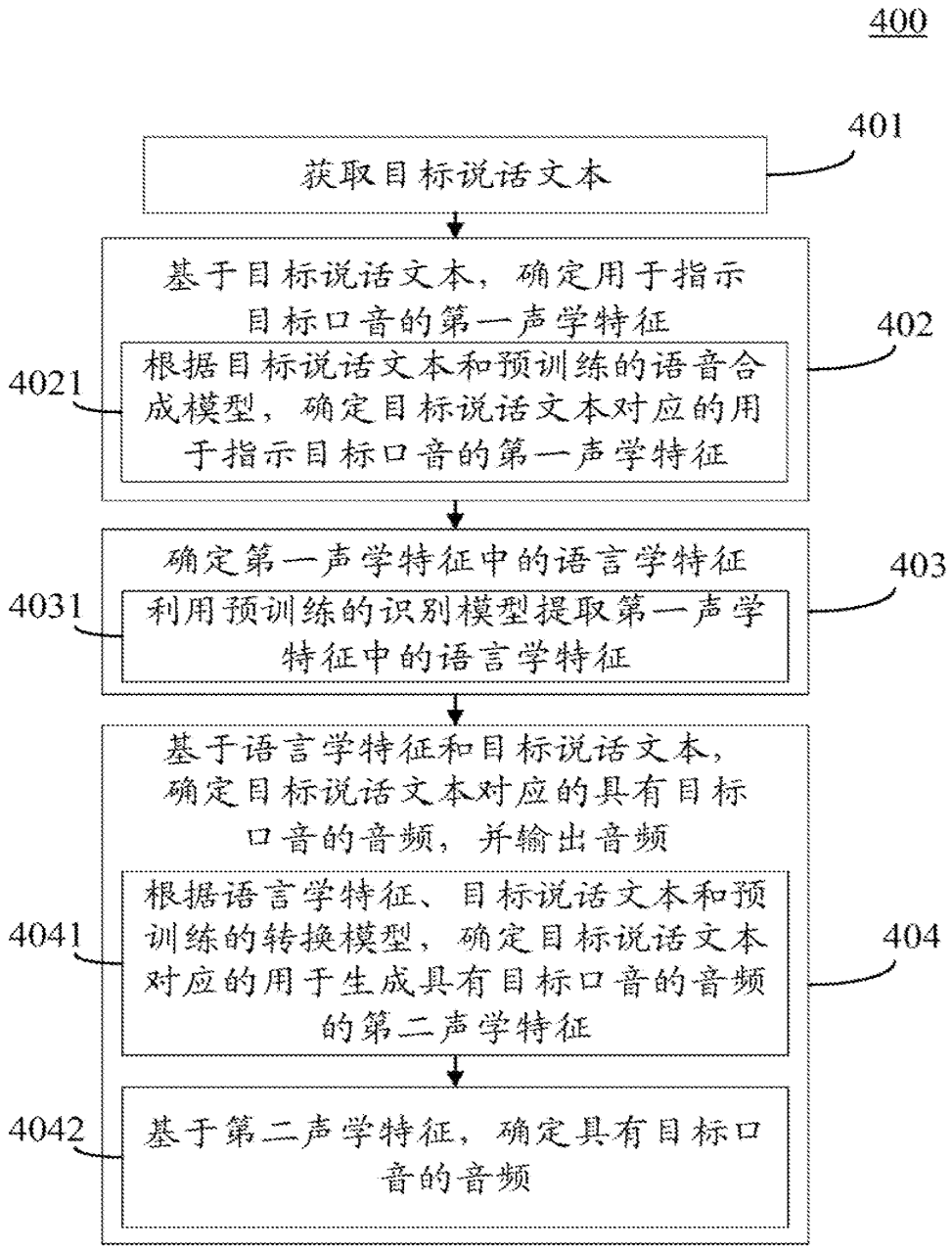


图4

IPR2026-00272
Krisp Technologies EX1002 Page 295

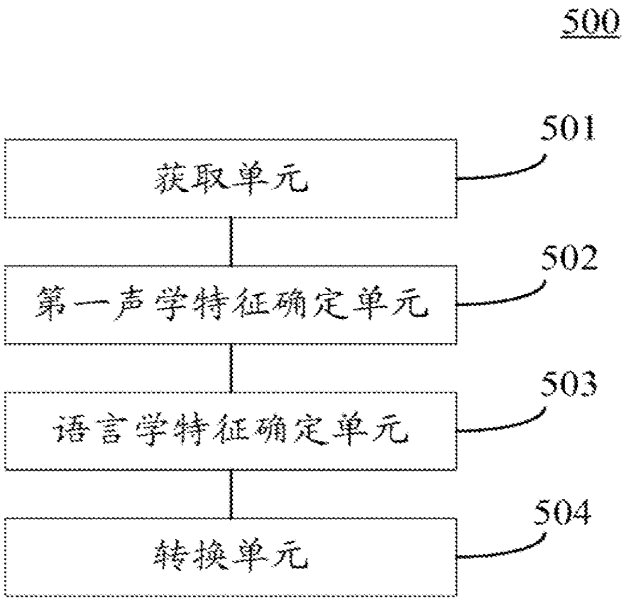


图5

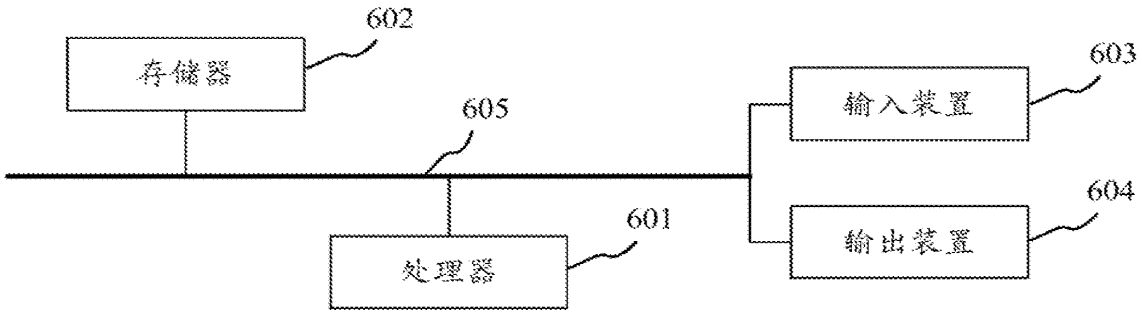


图6

IPR2026-00272
Krisp Technologies EX1002 Page 296

Espacenet

CN112382270A Speech synthesis method, device and equipment and storage medium

Applicants: BEIJING YOUZHUJU NETWORK TECH CO LTD

Inventors: TANG BENLAI,LI ZHONGHAO

Classifications:
IPC ***G10L13/02; G10L13/08; G10L25/03; G10L25/30;***

CPC **G10L13/02 (CN); G10L13/086 (CN); G10L25/03 (CN); G10L25/30 (CN);**

Priorities: CN202011273128A 2020-11-13

Application: CN202011273128A·2020-11-13

Publication: CN112382270A·2021-02-19

Published as: CN112382270A

Speech synthesis method, device and equipment and storage medium

Abstract

The invention discloses a speech synthesis method, device and equipment and a storage medium, and relates to the fields of speech synthesis, natural language processing, computer technology, artificial intelligence and deep learning. According to the specific implementation scheme, the method comprises the steps of acquiring a target text and a target audio; determining first acoustic features forindicating a target accent based on the target text; determining a target tone based on the target audio; determining linguistic features in the first acoustic features, and determining an audio withthe target accent corresponding to the target text based on the target tone, the linguistic features and the target text, and outputting the target audio. According to the implementation mode, the target audio with the target accent and the target tone corresponding to the target text can be accurately and quickly synthesized according to the linguistic features determined according to the obtained target text and the target tone determined according to the acquired target audio.

IPR2026-00272
Krisp Technologies EX1002 Page 297

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

DESCRIPTION CN112382270A

Speech synthesis method, apparatus, device and storage medium

[0001]

technical field

[0002]

The present application relates to the field of computer technology, in particular to the fields of speech synthesis, natural language processing, computer technology, artificial intelligence, and deep learning, and in particular to speech synthesis methods, devices, equipment, and storage media.

[0003]

Background technique

[0004]

In recent years, due to the rapid development of online education and online learning, speech synthesis technology has received extensive research and attention. Speech synthesis aims to synthesize a user's speech into different accents or different timbres or both audio.

Speech synthesis technology also has great application prospects in entertainment.

The speed of speech synthesis using the existing speech synthesis technology is slow and the results of speech synthesis are often inaccurate.

[0005]

SUMMARY OF THE INVENTION

[0006]

The present disclosure provides a speech synthesis method, apparatus, device, and storage medium.

[0007]

IPR2026-00272
Krisp Technologies EX1002 Page 298

According to an aspect of the present disclosure, there is provided a speech synthesis method, comprising: acquiring target text and target audio; determining a first acoustic feature for indicating a target accent based on the target text; determining a target timbre based on the target audio; determining The linguistic feature in the first acoustic feature; based on the target timbre, the linguistic feature and the target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre.

[0008]

According to another aspect of the present disclosure, there is provided a speech synthesis apparatus, comprising: an acquisition unit configured to acquire target text and target audio, and a first acoustic feature determination unit configured to determine, based on the target text a first acoustic feature of the target accent; a target timbre determination unit configured to determine a target timbre based on the target audio; a linguistic feature determination unit configured to determine a linguistic feature in the first acoustic feature; a conversion unit configured to be it is configured to determine the audio with the target accent and the target timbre corresponding to the target text based on the target timbre, the linguistic feature and the target text, and output the audio with the target accent and the target timbre.

[0009]

According to yet another aspect of the present disclosure, there is provided an electronic device for speech synthesis, comprising: at least one processor; and a memory communicatively connected to the at least one processor; wherein the memory stores instructions executable by the at least one processor, The instructions are executed by the at least one processor to enable the at least one processor to perform the speech synthesis method as described above.

[0010]

According to yet another aspect of the present disclosure, there is provided a non-transitory computer-readable storage medium storing computer instructions, wherein the computer instructions are used to cause a computer to execute the above-mentioned speech synthesis method.

[0011]

The technology according to the present application solves the problem that speech synthesis cannot be performed accurately and quickly. Through the linguistic features determined according to the acquired target text and the target timbre determined by the acquired target audio, it is possible to accurately and quickly synthesize the corresponding Target audio with target accent and target timbre.

[0012]

It should be understood that what is described in this section is not intended to identify key or critical features of embodiments of the disclosure, nor is it intended to limit the scope of the disclosure.

Other features of the present disclosure will become readily understood from the following description.

[0013]

Description of drawings

[0014]

The accompanying drawings are used for better understanding of the present solution, and do not constitute a limitation to the present application.

in:

[0015]

FIG. 1 is an exemplary system architecture diagram to which an embodiment of the present application may be applied;

IPR2026-00272
Krisp Technologies EX1002 Page 299

[0016]

Fig. 2 is a flow chart of an embodiment of a speech synthesis method according to the present application;

[0017]

3 is a schematic diagram of an application scenario of the speech synthesis method according to the present application;

[0018]

4 is a flowchart of another embodiment of a speech synthesis method according to the present application;

[0019]

5 is a schematic structural diagram of an embodiment of a speech synthesis apparatus according to the present application;

[0020]

FIG. 6 is a block diagram of an electronic device used to implement the speech synthesis method of the embodiment of the present application.

[0021]

Detailed ways

[0022]

Exemplary embodiments of the present application are described below with reference to the accompanying drawings, which include various details of the embodiments of the present application to facilitate understanding, and should be considered as exemplary only.

Accordingly, those of ordinary skill in the art will recognize that various changes and modifications of the embodiments described herein can be made without departing from the scope and spirit of the present application.

Also, descriptions of well-known functions and constructions are omitted from the following description for clarity and conciseness.

[0023]

It should be noted that the embodiments in the present application and the features of the embodiments may be combined with each other in the case of no conflict.

The present application will be described in detail below with reference to the accompanying drawings and in conjunction with the embodiments.

[0024]

FIG. 1 illustrates an exemplary system architecture 100 to which embodiments of the speech synthesis method or speech synthesis apparatus of the present application may be applied.

[0025]

As shown in FIG. 1 , the system architecture 100 may include terminal devices 101 , 102 , and 103 , a network 104 and a server 105 .

The network 104 is a medium used to provide a communication link between the terminal devices 101 , 102 , 103 and the server 105 .

IPR2026-00272
Krisp Technologies EX1002 Page 300

The network 104 may include various connection types, such as wired, wireless communication links, or fiber optic cables, among others.

[0026]

The user can use the terminal devices 101, 102, 103 to interact with the server 105 through the network 104 to receive or send messages and the like.

Various communication client applications, such as language synthesis applications, may be installed on the terminal devices 101 , 102 and 103 .

[0027]

The terminal devices 101, 102, and 103 may be hardware or software.

When the terminal devices 101, 102, and 103 are hardware, they can be various electronic devices, including but not limited to smart phones, tablet computers, in-vehicle computers, laptop computers, desktop computers, and the like.

When the terminal devices 101, 102, and 103 are software, they can be installed in the electronic devices listed above.

It can be implemented as a plurality of software or software modules, and can also be implemented as a single software or software module.

There is no specific limitation here.

[0028]

The server 105 may be a server that provides various services, such as a background server that processes the target text and target audio collected by the terminal devices 101 , 102 , and 103 .

The background server can acquire the target text and the target audio, and based on the target text, determine the first acoustic feature used to indicate the target accent; determine the target timbre based on the target audio; determine the linguistic feature in the first acoustic feature; based on the target Timbre, linguistic feature and target text, determine the audio with target accent and target timbre corresponding to the target text, and output the audio.

[0029]

It should be noted that the server 105 may be hardware or software.

When the server 105 is hardware, it can be implemented as a distributed server cluster composed of multiple servers, or can be implemented as a single server.

When the server 105 is software, it may be implemented as a plurality of software or software modules, or may be implemented as a single software or software module.

There is no specific limitation here.

[0030]

It should be noted that the speech synthesis method provided by the embodiment of the present application is generally executed by the server 105 .

Correspondingly, the speech synthesis device is generally provided in the server 105 .

[0031]

It should be understood that the numbers of terminal devices, networks and servers in FIG. 1 are merely illustrative.

There can be any number of terminal devices, networks and servers according to implementation needs.

[0032]

IPR2026-00272
Krisp Technologies EX1002 Page 301

Continuing to refer to FIG. 2, a flow 200 of an embodiment of a speech synthesis method according to the present application is shown.

The speech synthesis method of the present embodiment includes the following steps:

[0033]

Step 201, acquiring target text and target audio.

[0034]

In this embodiment, the execution body of the speech synthesis method (for example, the server 105 in FIG. 1) can obtain the target text locally, or can obtain the target text collected by the terminal device through recording or scanning through a wired connection or a wireless connection. , or any sentence entered by the user in the edit box of the terminal device can be obtained and determined as the target text.

Specifically, the target text may be the speech text converted from the speech uttered by any person with a definite identity who wants to convert the accent.

Spoken text, which can be text.

For example, if Xiao Ming said "I love the motherland" in Mandarin, the target text can be expressed as "I love the motherland" in literal form.

The target text may be Chinese or English, and the application does not specifically limit the language type of the target text.

The target audio can be the audio corresponding to the timbre that needs to be converted.

For example, the audio may be the audio recorded with the voice of the student A, or the audio recorded with the voice of the student B, and the application does not specifically limit the timbre of the target audio.

The target audio may be in the form of MP3 or MP4, and this application does not limit the storage form of the target audio.

It can be understood that the target audio may be the audio of a human being or the audio of other creatures in nature, and the source of the target audio is not specifically limited in this application.

[0035]

Step 202 , based on the target text, determine a first acoustic feature for indicating the target accent.

[0036]

After acquiring the target text, the executing subject may determine the first acoustic feature for indicating the target accent based on the target text.

Specifically, the first acoustic feature may be a speech feature parameter used to characterize the target accent, such as a Mel spectrum.

In this embodiment, the execution body may pre-emphasize, divide, and window the audio based on the target text, according to the target text and its corresponding audio, and then perform short-time Fourier transform (STFT) on the signal of each frame of audio. , short-time Fourier transform, or short-term Fourier transform) to obtain a short-term amplitude spectrum, the short-term amplitude spectrum obtains a Mel spectrum through a Mel filter bank.

The present application does not specifically limit the acquisition method of the Mel spectrum (ie, the first acoustic feature).

[0037]

Step 203: Determine the target timbre based on the target audio.

[0038]

After acquiring the target audio, the execution body can determine the target timbre based on the target audio.

IPR2026-00272
Krisp Technologies EX1002 Page 302

Audio contains multiple phonemes.

A phoneme is the smallest phonetic unit divided according to the natural properties of speech. It is analyzed according to the pronunciation action in the syllable, and an action constitutes a phoneme.

Phonemes are divided into vowels and consonants.

For example, the Chinese syllable ah (ā) has only one phoneme, love (ài) has two phonemes, and dai (dài) has three phonemes.

For example, in [ma-mi], two [m] have the same pronunciation action and are the same phoneme, and [a] and [i] have different pronunciation actions and are different phonemes. Timbre means that different sounds always have different characteristics in terms of waveform, and different objects have different characteristics of vibration. Different sound generators have different timbres due to their different materials and structures. For example, pianos, violins, and people make different sounds, and everyone makes different sounds. Therefore, timbre can be understood as the characteristic of sound. Timbre is one of the properties of a sound (i.e. loudness, pitch, timbre) and is mainly determined by its overtones. The difference between each person's voice and the sounds produced by various musical instruments is caused by different timbres. The vibration of the sounding body is composed of a variety of overtones, including fundamental and overtones. Different overtones determine a specific timbre. In addition to a 'fundamental tone', the sound also naturally adds many different 'frequency' (the number of times the vibrating object vibrates in 1 second) and the overtone 'interlacing', which determines different timbres, so that people can distinguish them after listening to them. is a different voice. Specifically, the execution subject can determine the corresponding identification of each phoneme in the target audio according to the target audio and the pre-trained classification model, wherein the pre-trained classification model is used to characterize the correspondence between the phoneme and the identification; Determine the fundamental tone and overtone in the target audio; determine the target timbre according to the determined fundamental tone and overtone in the target audio. Specifically, the execution subject can input the determined fundamental tone and overtone in the target audio into a pre-trained tone conversion model, and output the tone corresponding to the fundamental tone and overtone, wherein the pre-trained tone conversion model is used to represent fundamental tone, Correspondence between overtones and timbres. Exemplarily, the tone conversion model may be a pre-trained convolutional neural network (Convolutional Neural Networks, CNN).

[0039]

Step 204, determining linguistic features in the first acoustic features.

[0040]

After determining the first acoustic feature, the executing subject may determine the linguistic feature in the first acoustic feature.

Specifically, linguistic features may include prosodic features, syntax, discourse structure, information structure, and the like.

Among them, the prosodic feature can also be a supersonic feature or a supersegmental feature, which is a phonological structure of a language. The prosodic features can be divided into three main aspects: intonation, temporal distribution and stress, which are achieved by suprasegmental features. Suprasegmental features include pitch, intensity, and temporal properties, carried by phonemes or groups of phonemes. Prosody is a typical feature of human natural language, and has many common features across languages, such as pitch dip, stress, pause, etc., which are commonly found in different languages. Prosodic features are one of the important forms of language and emotional expression. The execution subject may compare the existing known linguistic features with each feature in the first acoustic feature, and determine a feature in the first acoustic feature whose similarity with the known linguistic feature is greater than a threshold value as the first acoustic feature. Linguistic features in acoustic features. Specifically, the execution body may convert each feature in the first acoustic feature into a first acoustic feature vector, and convert an existing known linguistic feature into a linguistic feature vector. Calculate the cosine similarity between each first acoustic feature vector and each linguistic feature vector. The closer the cosine similarity is to 1, it indicates that the first acoustic feature vector and the corresponding linguistic feature vector participate in the calculation of the cosine similarity. The more similar, the first acoustic feature vector may be determined as a linguistic feature vector, and accordingly, the first acoustic feature corresponding to the first acoustic feature vector is determined as the linguistic feature corresponding to the linguistic feature vector. In the same way, traverse each first acoustic feature and each linguistic feature until the absolute value of the difference between the cosine similarity and 1 of the corresponding vectors of each first acoustic feature and each linguistic feature is found to be less than The first acoustic feature corresponding to the first acoustic feature

IPR2026-00272
Krisp Technologies EX1002 Page 303

vector of the threshold value is determined, and the first acoustic feature is determined as a linguistic feature, according to which all linguistic features in the first acoustic feature can be obtained.

[0041]

Step 205 , based on the target timbre, the linguistic feature and the target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre.

[0042]

After determining the linguistic feature, the execution body may determine the audio with the target accent corresponding to the target text based on the target timbre, the linguistic feature and the target text, and output the audio.

Specifically, the executing subject may input the target timbre, linguistic features and target text into the pre-trained speech model to generate audio with the target accent and target timbre corresponding to the target text, and output the audio.

The pre-trained speech model is used to represent the target timbre, linguistic features, and the correspondence between the target text and the audio with the target accent and the target timbre, and may specifically be a pre-trained neural network model. When training the speech model, an initial neural network model can be obtained first; a training sample set can be obtained, wherein the training samples in the training sample set include target timbre, linguistic features, target text corresponding to the linguistic features, and labels corresponding to the target Timbre, linguistic features, and the audio with target accent and target timbre corresponding to the target text corresponding to linguistic features; the target text corresponding to the target timbre, linguistic features, and linguistic features of the training samples in the training sample set is used as the initial neural The input of the network model takes the audio with the target accent and the target timbre corresponding to the input linguistic features and the target text corresponding to the linguistic features as the expected output, and trains the initial neural network model; the trained initial neural network model Determined to be the above voice model. For example, the target text may be the text "I love the motherland" corresponding to the Mandarin audio of Xiao B, and the audio corresponding to the target text with the target accent and the target timbre may be the audio of "I love the motherland" of the northeastern accent of Xiao A, The audio may be in the form of MP3 or may also be in the form of MP4, and this application does not specifically limit the form of the output audio.

[0043]

Continue to refer to FIG. 3 , which shows a schematic diagram of an application scenario of the speech synthesis method according to the present application.

In the application scenario of FIG. 3 , the server 304 obtains the target text 301 and the target audio 302 through the network 303 . Based on the target text 301, the server 304 determines a first acoustic feature 305 indicative of the target accent. Server 304 determines target tone color 307 based on target audio 302 . The server 304 determines the linguistic features 306 in the first acoustic features 305 . The server 304 determines the audio 308 with the target accent and the target timbre corresponding to the target text 301 based on the target timbre 307, the linguistic feature 306 and the target text 301, and outputs the audio 308 with the target accent and the target timbre. This application is actually the conversion of target spoken text to audio with any accent, and the text content remains unchanged.

[0044]

In this embodiment, the target audio with the target accent and the target timbre corresponding to the target text can be synthesized accurately and quickly by using the linguistic features determined according to the acquired target text and the target timbre determined by the acquired target audio.

[0045]

Continuing to refer to FIG. 4, a flow 400 of another embodiment of the speech synthesis method according to the present application is shown.

As shown in Figure 4, the speech synthesis method of this embodiment may include the following steps:

[0046]

IPR2026-00272
Krisp Technologies EX1002 Page 304

Step 401, acquiring target text and target audio.

[0047]

Step 402 , based on the target text, determine a first acoustic feature for indicating the target accent.

[0048]

The principles of steps 401 to 402 are similar to those of steps 201 to 202, and are not repeated here.

[0049]

Specifically, step 402 can be implemented through step 4021:

[0050]

Step 4021: Determine, according to the target text and the pre-trained speech synthesis model, a first acoustic feature corresponding to the target text and used to indicate the target accent.

[0051]

In this embodiment, the pre-trained speech synthesis model is used to represent the correspondence between the spoken text and the first acoustic feature.

After obtaining the target text, the execution subject may determine, according to the target text and the pre-trained speech synthesis model, the first acoustic feature corresponding to the target text and used to indicate the target accent.

Specifically, the first acoustic feature may be a Mel spectrum used to indicate a target accent, such as a northeastern accent.

The first acoustic feature corresponding to the target text obtained through the speech synthesis model and used to indicate the target accent cannot be used for the output of the target accent at this time, and further feature recognition and feature conversion are required to obtain the target text that carries more information about the target text. Features, such as pitch, intensity, etc.

Then further speech synthesis is performed according to the obtained features that carry more target text information.

In this embodiment, specifically, the execution body can input the target text into a pre-trained speech synthesis model, and the pre-trained speech synthesis model extracts the first acoustic feature of the target text, and obtains the corresponding target text for indicating The first acoustic feature of the target accent.

[0052]

In this embodiment, according to the target text and the pre-trained speech synthesis model, the audio mel spectrum with the target accent can be obtained, so that the speech text of any user can be accurately targeted based on the obtained audio mel spectrum with the target accent. Accent conversion.

[0053]

Step 403: Determine the target timbre based on the target audio.

[0054]

The principle of step 403 is similar to that of step 203, and details are not repeated here.

[0055]

Specifically, step 403 can be implemented through steps 4031 to 4032:

IPR2026-00272
Krisp Technologies EX1002 Page 305

[0056]

Step 4031: Determine the identity vector corresponding to the target audio according to the target audio and the pre-trained authentication model.

[0057]

In this embodiment, the pre-trained authentication model is used to represent the correspondence between the audio and the identity vector.

After acquiring the target audio, the execution subject can determine the identity vector corresponding to the target audio according to the target audio and the pre-trained authentication model.

Specifically, the execution subject can input the target audio into the pre-trained authentication model, and output the identity vector corresponding to the target audio.

Specifically, the identity vector corresponding to the output target audio can be a set of multi-dimensional data, which can identify the timbre information of the speaker corresponding to the target audio. For example, the identity vector can be [0.3, 0.3, 0.5, 0.6, ...] such as A vector corresponding to a data sequence, the combination of one or more data in the data sequence can be used to represent a unique timbre.

[0058]

Step 4032: Determine the target timbre according to the identity vector.

[0059]

After determining the identity vector corresponding to the target audio, the execution subject can determine the target timbre according to the identity vector.

Specifically, the execution subject may determine the target timbre corresponding to the identity vector according to the identity vector and the preset correspondence between the identity vector and the timbre.

As another implementation manner, the execution body may also determine the timbre vector corresponding to the identity vector according to the similarity between the identity vector and each existing timbre vector, and determine the timbre corresponding to the timbre vector as the target timbre.

Specifically, in response to determining that the similarity between the identity vector and the corresponding timbre vector is greater than the preset threshold, the execution body determines the timbre corresponding to the timbre vector corresponding to the identity vector as the target timbre.

[0060]

In this embodiment, the target timbre is determined by the identity vector obtained from the identity verification model, which is not limited by the model training sample set, and the target text can be converted into any desired timbre, which improves the flexibility of timbre conversion and improves user experience.

[0061]

Step 404, determining linguistic features in the first acoustic features.

[0062]

The principle of step 404 is similar to that of step 204, and details are not repeated here.

[0063]

Specifically, step 404 can be implemented through step 4041:

IPR2026-00272
Krisp Technologies EX1002 Page 306

[0064]

Step 4041, using the pre-trained recognition model to extract linguistic features in the first acoustic features.

[0065]

The pre-trained recognition model is used to represent the correspondence between the first acoustic feature and the linguistic feature.

After obtaining the first acoustic feature, the executing subject may extract the linguistic feature in the first acoustic feature by using the pre-trained recognition model.

Specifically, the linguistic features in the first acoustic feature may include prosody features such as intonation, temporal distribution, stress, pitch, stress, pause, and the like.

Linguistic features are one of the important forms of language and emotional expression.

[0066]

In some optional implementations of this embodiment, the execution body may also determine a category identifier corresponding to the first acoustic feature according to the first acoustic feature and a pre-trained identification model, where the pre-trained identification model is used for The correspondence between the first acoustic feature and the class identifier is characterized.

The obtained category identifier may be an identifier used to characterize the category of each phoneme in the first acoustic feature, for example, each phoneme in the first acoustic feature may be a tone phoneme, a time-domain distribution phoneme, an accent phoneme, a pitch phoneme, Stressed phonemes and paused phonemes, for example, can be represented by signs 1, 2, 3, 4, 5, 6, and 7, respectively.

Then, the execution main body may determine, according to each phoneme in the first acoustic feature corresponding to each obtained logo and the preset corresponding relationship between the logo, the phoneme and the second acoustic feature, the method for generating the audio with the target accent. The second acoustic signature.

The second acoustic feature may be a Mel spectrum corresponding to each phoneme required for generating the target accent.

The execution body may determine audio with the target accent and the target timbre corresponding to the target text based on the second acoustic feature, and output the audio.

This implementation can enrich the Mel spectrum required for generating the audio of the target accent and the target timbre, and improve the accuracy of generating the audio of the target accent and the target timbre.

[0067]

In this embodiment, by using the pre-trained recognition model to extract the linguistic features in the first acoustic features, the relevant features of the corresponding language and emotional expression in the target text can be extracted, so as to improve the audio used to generate the target accent and target timbre features that improve the accuracy of generating audio with the target accent and timbre.

[0068]

Step 405 , based on the target timbre, linguistic features and target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre.

[0069]

The principle of step 405 is similar to that of step 205, and details are not repeated here.

[0070]

IPR2026-00272
Krisp Technologies EX1002 Page 307

Specifically, step 405 can be implemented through steps 4051 to 4052:

[0071]

Step 4051: Determine, according to the target timbre, linguistic features, target text and the pre-trained conversion model, a second acoustic feature corresponding to the target text for generating audio with the target accent and target timbre.

[0072]

Among them, the transformation model is used to characterize the timbre, linguistic features, and the correspondence between the text and the second acoustic feature.

Specifically, the second acoustic feature may be the Mel spectrum corresponding to the above-mentioned target timbre and linguistic feature.

Specifically, the second acoustic feature may be a Mel spectrum required for generating the audio of the target accent and the target timbre.

The execution body may input the target timbre, linguistic features and target text into the pre-trained conversion model, and output the second acoustic feature corresponding to the target text for generating audio with the target accent and target timbre.

[0073]

Step 4052 , based on the second acoustic feature, determine the audio corresponding to the target text with the target accent and the target timbre.

[0074]

After obtaining the second acoustic feature, the executing subject may determine the audio with the target accent and the target timbre based on the second acoustic feature.

Specifically, the execution body may input the obtained second acoustic feature into the vocoder, and obtain audio with the target accent and the target timbre after conversion by the vocoder.

The vocoder encodes and encrypts the second acoustic feature at its transmitting end to obtain a match with the channel, transmits it to the receiving end of the vocoder through the information channel, analyzes the received feature in the frequency domain, and identifies the clear signal. For voiced sound, the fundamental frequency of the voiced sound is measured, and then the unvoiced-voiced judgment, the fundamental frequency of the voiced sound and the spectral envelope are selected as characteristic parameters to transmit.

Of course, the analysis can also be performed in the time domain, using the periodic extraction of some second acoustic features to perform linear prediction to generate audio with the target accent and the target timbre corresponding to the second acoustic features.

Specifically, the above-mentioned vocoder may include a channel vocoder, a formant vocoder, a pattern vocoder, a linear prediction vocoder, a correlation vocoder and an orthogonal function vocoder. The type is not specifically limited.

[0075]

In this embodiment, the second acoustic feature corresponding to the target text is determined according to the target timbre, the linguistic feature, the target text and the pre-trained conversion model, and the audio with the target accent and the target timbre is determined based on the second acoustic feature, so that the For any spoken text of any user, it is not limited by the training samples of the conversion model, and the target text can be converted into audio with any timbre and accent that you want to convert, which enhances the fun of speech synthesis and enhances the user experience.

[0076]

In some optional implementations of this embodiment, the speech synthesis method further includes the following model training steps not shown in FIG. 4 : acquiring an initial neural network model; acquiring a training sample set, wherein the training

IPR2026-00272
Krisp Technologies EX1002 Page 308

samples in the training sample set including the timbre, linguistic features, the target text corresponding to the linguistic features, and the marked second acoustic feature corresponding to the target text corresponding to the timbre, linguistic features, and linguistic features; the timbre, language of the training samples in the training sample set The target text corresponding to the linguistic feature and the linguistic feature is used as the input of the initial neural network model, and the second acoustic feature corresponding to the target text corresponding to the input timbre, linguistic feature and linguistic feature is used as the expected output. Perform training; determine the initial neural network model after training as the transformation model.

[0077]

In this embodiment, the execution body may acquire the initial neural network model through wired connection or wireless connection.

The above-mentioned initial neural network model may include various artificial neural networks (Artificial Neural Network, ANN) including hidden layers.

In this embodiment, the above-mentioned execution body may also acquire a pre-stored initial model locally, or may acquire the above-mentioned initial model from a communication-connected electronic device, which is not limited herein.

[0078]

In this embodiment, the execution body may acquire the training sample set in various ways.

Specifically, the training samples in the training sample set may include timbre, linguistic features, target text corresponding to the linguistic features, and annotated second acoustic features corresponding to the timbre, linguistic features, and target text corresponding to the linguistic features.

The second acoustic feature corresponding to the target text corresponding to the timbre, linguistic feature, and linguistic feature marked in the training sample may be obtained from a local or communicatively connected electronic device through a wired or wireless connection, or it may be it is manually annotated in real time, or it can be obtained through automatic annotation first, and then manually supplemented and modified to correct the annotation error, which is not specifically limited in this application.

The timbres in the training samples can be obtained locally or from a communicatively connected electronic device.

Linguistic features in training samples can be extracted in real time, or acquired from local or communicatively connected electronic devices through wired or wireless connections.

The target text corresponding to the linguistic feature in the training sample may be collected in real time from different users, or may be obtained from a local or communication-connected electronic device through a wired or wireless connection, which is not covered in this application. Specific restrictions.

[0079]

In this embodiment, by acquiring the training sample set to train the initial neural network model, a conversion model capable of accurately generating the corresponding second acoustic feature according to the timbre, linguistic features and speech text can be obtained, so that any acquired The text on the target text is converted into audio with the target accent and the target timbre, and the quality of the generated audio with the target accent and the target timbre can be improved. The present application is not limited to the existing timbres in the training sample set for training the conversion model. For Other timbres that are not in the training sample set can also accurately synthesize audio with the timbre, which enhances the fun of speech synthesis and enhances the user experience.

[0080]

In some optional implementations of this embodiment, the training sample set does not include the target timbre; the second acoustic feature includes a Mel spectrum corresponding to the target accent and the target timbre; and based on the second acoustic feature, it is determined that the target text corresponds to The audio with the target accent and the target timbre includes: according to the Mel spectrum and the preset neural network vocoder, the audio with the target accent and the target timbre corresponding to the target text is synthesized.

[0081]

IPR2026-00272
Krisp Technologies EX1002 Page 309

Specifically, after obtaining the Mel spectrum corresponding to the target accent and the target timbre, the executing subject can automatically input the Mel spectrum into a preset neural network vocoder, so as to obtain the Mel spectrum based on the neural network vocoder and the Mel spectrum. Spectral synthesis of audio with target accents and target timbres.

The audio with the target accent and the target timbre includes at least one of the following: speaking audio with the target accent and the target timbre, and singing audio with the target accent and the target timbre.

[0082]

This implementation can make the synthesized audio with the target accent and the target timbre more accurate by using the Mel spectrum corresponding to the target accent and the target timbre and the preset neural network vocoder.

[0083]

Referring further to FIG. 5 , as an implementation of the methods shown in the above figures, the present application provides an embodiment of a speech synthesis apparatus. The apparatus embodiment corresponds to the method embodiment shown in FIG. 2 . Specifically, the apparatus may Used in various electronic devices.

[0084]

As shown in FIG. 5 , the speech synthesis apparatus 500 of this embodiment includes: an acquisition unit 501 , a first acoustic feature determination unit 502 , a target timbre determination unit 503 , a linguistic feature determination unit 504 , and a conversion unit 505 .

[0085]

The acquiring unit 501 is configured to acquire target text and target audio.

[0086]

The first acoustic feature determination unit 502 is configured to determine, based on the target text, a first acoustic feature for indicating the target accent.

[0087]

The target timbre determination unit 503 is configured to determine the target timbre based on the target audio.

[0088]

The linguistic feature determination unit 504 is configured to determine linguistic features in the first acoustic features.

[0089]

The conversion unit 505 is configured to, based on the target timbre, the linguistic feature and the target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre.

[0090]

In some optional implementations of this embodiment, the first acoustic feature determining unit 502 is further configured to: determine, according to the target text and the pre-trained speech synthesis model, a first acoustic feature indicating the target accent corresponding to the target text Acoustic features, wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0091]

IPR2026-00272
Krisp Technologies EX1002 Page 310

In some optional implementations of this embodiment, the target timbre determining unit 503 is further configured to: determine an identity vector corresponding to the target audio according to the target audio and a pre-trained authentication model, wherein the pre-trained authentication model It is used to characterize the correspondence between the audio and the identity vector; according to the identity vector, the target timbre is determined.

[0092]

In some optional implementations of this embodiment, the linguistic feature determining unit 503 is further configured to: extract linguistic features in the first acoustic feature by using a pre-trained recognition model, wherein the pre-trained recognition model uses it is used to characterize the correspondence between the first acoustic feature and the linguistic feature.

[0093]

In some optional implementations of this embodiment, the conversion unit 505 is further configured to: according to the target timbre, linguistic features, target text and the pre-trained conversion model, determine the corresponding target text for generating the target accent and The second acoustic feature of the audio of the target timbre, wherein the conversion model is used to represent the timbre, linguistic features and the correspondence between the text and the second acoustic feature; based on the second acoustic feature, determine the target text corresponding to the target accent and the target timbre 's audio.

[0094]

In some optional implementations of this embodiment, the speech synthesis apparatus further includes a training unit not shown in FIG. 5, and is configured to: acquire an initial neural network model; acquire a training sample set, wherein the The training samples include timbre, linguistic features, target text corresponding to the linguistic features, and annotated second acoustic features corresponding to the target text corresponding to the timbre, linguistic features, and linguistic features; the timbre of the training samples in the training sample set is , linguistic features and the target text corresponding to the linguistic features are used as the input of the initial neural network model, and the second acoustic feature corresponding to the target text corresponding to the input timbre, linguistic features and linguistic features is used as the expected output. The network model is trained, the trained initial neural network model is determined as the transformation model.

[0095]

In some optional implementations of this embodiment, the training sample set does not include the target timbre; the second acoustic feature includes a Mel spectrum corresponding to the target accent and the target timbre; and the converting unit 505 is further configured to: according to the Mel spectrum Spectral and preset neural network vocoders to synthesize audio corresponding to the target text with the target accent and the target timbre.

[0096]

It should be understood that the units 501 to 505 described in the speech synthesis apparatus 500 respectively correspond to the respective steps in the method described with reference to FIG. 2 .

Therefore, the operations and features described above with respect to the speech synthesis method are also applicable to the apparatus 500 and the units included therein, and details are not repeated here.

[0097]

According to the embodiments of the present application, the present application further provides an electronic device and a readable storage medium.

[0098]

As shown in FIG. 6 , it is a block diagram of an electronic device according to a speech synthesis method according to an embodiment of the present application.

IPR2026-00272
Krisp Technologies EX1002 Page 311

Electronic devices are intended to represent various forms of digital computers, such as laptops, desktops, workstations, personal digital assistants, servers, blade servers, mainframe computers, and other suitable computers.

Electronic devices may also represent various forms of mobile devices, such as personal digital processors, cellular phones, smart phones, wearable devices, and other similar computing devices.

The components shown herein, their connections and relationships, and their functions are by way of example only, and are not intended to limit implementations of the application described and/or claimed herein.

[0099]

As shown in FIG. 6, the electronic device includes: one or more processors 601, a memory 602, and interfaces for connecting various components, including a high-speed interface and a low-speed interface.

The various components are interconnected using different buses 605 and may be mounted on a common motherboard or otherwise as desired.

The processor may process instructions executed within the electronic device, including instructions stored in or on memory to display graphical information of the GUI on an external input/output device, such as a display device coupled to the interface.

In other embodiments, multiple processors and/or multiple buses 605 may be used with multiple memories and multiple memories, if desired.

Likewise, multiple electronic devices may be connected, each providing some of the necessary operations (eg, as a server array, a group of blade servers, or a multiprocessor system).

A processor 601 is taken as an example in FIG. 6 .

[0100]

The memory 602 is the non-transitory computer-readable storage medium provided by the present application.

The memory stores instructions executable by at least one processor, so that the at least one processor executes the speech synthesis method provided by the present application.

The non-transitory computer-readable storage medium of the present application stores computer instructions for causing a computer to execute the speech synthesis method provided by the present application.

[0101]

As a non-transitory computer-readable storage medium, the memory 602 can be used to store non-transitory software programs, non-transitory computer-executable programs, and units, such as program instructions/units corresponding to the speech synthesis method in the embodiments of the present application (for example, appendix). The acquisition unit 501, the first acoustic feature determination unit 502, the target timbre determination unit 503, the linguistic feature determination unit 504, and the conversion unit 505 shown in FIG. 5).

The processor 601 executes various functional applications and data processing of the server by running the non-transitory software programs, instructions and modules stored in the memory 602, ie, implements the speech synthesis method in the above method embodiments.

[0102]

The memory 602 may include a storage program area and a storage data area, wherein the storage program area may store an operating system, an application program required for at least one function; the storage data area may store data created according to the use of the speech synthesis electronic device, and the like.

Additionally, memory 602 may include high-speed random access memory, and may also include non-transitory memory, such as at least one magnetic disk storage device, flash memory device, or other non-transitory solid state storage device.

In some embodiments, the memory 602 may optionally include memory located remotely from the processor 601, and these remote memories may be connected to the speech synthesis electronics via a network.

IPR2026-00272
Krisp Technologies EX1002 Page 312

Examples of such networks include, but are not limited to, the Internet, an intranet, a local area network, a mobile communication network, and combinations thereof.

[0103]

The electronic device for speech synthesis may further include: an input device 603 and an output device 604 .

The processor 601, the memory 602, the input device 603, and the output device 604 may be connected through a bus 605 or other means, and the connection through the bus 605 is taken as an example in FIG. 6 .

[0104]

The input device 603 can receive input numerical or character information, and generate key signal input related to user settings and function control of the speech synthesis electronic device, such as a touch screen, a keypad, a mouse, a trackpad, a touchpad, a pointing stick, an or Multiple input devices such as mouse buttons, trackballs, joysticks, etc.

Output devices 604 may include display devices, auxiliary lighting devices (eg, LEDs), haptic feedback devices (eg, vibration motors), and the like.

The display device may include, but is not limited to, a liquid crystal display (LCD), a light emitting diode (LED) display, and a plasma display.

In some implementations, the display device may be a touch screen.

[0105]

Various implementations of the systems and techniques described herein can be implemented in digital electronic circuitry, integrated circuit systems, application specific ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof.

These various embodiments may include being implemented in one or more computer programs executable and/or interpretable on a programmable system including at least one programmable processor that The processor, which may be a special purpose or general-purpose programmable processor, may receive data and instructions from a storage system, at least one input device, and at least one output device, and transmit data and instructions to the storage system, the at least one input device, and the at least one output device an output device.

[0106]

These computational programs (also referred to as programs, software, software applications, or codes) include machine instructions for programmable processors, and may be implemented using high-level procedural and/or object-oriented programming languages, and/or assembly/machine languages calculation program.

As used herein, the terms "machine-readable medium" and "computer-readable medium" refer to any computer program product, apparatus, and/or apparatus for providing machine instructions and/or data to a programmable processor (For example, magnetic disks, optical disks, memories, programmable logic devices (PLDs), including machine-readable media that receive machine instructions as machine-readable signals.

The term "machine-readable signal" refers to any signal used to provide machine instructions and/or data to a programmable processor.

[0107]

To provide interaction with a user, the systems and techniques described herein may be implemented on a computer having a display device (eg, a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user); and a keyboard and pointing device (eg, a mouse or trackball) through which a user can provide input to the computer.

Other kinds of devices can also be used to provide interaction with the user; for example, the feedback provided to the user can be any form of sensory feedback (eg, visual feedback, auditory feedback, or tactile feedback); and can be in any form (including acoustic input, voice input, or tactile input) to receive input from the user.

IPR2026-00272
Krisp Technologies EX1002 Page 313

[0108]

The systems and techniques described herein may be implemented on a computing system that includes back-end components (eg, as a data server), or a computing system that includes middleware components (eg, an application server), or a computing system that includes front-end components (eg, a user's computer having a graphical user interface or web browser through which a user may interact with implementations of the systems and techniques described herein), or including such backend components, middleware components, Or any combination of front-end components in a computing system.

The components of the system may be interconnected by any form or medium of digital data communication (eg, a communication network).

Examples of communication networks include: Local Area Networks (LANs), Wide Area Networks (WANs), and the Internet.

[0109]

A computer system can include clients and servers.

Clients and servers are generally remote from each other and usually interact through a communication network.

The relationship of client and server arises by computer programs running on the respective computers and having a client-server relationship to each other.

[0110]

According to the technical solutions of the embodiments of the present application, through the linguistic features determined according to the acquired target text and the target timbre determined by the acquired target audio, the target audio corresponding to the target text with the target accent and the target timbre can be synthesized accurately and quickly .

[0111]

According to one or more embodiments of the present disclosure, the present disclosure provides a speech synthesis method, including: acquiring target text and target audio; determining a first acoustic feature for indicating a target accent based on the target text; based on the target audio , determine the target timbre; determine the linguistic feature in the first acoustic feature; based on the target timbre, the linguistic feature and the target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre 's audio.

[0112]

According to one or more embodiments of the present disclosure, determining, based on the target text, the first acoustic feature used to indicate the target accent includes: according to the target text and a pretrained speech synthesis model, determining the corresponding target text and used to indicate the target The first acoustic feature of the accent, wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0113]

According to one or more embodiments of the present disclosure, determining a target timbre based on the target audio includes: determining an identity vector corresponding to the target audio according to the target audio and a pre-trained authentication model, wherein the pre-trained authentication model is used for Characterize the correspondence between the audio and the identity vector; according to the identity vector, determine the target timbre.

[0114]

According to one or more embodiments of the present disclosure, determining the linguistic feature in the first acoustic feature includes: extracting the linguistic feature in the first acoustic feature by using a pretrained recognition model, wherein the pretrained recognition model Used to characterize the correspondence between the first acoustic feature and the linguistic feature.

[0115]

IPR2026-00272
Krisp Technologies EX1002 Page 314

According to one or more embodiments of the present disclosure, based on the target timbre, linguistic features, and target text, determining the audio corresponding to the target text with the target accent and the target timbre includes: according to the target timbre, linguistic features, target text, and preset The conversion model of the training determines the second acoustic feature corresponding to the target text for generating the audio with the target accent and the target timbre, wherein the conversion model is used to characterize the correspondence between the timbre, the linguistic feature and the text and the second acoustic feature; Based on the second acoustic feature, the audio corresponding to the target text with the target accent and the target timbre is determined.

[0116]

According to one or more embodiments of the present disclosure, the speech synthesis method further includes: obtaining an initial neural network model; obtaining a training sample set, wherein the training samples in the training sample set include timbre, linguistic features, and targets corresponding to the linguistic features The text and the marked second acoustic feature corresponding to the target text corresponding to the timbre, linguistic features and linguistic features; the target text corresponding to the timbre, linguistic features and linguistic features of the training samples in the training sample set is used as the initial neural As the input of the network model, the second acoustic feature corresponding to the target text corresponding to the input timbre, linguistic feature and linguistic feature is used as the expected output to train the initial neural network model; the initial neural network model after training is determined as Convert the model.

[0117]

According to one or more embodiments of the present disclosure, the training sample set does not include the target timbre; the second acoustic feature includes a mel spectrum corresponding to the target accent and the target timbre; and based on the second acoustic feature, it is determined that the target text has a target Accent and target timbre audio, including: synthesizing audio with target accent and target timbre corresponding to the target text according to the Mel spectrum and a preset neural network vocoder.

[0118]

According to one or more embodiments of the present disclosure, the present disclosure provides a speech synthesis apparatus, comprising: an acquisition unit configured to acquire target text and target audio; a first acoustic feature determination unit configured to be based on the target text , determine the first acoustic feature used to indicate the target accent; the target timbre determination unit is configured to determine the target timbre based on the target audio; the linguistic feature determination unit is configured to determine the linguistic feature in the first acoustic feature The conversion unit is configured to determine the audio with the target accent and the target timbre corresponding to the target text based on the target timbre, the linguistic feature and the target text, and output the audio with the target accent and the target timbre.

[0119]

According to one or more embodiments of the present disclosure, the first acoustic feature determining unit is further configured to: determine, according to the target text and the pre-trained speech synthesis model, a first acoustic feature corresponding to the target text and used to indicate the target accent , wherein the pre-trained speech synthesis model is used to characterize the correspondence between the spoken text and the first acoustic feature.

[0120]

According to one or more embodiments of the present disclosure, the target timbre determination unit is further configured to: determine an identity vector corresponding to the target audio according to the target audio and a pre-trained authentication model, wherein the pre-trained authentication model is used to represent Correspondence between the audio and the identity vector; according to the identity vector, determine the target timbre.

[0121]

IPR2026-00272
Krisp Technologies EX1002 Page 315

According to one or more embodiments of the present disclosure, the linguistic feature determination unit is further configured to: extract linguistic features in the first acoustic feature using a pretrained recognition model, wherein the pretrained recognition model is used to characterize the first acoustic feature. A correspondence between acoustic features and linguistic features.

[0122]

According to one or more embodiments of the present disclosure, the conversion unit is further configured to: determine, according to the target timbre, the linguistic feature, the target text and the pre-trained conversion model, the corresponding target text for generating the target accent and the target timbre. The second acoustic feature of the audio, wherein the conversion model is used to represent the timbre, the linguistic feature, and the correspondence between the text and the second acoustic feature; based on the second acoustic feature, the audio corresponding to the target text with the target accent and the target timbre is determined.

[0123]

According to one or more embodiments of the present disclosure, the speech synthesis apparatus further includes a training unit not shown in FIG. 5, configured to: obtain an initial neural network model; obtain a training sample set, wherein the training samples in the training sample set including the timbre, linguistic features, the target text corresponding to the linguistic features, and the marked second acoustic feature corresponding to the target text corresponding to the timbre, linguistic features, and linguistic features; the timbre, language of the training samples in the training sample set The target text corresponding to the linguistic feature and the linguistic feature is used as the input of the initial neural network model, and the second acoustic feature corresponding to the target text corresponding to the input timbre, linguistic feature and linguistic feature is used as the expected output. Perform training; determine the initial neural network model after training as the transformation model.

[0124]

According to one or more embodiments of the present disclosure, the training sample set does not include the target timbre; the second acoustic feature includes a mel spectrum corresponding to the target accent and the target timbre; and the converting unit is further configured to: according to the mel spectrum and the pre-determined timbre The designed neural network vocoder synthesizes the audio corresponding to the target text with the target accent and the target timbre.

[0125]

It should be understood that the above embodiments are only exemplary embodiments, but are not limited thereto, and also include other methods known in the art that can implement speech synthesis.

Steps can be reordered, added, or removed using the various forms of flow shown above.

For example, the steps described in the present application can be executed in parallel, sequentially or in different orders, as long as the desired results of the technical solutions disclosed in the present application can be achieved, no limitation is imposed herein.

[0126]

The above-mentioned specific embodiments do not constitute a limitation on the protection scope of the present application.

It should be understood by those skilled in the art that various modifications, combinations, sub-combinations and substitutions may occur depending on design requirements and other factors.

Any modifications, equivalent replacements and improvements made within the spirit and principles of this application shall be included within the protection scope of this application.

IPR2026-00272
Krisp Technologies EX1002 Page 316

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

CLAIMS CN112382270A

1.

A speech synthesis method, comprising:

Get target text and target audio;

Based on the target text, determining a first acoustic feature indicative of a target accent;

Based on the target audio, determine the target timbre;

determining linguistic features in the first acoustic features;

Based on the target timbre, the linguistic feature and the target text, the audio with the target accent and the target timbre corresponding to the target text is determined, and the audio with the target accent and the target timbre is output.

2.

The method of claim 1, wherein the determining, based on the target text, a first acoustic feature indicative of a target accent comprises:

According to the target text and a pre-trained speech synthesis model, a first acoustic feature corresponding to the target text and used to indicate the target accent is determined, wherein the pre-trained speech synthesis model is used to characterize the difference between the spoken text and the first acoustic feature. Correspondence of acoustic features.

3.

The method according to claim 1, wherein the determining a target timbre based on the target audio comprises:

Determine the identity vector corresponding to the target audio according to the target audio and a pre-trained authentication model, wherein the pre-trained authentication model is used to represent the correspondence between the audio and the identity vector;

According to the identity vector, the target timbre is determined.

4.

The method of claim 1, wherein said determining a linguistic feature in the first acoustic feature comprises:

The linguistic features in the first acoustic features are extracted by using a pre-trained recognition model, wherein the pre-trained recognition model is used to represent the correspondence between the first acoustic features and the linguistic features.

IPR2026-00272
Krisp Technologies EX1002 Page 317

5.

The method according to claim 1, wherein the determining, based on the target timbre, the linguistic feature and the target text, the audio with the target accent and the target timbre corresponding to the target text, comprising:

According to the target timbre, the linguistic feature, the target text and the pre-trained conversion model, a second acoustic feature corresponding to the target text for generating audio with the target accent and target timbre is determined, wherein the The transformation model described above is used to characterize the timbre, linguistic features, and the correspondence between text and second acoustic features;

Based on the second acoustic feature, the audio corresponding to the target text with the target accent and the target timbre is determined.

6.

The method of claim 5, wherein the method further comprises:

Get the initial neural network model;

Obtain a training sample set, wherein the training samples in the training sample set include timbre, linguistic features, target texts corresponding to linguistic features, and marked timbres, linguistic features, and linguistic features corresponding to the linguistic features The second acoustic feature corresponding to the target text of ;

The target text corresponding to the timbre, linguistic features and linguistic features of the training samples in the training sample set is used as the input of the initial neural network model, and the target text corresponding to the input timbre, linguistic features and linguistic features The second acoustic feature corresponding to the text is used as the desired output, and the initial neural network model is trained;

The trained initial neural network model is determined as the conversion model.

7.

The method of claim 6, wherein the set of training samples does not include the target timbre; the second acoustic feature comprises a mel spectrum corresponding to the target accent and the target timbre; and

The determining, based on the second acoustic feature, the audio with the target accent and the target timbre corresponding to the target text, including:

According to the Mel spectrum and the preset neural network vocoder, the audio corresponding to the target text with the target accent and the target timbre is synthesized.

8.

A speech synthesis device, comprising:

an acquisition unit, configured to acquire target text and target audio;

a first acoustic feature determination unit configured to determine, based on the target text, a first acoustic feature indicative of a target accent;

a target timbre determining unit, configured to determine a target timbre based on the target audio;

a linguistic feature determination unit configured to determine a linguistic feature in the first acoustic feature;

The conversion unit is configured to, based on the target timbre, the linguistic feature and the target text, determine the audio with the target accent and the target timbre corresponding to the target text, and output the audio with the target accent and the target timbre audio.

9.

An electronic device for speech synthesis, comprising:

IPR2026-00272
Krisp Technologies EX1002 Page 318

at least one processor; and

a memory communicatively coupled to the at least one processor; wherein,

The memory stores instructions executable by the at least one processor, the instructions being executed by the at least one processor to enable the at least one processor to perform the execution of any of claims 1-7 Methods.

10.

A non-transitory computer-readable storage medium storing computer instructions, wherein the computer instructions are used to cause the computer to perform the method of any one of claims 1-7.

IPR2026-00272
Krisp Technologies EX1002 Page 319

(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112382270 A
(43)申请公布日 2021.02.19

(21)申请号 202011273128.7

(22)申请日 2020.11.13

(71)申请人 北京有竹居网络技术有限公司
地址 101299 北京市平谷区林荫北街13号信息大厦802室

(72)发明人 汤本来 李忠豪

(74)专利代理机构 北京海智友知识产权代理事务所(普通合伙) 11455
代理人 吴京顺

(51)Int.Cl.
G10L 13/02(2013.01)
G10L 13/08(2013.01)
G10L 25/03(2013.01)
G10L 25/30(2013.01)

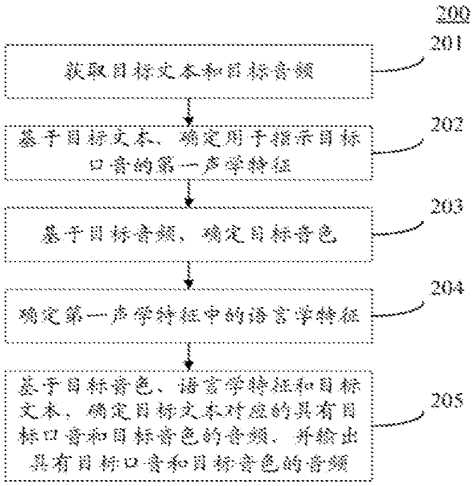
权利要求书2页 说明书13页 附图4页

(54)发明名称

语音合成方法、装置、设备以及存储介质

(57)摘要

本申请公开了一种语音合成方法、装置、设备以及存储介质，涉及语音合成、自然语言处理、计算机技术、人工智能、深度学习领域。具体实现方案为：获取目标文本和目标音频；基于目标文本，确定用于指示目标口音的第一声学特征；基于目标音频，确定目标音色；确定第一声学特征中的语言学特征；基于目标音色、语言学特征和目标文本，确定目标文本对应的目标具有目标口音的音频，并输出目标音频。本实现方式通过根据获取的目标文本确定的语言学特征以及通过获取的目标音频确定的目标音色，可以准确、快速地合成目标文本对应的具有目标口音和目标音色的目标音频。



CN 112382270 A

IPR2026-00272
Krisp Technologies EX1002 Page 320

1.一种语音合成方法，包括：

获取目标文本和目标音频；

基于所述目标文本，确定用于指示目标口音的第一声学特征；

基于所述目标音频，确定目标音色；

确定所述第一声学特征中的语言学特征；

基于所述目标音色、所述语言学特征和所述目标文本，确定所述目标文本对应的具有目标口音和目标音色的音频，并输出所述具有目标口音和目标音色的音频。

2.根据权利要求1所述的方法，其中，所述基于所述目标文本，确定用于指示目标口音的第一声学特征，包括：

根据所述目标文本和预训练的语音合成模型，确定所述目标文本对应的用于指示目标口音的第一声学特征，其中，所述预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

3.根据权利要求1所述的方法，其中，所述基于所述目标音频，确定目标音色，包括：

根据所述目标音频和预训练的身份验证模型，确定所述目标音频对应的身份向量，其中，所述预训练的身份验证模型用于表征音频与身份向量之间的对应关系；

根据所述身份向量，确定目标音色。

4.根据权利要求1所述的方法，其中，所述确定所述第一声学特征中的语言学特征，包括：

利用预训练的识别模型提取所述第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

5.根据权利要求1所述的方法，其中，所述基于所述目标音色、所述语言学特征和所述目标文本，确定所述目标文本对应的具有目标口音和目标音色的音频，包括：

根据所述目标音色、所述语言学特征、所述目标文本和预训练的转换模型，确定所述目标文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征，其中，所述转换模型用于表征音色、语言学特征和文本与第二声学特征的对应关系；

基于所述第二声学特征，确定所述目标文本对应的具有目标口音和目标音色的音频。

6.根据权利要求5所述的方法，其中，所述方法还包括：

获取初始神经网络模型；

获取训练样本集合，其中，所述训练样本集合中的训练样本包括音色、语言学特征、语言学特征对应的目标文本以及标注的与所述音色、所述语言学特征和所述语言学特征对应的目标文本对应的第二声学特征；

将所述训练样本集合中的训练样本的音色、语言学特征和语言学特征对应的目标文本作为所述初始神经网络模型的输入，将与输入的音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征作为期望输出，对所述初始神经网络模型进行训练；

将训练后的所述初始神经网络模型确定为所述转换模型。

7.根据权利要求6所述的方法，其中，所述训练样本集合不包括所述目标音色；所述第二声学特征包括对应于所述目标口音和所述目标音色的梅尔频谱；以及

所述基于所述第二声学特征，确定所述目标文本对应的具有目标口音和目标音色的音频，包括：

IPR2026-00272
Krisp Technologies EX1002 Page 321

根据所述梅尔频谱和预设的神经网络声码器，合成所述目标文本对应的具有目标口音和目标音色的音频。

8.一种语音合成装置，包括：

获取单元，被配置成获取目标文本和目标音频；

第一声学特征确定单元，被配置成基于所述目标文本，确定用于指示目标口音的第一声学特征；

目标音色确定单元，被配置成基于所述目标音频，确定目标音色；

语言学特征确定单元，被配置成确定所述第一声学特征中的语言学特征；

转换单元，被配置成基于所述目标音色、所述语言学特征和所述目标文本，确定所述目标文本对应的具有目标口音和目标音色的音频，并输出所述具有目标口音和目标音色的音频。

9.一种语音合成电子设备，其特征在于，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-7中任一项所述的方法。

10.一种存储有计算机指令的非瞬时计算机可读存储介质，其特征在于，所述计算机指令用于使所述计算机执行权利要求1-7中任一项所述的方法。

IPR2026-00272
Krisp Technologies EX1002 Page 322

语音合成方法、装置、设备以及存储介质

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及语音合成、自然语言处理、计算机技术、人工智能、深度学习领域，尤其涉及语音合成方法、装置、设备以及存储介质。

背景技术

[0002] 近年来，由于在线教育和在线学习的迅猛发展，语音合成技术得到了广泛的研究与关注，语音合成旨在于将某一用户的语音合成为不同口音或不同音色或既不同口音又不同音色等的音频。语音合成技术在娱乐方面也有很大的应用前景。利用现有的语音合成技术进行语音合成速度慢而且语音合成的结果往往不准确。

发明内容

[0003] 本公开提供了一种语音合成方法、装置、设备以及存储介质。

[0004] 根据本公开的一方面，提供了一种语音合成方法，包括：获取目标文本和目标音频；基于目标文本，确定用于指示目标口音的第一声学特征；基于目标音频，确定目标音色；确定第一声学特征中的语言学特征；基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0005] 根据本公开的另一方面，提供了一种语音合成装置，包括：获取单元，被配置成获取目标文本和目标音频；第一声学特征确定单元，被配置成基于目标文本，确定用于指示目标口音的第一声学特征；目标音色确定单元，被配置成基于目标音频，确定目标音色；语言学特征确定单元，被配置成确定第一声学特征中的语言学特征；转换单元，被配置成基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0006] 根据本公开的再一方面，提供了一种语音合成电子设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行如上述语音合成方法。

[0007] 根据本公开的又一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，上述计算机指令用于使计算机执行如上述语音合成方法。

[0008] 根据本申请的技术解决了无法准确、快速地进行语音合成的问题，通过根据获取的目标文本确定的语言学特征以及通过获取的目标音频确定的目标音色，可以准确、快速地合成目标文本对应的具有目标口音和目标音色的目标音频。

[0009] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0010] 附图用于更好地理解本方案，不构成对本申请的限定。其中：

[0011] 图1是本申请的一个实施例可以应用于其中的示例性系统架构图；

IPR2026-00272
Krisp Technologies EX1002 Page 323

[0012] 图2是根据本申请的语音合成方法的一个实施例的流程图；

[0013] 图3是根据本申请的语音合成方法的一个应用场景的示意图；

[0014] 图4是根据本申请的语音合成方法的另一个实施例的流程图；

[0015] 图5是根据本申请的语音合成装置的一个实施例的结构示意图；

[0016] 图6是用来实现本申请实施例的语音合成方法的电子设备的框图。

具体实施方式

[0017] 以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0018] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0019] 图1示出了可以应用本申请的语音合成方法或语音合成装置的实施例的示例性系统架构100。

[0020] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0021] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如语言合成类应用等。

[0022] 终端设备101、102、103可以是硬件，也可以是软件。当终端设备101、102、103为硬件时，可以是各种电子设备，包括但不限于智能手机、平板电脑、车载电脑、膝上型便携计算机和台式计算机等等。当终端设备101、102、103为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

[0023] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103采集的目标文本和目标音频进行处理的后台服务器。后台服务器可以获取目标文本和目标音频，并基于目标文本，确定用于指示目标口音的第一声学特征；基于目标音频，确定目标音色；确定第一声学特征中的语言学特征；基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出该音频。

[0024] 需要说明的是，服务器105可以是硬件，也可以是软件。当服务器105为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器105为软件时，可以实现成多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

[0025] 需要说明的是，本申请实施例所提供的语音合成方法一般由服务器105执行。相应地，语音合成装置一般设置于服务器105中。

[0026] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

IPR2026-00272
Krisp Technologies EX1002 Page 324

[0027] 继续参考图2，示出了根据本申请的语音合成方法的一个实施例的流程200。本实施例的语音合成方法，包括以下步骤：

[0028] 步骤201，获取目标文本和目标音频。

[0029] 本实施例中，语音合成方法的执行主体(例如图1中的服务器105)可以从本地获取目标文本，也可以通过有线连接或无线连接的方式获取终端设备通过录音或扫描的方式采集的目标文本，也可以获取用户在终端设备的编辑框中输入的任意一句话并将其确定为目标文本。具体地，目标文本可以是具有确定身份的任意一个想要转换口音的人所发出的语音转换成的说话文本。说话文本，可以是文字。比如，小明说了一句普通话“我爱祖国”，则目标文本就可以是以文字形式表示的“我爱祖国”。目标文本可以是中文也可以是英文，本申请对目标文本的语言类型不做具体限定。目标音频可以是需要转换成的音色对应的音频。例如，可以是录有小A同学的声音的音频，也可以是录有小B同学的声音的音频，本申请对目标音频的音色不做具体限定。该目标音频可以是MP3形式，也可以是MP4形式，本申请对目标音频的存储形式不做限定。可以理解的是，目标音频可以是人的音频，也可以是自然界其他生物的音频，本申请对目标音频的来源不做具体限定。

[0030] 步骤202，基于目标文本，确定用于指示目标口音的第一声学特征。

[0031] 执行主体在获取目标文本后，可以基于目标文本，确定用于指示目标口音的第一声学特征。具体地，第一声学特征可以是用于表征目标口音的语音特征参数，例如可以是梅尔频谱。本实施例中，执行主体可以基于目标文本，根据目标文本和其对应的音频，对音频进行预加重、分帧和加窗，然后可以对每帧音频的信号进行短时傅里叶变换(STFT，short-time Fourier transform，或short-term Fourier transform)，得到短时幅度谱；短时幅度谱通过梅尔滤波器组得到梅尔频谱。本申请对梅尔频谱(即第一声学特征)的获取方式不做具体限定。

[0032] 步骤203，基于目标音频，确定目标音色。

[0033] 执行主体在获取目标音频后，可以基于目标音频，确定目标音色。音频中包含多个音素。音素，是根据语音的自然属性划分出来的最小语音单位，依据音节里的发音动作来分析，一个动作构成一个音素。音素分为元音和辅音两大类。例如，如汉语音节啊(ā)只有一个音素，爱(ài)有两个音素，代(dài)有三个音素等。如[ma-mi]中，两个[m]发音动作相同，是相同音素，[a][i]发音动作不同，是不同音素。音色(Timbre)是指不同声音表现在波形方面总是有与众不同的特性，不同的物体振动都有不同的特点。不同的发声体由于其材料、结构不同，则发出声音的音色也不同。例如钢琴、小提琴和人发出的声音不一样，每一个人发出的声音也不一样。因此，可以把音色理解为声音的特征。音色是声音的属性(即响度、音调、音色)之一，主要由其泛音决定。每个人的声音以及各种乐器所发出的声音的区别，就是由音色不同造成的。发音体的振动是由多种泛音组成的，其中有基音和泛音，泛音不同，决定着某一个特定的音色。声音除了有一个‘基音’外，还自然而然加上许多不同‘频率’(振动的物体1秒钟振动的次数)与泛音‘交织’，就决定了不同的音色，使人听了以后能辨别出是不同的声音。具体地，执行主体可以根据目标音频和预训练的分类模型，确定目标音频中各音素对应的标识，其中，预训练的分类模型用于表征音素与标识的对应关系；根据确定出的各音素对应的标识，确定目标音频中的基音和泛音；根据确定出的目标音频中的基音和泛音，确定目标音色。具体地，执行主体可以将确定出的目标音频中的基音和泛音输入至与预训

IPR2026-00272
Krisp Technologies EX1002 Page 325

练的音色转换模型中，输出该基音和泛音对应的音色，其中，预训练的音色转换模型用于表征基音、泛音与音色的对应关系。示例的，该音色转换模型可以是预训练的卷积神经网络(Convolutional Neural Networks,CNN)。

[0034] 步骤204，确定第一声学特征中的语言学特征。

[0035] 执行主体在确定第一声学特征后，可以确定第一声学特征中的语言学特征。具体地，语言学特征可以包括韵律特征、句法、语篇结构、信息结构等。其中，韵律特征又可以是超音质特征或超音段特征，是语言的一种音系结构。韵律特征可以分为三个主要方面：语调、时域分布和重音，通过超音段特征实现。超音段特征包括音高、强度以及时间特性，由音位或音位群负载。韵律是人类自然语言的一个典型特征，具有许多跨语言的共同特点，比如：音高下倾、重读、停顿等都普遍存在于不同的语言之中。韵律特征是语言和情绪表达的重要形式之一。执行主体可以用现有的已知的语言学特征跟第一声学特征中的各特征对比，将第一声学特征中与已知的语言学特征的相似度大于阈值的特征确定为第一声学特征中的语言学特征。具体地，执行主体可以分别将第一声学特征中的各特征转化为第一声学特征向量，将现有的已知的语言学特征转化为语言学特征向量。将各第一声学特征向量与各语言学特征向量进行余弦相似度计算，余弦相似度越趋近于1，表明参与计算该余弦相似度的第一声学特征向量和对应的语言学特征向量越相似，可以将该第一声学特征向量确定为语言学特征向量，相应地，将该第一声学特征向量对应的第一声学特征确定为该语言学特征向量对应的语言学特征。同理，遍历每个第一声学特征和每个语言学特征，直至找到每个第一声学特征和每个语言学特征各自对应的向量的余弦相似度与1的差值的绝对值小于阈值的第一声学特征向量对应的第一声学特征，并将该第一声学特征确定为语言学特征，据此可以得到第一声学特征中的所有语言学特征。

[0036] 步骤205，基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0037] 执行主体在确定出语言学特征后，可以基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音的音频，并输出音频。具体地，执行主体可以将目标音色、语言学特征和目标文本输入预训练的语音模型以生成与目标文本对应的具有目标口音和目标音色的音频，并输出该音频。其中，该预训练的语音模型用于表征目标音色、语言学特征和目标文本与具有目标口音和目标音色的音频的对应关系，具体可以是预训练的神经网络模型。在对语音模型进行训练时，可以首先获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括目标音色、语言学特征、语言学特征对应的目标文本以及标注的与目标音色、语言学特征和语言学特征对应的目标文本对应的具有目标口音和目标音色的音频；将训练样本集合中的训练样本的目标音色、语言学特征和语言学特征对应的目标文本作为初始神经网络模型的输入，将与输入的语言学特征和语言学特征对应的目标文本对应的具有目标口音和目标音色的音频作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为上述语音模型。示例的，目标文本可以为小B的普通话音频对应的文本“我爱祖国”，目标文本对应的具有目标口音和目标音色的音频，可以是小A的东北口音的“我爱祖国”的音频，该音频可以是MP3形式或者也可以是MP4形式，本申请对输出音频的形式不做具体限定。

[0038] 继续参考图3，其示出了根据本申请的语音合成方法的一个应用场景的示意图。在

IPR2026-00272
Krisp Technologies EX1002 Page 326

图3的应用场景中，服务器304通过网络303获取目标文本301和目标音频302。服务器304基于目标文本301，确定用于指示目标口音的第一声学特征305。服务器304基于目标音频302确定目标音色307。服务器304确定第一声学特征305中的语言学特征306。服务器304基于目标音色307、语言学特征306和目标文本301，确定目标文本301对应的具有目标口音和目标音色的音频308，并输出具有目标口音和目标音色的音频308。本申请实际上是目标说话文本向具有任意口音的音频的转换，文本内容不变。

[0039] 本实施例通过根据获取的目标文本确定的语言学特征以及通过获取的目标音频确定的目标音色，可以准确、快速地合成定目标文本对应的具有目标口音和目标音色的目标音频。

[0040] 继续参考图4，其示出了根据本申请的语音合成方法的另一个实施例的流程400。如图4所示，本实施例的语音合成方法，可以包括以下步骤：

[0041] 步骤401，获取目标文本和目标音频。

[0042] 步骤402，基于目标文本，确定用于指示目标口音的第一声学特征。

[0043] 步骤401～步骤402的原理与步骤201～步骤202的原理类似，此处不再赘述。

[0044] 具体地，步骤402可以通过步骤4021实现：

[0045] 步骤4021，根据目标文本和预训练的语音合成模型，确定目标文本对应的用于指示目标口音的第一声学特征。

[0046] 本实施例中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。执行主体在得到目标文本后，可以根据目标文本和预训练的语音合成模型，确定目标文本对应的用于指示目标口音的第一声学特征。具体地，第一声学特征可以是用于指示目标口音(比如东北口音)的梅尔频谱。经过语音合成模型得到的目标文本对应的用于指示目标口音的第一声学特征此时还不能用于目标口音的输出，还需要进一步的特征识别和特征转换得到携带有较多目标文本信息的特征，比如，音高、音强等。然后根据得到的这些携带有较多目标文本信息的特征进行进一步的语音合成。本实施例，具体地，执行主体可以将目标文本输入到预训练的语音合成模型中，由预训练的语音合成模型对目标文本进行第一声学特征的提取，得到目标文本对应的用于指示目标口音的第一声学特征。

[0047] 本实施例通过根据目标文本和预训练的语音合成模型，可以得到具有目标口音的音频梅尔频谱，从而可以基于得到的具有目标口音的音频梅尔频谱准确地对任意一个用户的说话文本进行目标口音的转换。

[0048] 步骤403，基于目标音频，确定目标音色。

[0049] 步骤403的原理与步骤203的原理类似，此处不再赘述。

[0050] 具体地，步骤403可以通过步骤4031～步骤4032实现：

[0051] 步骤4031，根据目标音频和预训练的身份验证模型，确定目标音频对应的身份向量。

[0052] 本实施例，预训练的身份验证模型用于表征音频与身份向量之间的对应关系。执行主体在获取目标音频后，可以根据目标音频和预训练的身份验证模型，确定目标音频对应的身份向量。具体地，执行主体可以将目标音频输入预训练的身份验证模型，输出目标音频对应的身份向量。具体地，输出的目标音频对应的身份向量可以是一组多维的数据，可以标识目标音频对应的说话人的音色信息，例如，身份向量可以是[0.3,0.3,0.5,0.6,…]这

IPR2026-00272
Krisp Technologies EX1002 Page 327

样的数据序列对应的向量，该数据序列中的一个或多个数据的组合可以用来表征唯一的音色。

[0053] 步骤4032，根据身份向量，确定目标音色。

[0054] 执行主体在确定目标音频对应的身份向量后，可以根据身份向量，确定目标音色。具体地，执行主体可以根据身份向量和预设的身份向量与音色的对应关系，确定该身份向量对应的目标音色。作为另一种实现方式，执行主体还可以根据身份向量与各已有的音色向量的相似度，确定该身份向量对应的音色向量，并将该音色向量对应的音色确定为目标音色。具体地，执行主体响应于确定身份向量与对应音色向量的相似度大于预设阈值，将该身份向量对应的音色向量所对应的音色确定为目标音色。

[0055] 本实施例通过由身份验证模型得到的身份向量确定目标音色，可以不受模型训练样本集的限制，可以将目标文本转换成任意想转换的音色，提高音色转换的灵活性，提升用户体验。

[0056] 步骤404，确定第一声学特征中的语言学特征。

[0057] 步骤404的原理与步骤204的原理类似，此处不再赘述。

[0058] 具体地，步骤404可以通过步骤4041实现：

[0059] 步骤4041，利用预训练的识别模型提取第一声学特征中的语言学特征。

[0060] 其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。执行主体在得到第一声学特征后，可以利用预训练的识别模型提取第一声学特征中的语言学特征。具体地，第一声学特征中的语言学特征可以包括语调、时域分布、重音、音高、重读、停顿等韵律特征。语言学特征是语言和情绪表达的重要形式之一。

[0061] 在本实施例的一些可选的实现方式中，执行主体还可以根据第一声学特征和预训练的识别模型，确定第一声学特征对应的类别标识，其中，预训练的识别模型用于表征第一声学特征和类别标识的对应关系。得到的类别标识可以是用于表征第一声学特征中各音素的类别的标识，例如，第一声学特征中的各音素可以是语调音素、时域分布音素、重音音素、音高音素、重读音素、停顿音素，例如可以分别用标识1、2、3、4、5、6、7来表示。然后，然后执行主体可以根据第一声学特征中与得到的各标识对应的各音素以及预设的标识、音素与第二声学特征之间的对应关系，确定用于生成具有目标口音的音频的第二声学特征。第二声学特征可以是用于生成目标口音所需的各音素对应的梅尔频谱。执行主体可以基于该第二声学特征，确定与目标文本对应的具有目标口音和目标音色的音频，并输出该音频。本实现方式可以丰富用于生成目标口音和目标音色的音频所需的梅尔频谱，并提高生成目标口音和目标音色的音频的准确性。

[0062] 本实施例通过利用预训练的识别模型提取第一声学特征中的语言学特征，可以将目标文本中对应语言和情绪表达的相关特征提取出来，以完善用于生成目标口音和目标音色的音频的特征，提高生成具有目标口音和目标音色的音频的准确性。

[0063] 步骤405，基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0064] 步骤405的原理与步骤205的原理类似，此处不再赘述。

[0065] 具体地，步骤405可以通过步骤4051～步骤4052实现：

[0066] 步骤4051，根据目标音色、语言学特征、目标文本和预训练的转换模型，确定目标

IPR2026-00272
Krisp Technologies EX1002 Page 328

文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征。

[0067] 其中，转换模型用于表征音色、语言学特征和文本与第二声学特征的对应关系。具体地，第二声学特征可以为上述目标音色、语言学特征对应的梅尔频谱。具体地，第二声学特征可以是用于生成目标口音和目标音色的音频所需的梅尔频谱。执行主体可以将目标音色、语言学特征和目标文本输入预训练的转换模型，输出目标文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征。

[0068] 步骤4052，基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频。

[0069] 执行主体在得到第二声学特征后，可以基于第二声学特征，确定具有目标口音和目标音色的音频。具体地，执行主体可以将得到的第二声学特征输入到声码器，经过声码器的转换得到具有目标口音和目标音色的音频。声码器在其发送端对第二声学特征进行编码和加密，以取得和信道的匹配，经信息通道传递到声码器的接受端，在频域中对接收到的特征进行分析，鉴别清浊音，测定浊音基频，进而选取清-浊判断、浊音基频和频谱包络作为特征参量加以传送。当然，该分析也可以在时域中进行，利用其周期性提取一些第二声学特征进行线性预测，生成与第二声学特征对应的具有目标口音和目标音色的音频。具体地，上述声码器可以包括通道式声码器、共振峰声码器、图案声码器、线性预测声码器、相关声码器和正交函数声码器，本申请对声码器的类型不做具体限定。

[0070] 本实施例通过根据目标音色、语言学特征、目标文本和预训练的转换模型，确定目标文本对应的第二声学特征，并基于第二声学特征确定具有目标口音和目标音色的音频，可以使得针对于任意用户的任意一个说话文本，可以不受转换模型的训练样本的限制，将目标文本转换成任意想转换成的音色和口音的音频，提升语音合成的趣味性，提升用户体验。

[0071] 在本实施例的一些可选的实现方式中，语音合成方法还包括图4中未示出的以下模型训练步骤：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括音色、语言学特征、语言学特征对应的目标文本以及标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征；将训练样本集合中的训练样本的音色、语言学特征和语言学特征对应的目标文本作为初始神经网络模型的输入，将与输入的音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0072] 本实施例中，执行主体可以通过有线连接方式或者无线连接方式获取初始神经网络模型。上述初始神经网络模型可以包括各种包含隐藏层的人工神经网络(Artificial Neural Network，ANN)。在本实施例中，上述执行主体还可以从本地获取预先存储的初始模型，也可以从通信连接的电子设备获取上述初始模型，在此不作限定。

[0073] 本实施例中，执行主体可以通过各种方式获取训练样本集合。具体地，训练样本集合中的训练样本可以包括音色、语言学特征、语言学特征对应的目标文本以及标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征。其中，训练样本中标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的、也可以是人工实时标注的、也可以是先通过自动标注后，又人工补充修改纠正标注错误后得到的，本申请对此不做具体

IPR2026-00272
Krisp Technologies EX1002 Page 329

限定。训练样本中的音色可以是从本地或通信连接的电子设备上获取的。训练样本中的语言学特征可以是实时提取的，也可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的。训练样本中的与语言学特征对应的目标文本可以是对不同的用户实时采集的，也可以是通过有线或无线连接的方式从本地或通信连接的电子设备上获取的，本申请对此不做具体限定。

[0074] 本实施例通过获取训练样本集对初始神经网络模型进行训练，可以得到能够根据音色、语言学特征和说话文本准确地生成对应的第二声学特征的能力的转换模型，从而可以实现将获取的任意目标文本上的文字转换成具有目标口音和目标音色的音频，并可以提高生成的具有目标口音和目标音色的音频的质量，本申请可以不限于训练转换模型的训练样本集中已有的音色，对于其他的、训练样本集中没有的音色也可以精确合成带有该音色的音频，提升语音合成的趣味性，提升用户体验。

[0075] 在本实施例的一些可选的实现方式中，训练样本集合不包括目标音色；第二声学特征包括对应于目标口音和目标音色的梅尔频谱；以及基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频，包括：根据梅尔频谱和预设的神经网络声码器，合成目标文本对应的具有目标口音和目标音色的音频。

[0076] 具体地，执行主体在得到对应于目标口音和目标音色的梅尔频谱后，可以将该梅尔频谱自动输入预设的神经网络声码器中，以基于该神经网络声码器和该梅尔频谱合成具有目标口音和目标音色的音频。具有目标口音和目标音色的音频，包括以下至少一项：具有目标口音和目标音色的说话音频、具有目标口音和目标音色的歌唱音频。

[0077] 本实现方式通过根据对应于目标口音和目标音色的梅尔频谱和预设的神经网络声码器，可以使得合成的具有目标口音和目标音色的音频更精确。

[0078] 进一步参考图5，作为对上述各图所示方法的实现，本申请提供了一种语音合成装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0079] 如图5所示，本实施例的语音合成装置500包括：获取单元501、第一声学特征确定单元502、目标音色确定单元503、语言学特征确定单元504和转换单元505。

[0080] 获取单元501，被配置成获取目标文本和目标音频。

[0081] 第一声学特征确定单元502，被配置成基于目标文本，确定用于指示目标口音的第一声学特征。

[0082] 目标音色确定单元503，被配置成基于目标音频，确定目标音色。

[0083] 语言学特征确定单元504，被配置成确定第一声学特征中的语言学特征。

[0084] 转换单元505，被配置成基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0085] 在本实施例的一些可选的实现方式中，第一声学特征确定单元502进一步被配置成：根据目标文本和预训练的语音合成模型，确定目标文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0086] 在本实施例的一些可选的实现方式中，目标音色确定单元503进一步被配置成：根据目标音频和预训练的身份验证模型，确定目标音频对应的身份向量，其中，预训练的身份验证模型用于表征音频与身份向量之间的对应关系；根据身份向量，确定目标音色。

IPR2026-00272
Krisp Technologies EX1002 Page 330

[0087] 在本实施例的一些可选的实现方式中，语言学特征确定单元503进一步被配置成：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0088] 在本实施例的一些可选的实现方式中，转换单元505进一步被配置成：根据目标音色、语言学特征、目标文本和预训练的转换模型，确定目标文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征，其中，转换模型用于表征音色、语言学特征和文本与第二声学特征的对应关系；基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频。

[0089] 在本实施例的一些可选的实现方式中，语音合成装置还包括图5中未示出的训练单元，被配置成：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括音色、语言学特征、语言学特征对应的目标文本以及标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征；将训练样本集合中的训练样本的音色、语言学特征和语言学特征对应的目标文本作为初始神经网络模型的输入，将与输入的音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0090] 在本实施例的一些可选的实现方式中，训练样本集合不包括目标音色；第二声学特征包括对应于目标口音和目标音色的梅尔频谱；以及转换单元505进一步被配置成：根据梅尔频谱和预设的神经网络声码器，合成目标文本对应的具有目标口音和目标音色的音频。

[0091] 应当理解，语音合成装置500中记载的单元501至单元505分别与参考图2中描述的方法中的各个步骤相对应。由此，上文针对语音合成方法描述的操作和特征同样适用于装置500及其中包含的单元，在此不再赘述。

[0092] 根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

[0093] 如图6所示，是根据本申请实施例的语音合成方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0094] 如图6所示，该电子设备包括：一个或多个处理器601、存储器602，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线605互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置（诸如，耦合至接口的显示设备）上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线605与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作（例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统）。图6中以一个处理器601为例。

[0095] 存储器602即为本申请所提供的非瞬时计算机可读存储介质。其中，存储器存储有可由至少一个处理器执行的指令，以使至少一个处理器执行本申请所提供的语音合成方

IPR2026-00272
Krisp Technologies EX1002 Page 331

法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的语音合成方法。

[0096] 存储器602作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及单元，如本申请实施例中的语音合成方法对应的程序指令/单元(例如，附图5所示的获取单元501、第一声学特征确定单元502、目标音色确定单元503、语言学特征确定单元504和转换单元505)。处理器601通过运行存储在存储器602中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的语音合成方法。

[0097] 存储器602可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据语音合成电子设备的使用所创建的数据等。此外，存储器602可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器602可选包括相对于处理器601远程设置的存储器，这些远程存储器可以通过网络连接至语音合成电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0098] 语音合成电子设备还可以包括：输入装置603和输出装置604。处理器601、存储器602、输入装置603和输出装置604可以通过总线605或者其他方式连接，图6中以通过总线605连接为例。

[0099] 输入装置603可接收输入的数字或字符信息，以及产生与语音合成电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置604可以包括显示设备、辅助照明装置(例如，LED)和触觉反馈装置(例如，振动电机)等。该显示设备可以包括但不限于，液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

[0100] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0101] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令，并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的，术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如，磁盘、光盘、存储器、可编程逻辑装置(PLD))，包括，接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0102] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机

IPR2026-00272
Krisp Technologies EX1002 Page 332

具有：用于向用户显示信息的显示装置（例如，CRT（阴极射线管）或者LCD（液晶显示器）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或者、触觉输入）来接收来自用户的输入。

[0103] 可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（LAN）、广域网（WAN）和互联网。

[0104] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0105] 根据本申请实施例的技术方案，通过根据获取的目标文本确定的语言学特征以及通过获取的目标音频确定的目标音色，可以准确、快速地合成目标文本对应的具有目标口音和目标音色的目标音频。

[0106] 根据本公开的一个或多个实施例，本公开提供了一种语音合成方法，包括：获取目标文本和目标音频；基于目标文本，确定用于指示目标口音的第一声学特征；基于目标音频，确定目标音色；确定第一声学特征中的语言学特征；基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0107] 根据本公开的一个或多个实施例，基于目标文本，确定用于指示目标口音的第一声学特征，包括：根据目标文本和预训练的语音合成模型，确定目标文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0108] 根据本公开的一个或多个实施例，基于目标音频，确定目标音色，包括：根据目标音频和预训练的身份验证模型，确定目标音频对应的身份向量，其中，预训练的身份验证模型用于表征音频与身份向量之间的对应关系；根据身份向量，确定目标音色。

[0109] 根据本公开的一个或多个实施例，确定第一声学特征中的语言学特征，包括：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0110] 根据本公开的一个或多个实施例，基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，包括：根据目标音色、语言学特征、目标文本和预训练的转换模型，确定目标文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征，其中，转换模型用于表征音色、语言学特征和文本与第二声学特征的对应关系；基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频。

[0111] 根据本公开的一个或多个实施例，语音合成方法还包括：获取初始神经网络模型；

IPR2026-00272
Krisp Technologies EX1002 Page 333

获取训练样本集合，其中，训练样本集合中的训练样本包括音色、语言学特征、语言学特征对应的目标文本以及标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征；将训练样本集合中的训练样本的音色、语言学特征和语言学特征对应的目标文本作为初始神经网络模型的输入，将与输入的音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0112] 根据本公开的一个或多个实施例，训练样本集合不包括目标音色；第二声学特征包括对应于目标口音和目标音色的梅尔频谱；以及基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频，包括：根据梅尔频谱和预设的神经网络声码器，合成目标文本对应的具有目标口音和目标音色的音频。

[0113] 根据本公开的一个或多个实施例，本公开提供了一种语音合成装置，包括：获取单元，被配置成获取目标文本和目标音频；第一声学特征确定单元，被配置成基于目标文本，确定用于指示目标口音的第一声学特征；目标音色确定单元，被配置成基于目标音频，确定目标音色；语言学特征确定单元，被配置成确定第一声学特征中的语言学特征；转换单元，被配置成基于目标音色、语言学特征和目标文本，确定目标文本对应的具有目标口音和目标音色的音频，并输出具有目标口音和目标音色的音频。

[0114] 根据本公开的一个或多个实施例，第一声学特征确定单元进一步被配置成：根据目标文本和预训练的语音合成模型，确定目标文本对应的用于指示目标口音的第一声学特征，其中，预训练的语音合成模型用于表征说话文本与第一声学特征的对应关系。

[0115] 根据本公开的一个或多个实施例，目标音色确定单元进一步被配置成：根据目标音频和预训练的身份验证模型，确定目标音频对应的身份向量，其中，预训练的身份验证模型用于表征音频与身份向量之间的对应关系；根据身份向量，确定目标音色。

[0116] 根据本公开的一个或多个实施例，语言学特征确定单元进一步被配置成：利用预训练的识别模型提取第一声学特征中的语言学特征，其中，预训练的识别模型用于表征第一声学特征与语言学特征的对应关系。

[0117] 根据本公开的一个或多个实施例，转换单元进一步被配置成：根据目标音色、语言学特征、目标文本和预训练的转换模型，确定目标文本对应的用于生成具有目标口音和目标音色的音频的第二声学特征，其中，转换模型用于表征音色、语言学特征和文本与第二声学特征的对应关系；基于第二声学特征，确定目标文本对应的具有目标口音和目标音色的音频。

[0118] 根据本公开的一个或多个实施例，语音合成装置还包括图5中未示出的训练单元，被配置成：获取初始神经网络模型；获取训练样本集合，其中，训练样本集合中的训练样本包括音色、语言学特征、语言学特征对应的目标文本以及标注的与音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征；将训练样本集合中的训练样本的音色、语言学特征和语言学特征对应的目标文本作为初始神经网络模型的输入，将与输入的音色、语言学特征和语言学特征对应的目标文本对应的第二声学特征作为期望输出，对初始神经网络模型进行训练；将训练后的初始神经网络模型确定为转换模型。

[0119] 根据本公开的一个或多个实施例，训练样本集合不包括目标音色；第二声学特征包括对应于目标口音和目标音色的梅尔频谱；以及转换单元进一步被配置成：根据梅尔频

IPR2026-00272
Krisp Technologies EX1002 Page 334

谱和预设的神经网络声码器，合成目标文本对应的具有目标口音和目标音色的音频。

[0120] 应该理解，以上实施例仅是示例性实施例，但不限于此，还包括本领域已知的其他可以实现语音合成方法。可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

[0121] 上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

IPR2026-00272
Krisp Technologies EX1002 Page 335

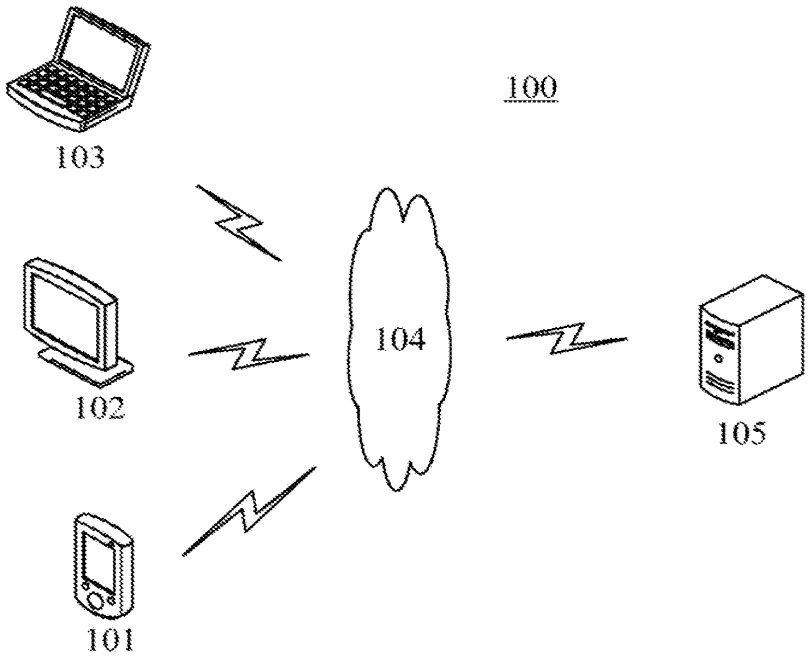


图1

IPR2026-00272
Krisp Technologies EX1002 Page 336

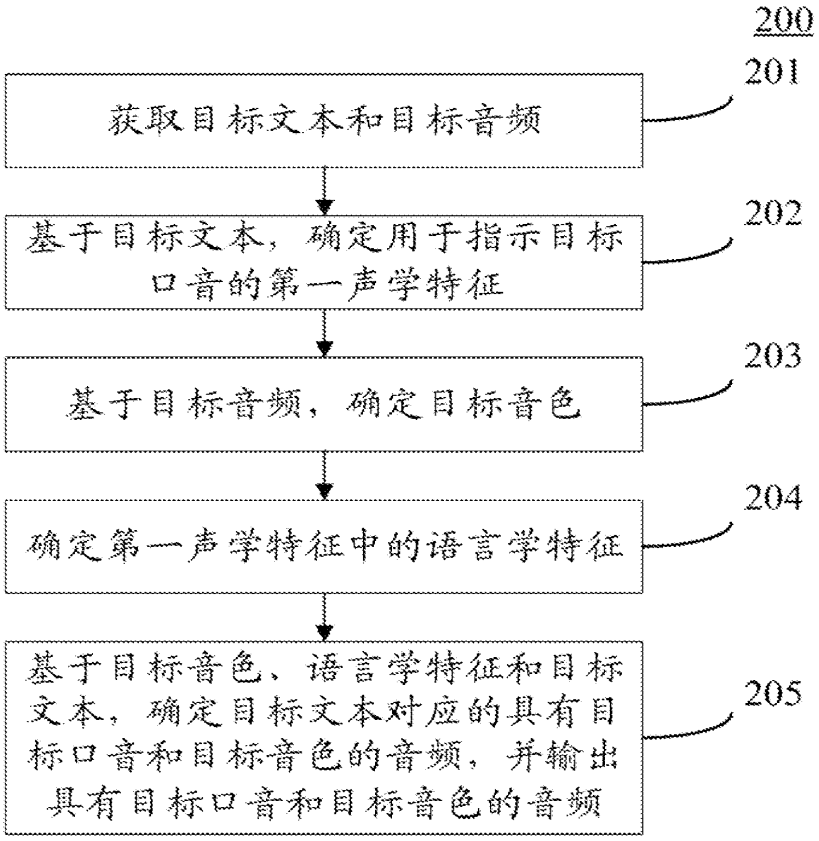


图2

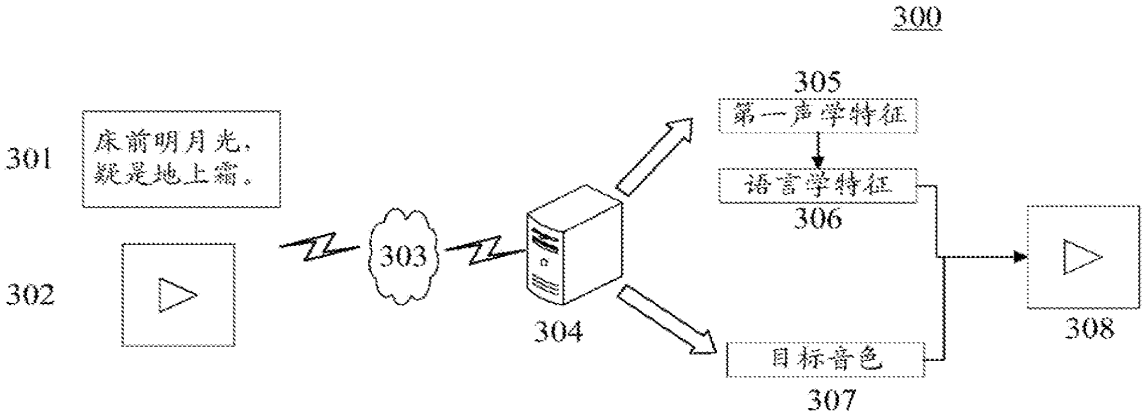


图3

IPR2026-00272
Krisp Technologies EX1002 Page 337

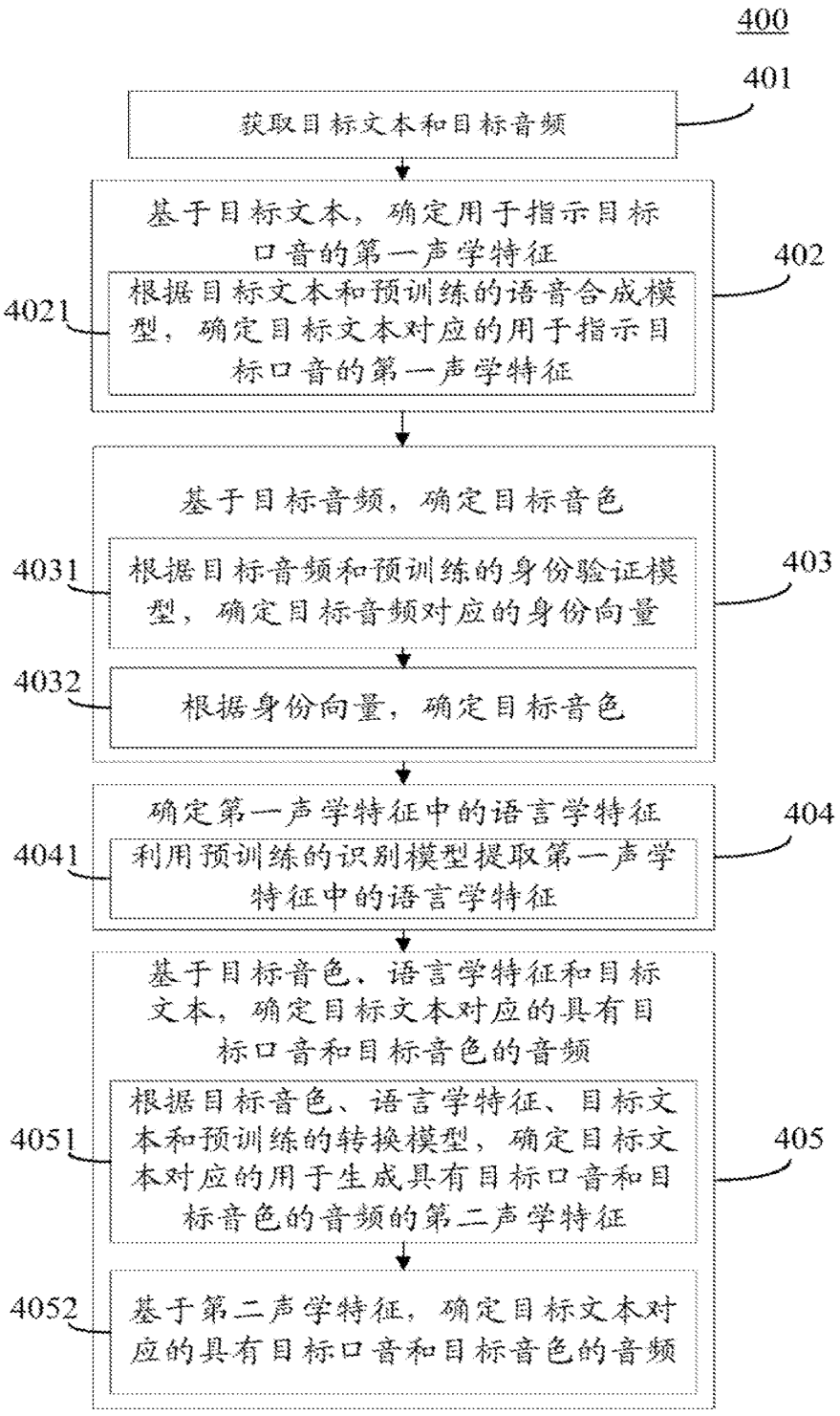


图4

IPR2026-00272
Krisp Technologies EX1002 Page 338

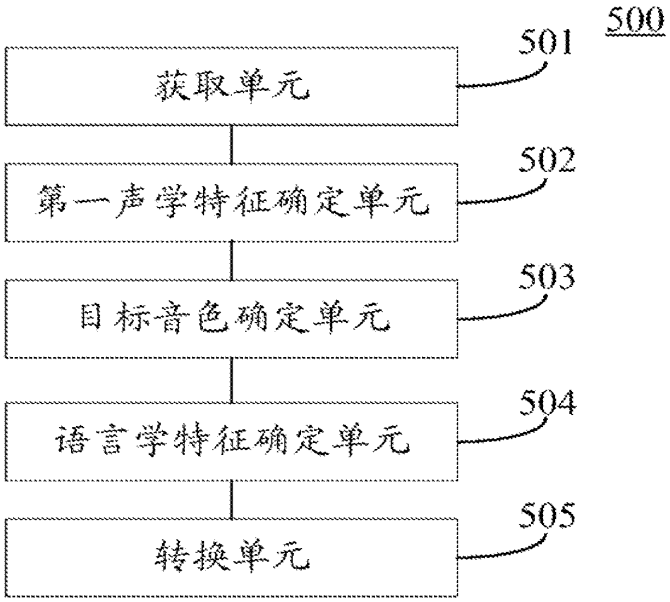


图5

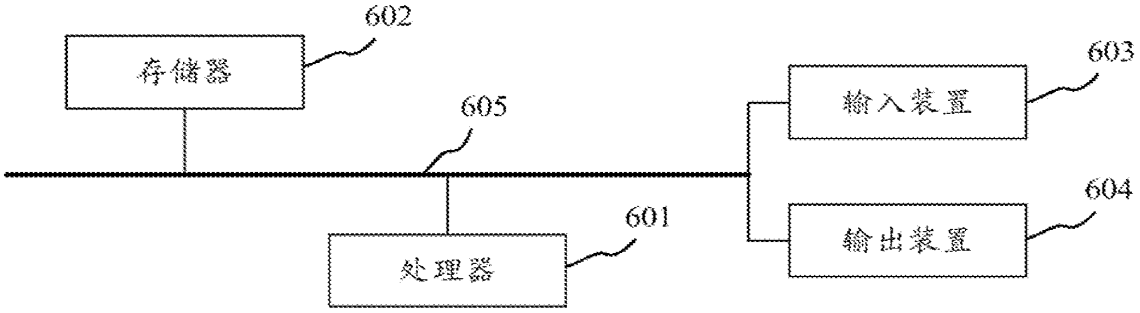


图6

IPR2026-00272
Krisp Technologies EX1002 Page 339

Espacenet

JP2005234418A METHOD AND COMPUTER PROGRAM FOR SYNTHESIZING F0-CONTOURS

Applicants: ATR ADVANCED TELECOMM RES INST

Inventors: JINFU NI,KAWAI HISASHI

Classifications:
IPC ***G10L13/08;*** (IPC1-7): G10L13/08;

Priorities: JP2004045855A 2004-02-23

Application: JP2004045855A·2004-02-23

Publication: JP2005234418A·2005-09-02

Published as: JP2005234418A;**JP4223416B2**

METHOD AND COMPUTER PROGRAM FOR SYNTHESIZING F0-CONTOURS

Abstract

PROBLEM TO BE SOLVED: To synthesize a speech with natural tones in a tone language, such as Chinese, with high reliability. SOLUTION: A method of synthesizing F0-contours from Chinese linguistic information 32 includes a step of preparing a machine-learning model 36 for modeling internal dependency of tonal F0peak and valley parameters with linguistic information of Chinese, a step 70 of predicting F0peak and valley parameters corresponding to input linguistic information 32 by use of the machine learning model 36, and steps 72 and 74 of generating F0-contours based on the predicted F0peak and valley parameters. COPYRIGHT: (C)2005,JPO&NCIPI

IPR2026-00272
Krisp Technologies EX1002 Page 340

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

DESCRIPTION JP2005234418A

Method and computer program for synthesizing F0 contours

[0001]

TECHNICAL FIELD The present invention relates to a speech synthesis method, and more particularly to a tonal language speech synthesis method and system.

[0002]

Efforts have been made to improve the performance of prosody generation to reliably convey linguistic information derived from text analysis in order to achieve high quality synthesized speech. The most difficult aspect of prosody generation is how to synthesize speech with appropriate tone and intonation. This prosodic element is particularly important in tonal languages.

[0003]

Chinese is one of the typical tonal languages.

In the Chinese prosody complex, the smallest characteristic component is the tone, especially focusing on the fundamental frequency (F0) contour. Tones are fundamental because larger structures, including the prosody system itself, are specific variations of one or more tones. From this point of view, the synthesis of Chinese F0 contours can focus on tone and its variation with context.

[0004]

チャオ、Ｙ．Ｒ．１９６８、「中国語話し言葉の文法」、バークレー、ＣＡ、カリフォルニア大学出版局（Chao, Y. R., 1968. A Grammar of Spoken Chinese. Berkeley, CA. University of California Press.）

シェン、Ｘ．Ｓ．１９９０、「標準中国語の韻律」、カリフォルニア大学出版局。（Shen, X. S., 1990. The Prosody of Mandarin Chinese. University of California Publications.）

シェン、「漢語語調和語調類型」、中国 語文、１９９４、３、２２１-２２８

[0005]

シュイ、Ｙ．１９９９、「Ｆ０輪郭の形成及びアライメントに対する声調と焦点の効果」、音声学ジャーナル、２７、５５-１０５。（Xu, Y., 1999. Effects of Tone and Focus on the Formation and Alignment of F0 Contours. Journal of Phonetics, 27, 55-105.）

IPR2026-00272
Krisp Technologies EX1002 Page 341

ニ、Ｊ及びヒロセ、Ｋ.２０００、「標準中国語文の基本周波数輪郭の機能的モデリングに対する実験的評価」、ＩＳＣＳＬＰ２０００、北京、３１９－３２２。（Ni, J. and Hirose, K., 2000. Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences. ISCSLP2000. Beijing, 319-322.）

ニ、Ｊ及びカワイ、Ｈ.、「パラメータ的モデリングによる声調特徴の抽出と合成ベース分析によるパターンマッチング」、ＩＣＡＳＳＰ２００３、ｐｐ.７２－７５、２００３。（Ni, J. and Kawai, H., "Tone Feature Extraction through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching," ICASSP2003, pp. 72-75, 2003.）

カワハラ、Ｈ.、イクヨ、Ｍ．Ｋ.、チェイニー、Ａ．１９９９、「ピッチ適応時間周波数平滑化及び瞬時周波数ベースのＦ０抽出を用いた音声表現の再構築：音声における反復構造の果たし得る役割」、音声コミュニケーション、２７、１８７－２０７。（Kawahara, H., Ikuyo, M. K., Cheneigne, A., 1999. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetilive Structure in Sounds. Speech Communication, 27, 187-207.）

コロベール、Ｒ.、ベンジオ、Ｓ.及びマリソン、Ｊ.、「トーチ：モジュラー機械学習ソフトウェアライブラリ」、技術報告ＩＤＩＡＰ、ｐｐ.１－９、２００２。（Collobert, R., Bengio, S., and Marielhon, J., "Torch: a Modular Machine Learning Software Library," Technical Report IDIAP, pp. 1-9, 2002.）

[0006]

Pitch targets are believed to play an important role in expressing Chinese tone and intonation.

Pitch targets basically include high and low, which are commonly used to represent intonation in accented languages such as English and Japanese.

In Chinese, there are four tones called tones 1 through 4, and an intermediate tone called tone 0.

If the speaker's voice is: 1.

2. low;

3. half low;

intermediate, 4.

5. half height;

If divided into four equal intervals of high scores, the first through fourth tones are denoted as 55, 35, 214, and 51, respectively.

Since both actual spacing and absolute pitch are relative to individual voices and speaking moods, the term "pitch target" as used herein refers to F0 (basic pitch) over time. frequency) peaks and valleys.

[0007]

Tone and intonation patterns, on the other hand, are particularly relevant to F0 contours.

The temporal range of tone patterns is limited to the size of a syllable, whereas the temporal range of intonation patterns covers more than syllables and may span entire utterances.

[0008]

There is a lot of research on tone and intonation, for example, documents such as Non-Patent Document 1 to Non-Patent Document 4.

Previous perceptual tests and instrumental analyzes have yielded a consensus view that the F0 contour of speech can represent both tone and intonation.

[0009]

However, some basic questions, at least in practice, remain elusive.

For example, are pitch targets sufficient to represent Chinese tone and intonation, or what are the essential features needed for tone and intonation synthesis to achieve natural sound in text-to-speech conversion? It is not clear whether

IPR2026-00272
Krisp Technologies EX1002 Page 342

Therefore, there has been no reliable method of synthesizing natural Chinese speech.

[0010]

SUMMARY OF THE INVENTION Accordingly, one of the objects of the present invention is to provide a method and a computer program for synthesizing natural speech with high reliability.

[0011]

According to one aspect of the present invention, a method for synthesizing fundamental frequency (F0) contours from given tonal language linguistic information comprises the steps of tonal fundamental frequency peak and valley parameters to tonal language linguistic features: preparing a machine learning model for modeling internal dependencies; using the machine learning model to predict peak and valley parameters of F0 corresponding to input linguistic information; generating an F0 contour based on the peak and valley parameters of .

[0012]

Preferably, the preparing step comprises preparing a training data set comprising linguistic information and associated speech data; extracting F0 contour model parameters from the speech data in the training data set; estimating the F0 peak and valley parameters from the F0 contour model parameters extracted in the step; using the F0 peak and valley parameters estimated in the estimating step and the linguistic information in the training data set; , training the machine learning model so that it learns the internal dependence of the F0 peak and valley parameters on linguistic information.

[0013]

More preferably, the step of estimating F0 peak and trough parameters includes RONDO-F0 contours, wherein the F0 contours of each utterance in the training data set are represented by a continuous chevron pattern serially aligned along the time axis. identifying the F0 crests in the RONDO-F0 contour obtained in the converting step, and identifying all adjacent F0 crests in the RONDO-F0 contour obtained in the converting step. finding an F0 trough such that adjacent F0 peaks and F0 troughs satisfy a predetermined condition in between.

[0014]

The finding step includes finding an initial candidate for the F0 valley (tvfi, λvfi) between the i-th F0 peak and the next peak in the RONDO-F0 contour; , λvfi) by subtracting tvfi at predetermined time intervals until the RONDO-F0 contour satisfies a predetermined condition for the valley of F0.

[0015]

Preferably, the step of finding the initial candidates comprises setting the F0 valleys (tvfi, λvfi) of the initial candidates to the lowest depressions denoted by (/tvi, /λvi).

[0016]

More preferably, the searching step starts from the initial candidates (/tvi, /λvi) and continues for a predetermined time until λvfi - λpi ≤ (/λvi - λpi) × C, where C is a predetermined constant, or tvfi = tpi. Finding the F0 valleys on the RONDO-F0 contour by subtracting tvfi with the interval.

[0017]

The constant C may be chosen to be approximately 0.95.

[0018]

More preferably, the F0 contour F0(t) is a function of time t

IPR2026-00272
Krisp Technologies EX1002 Page 343

[0019]

defined as

[0020]

Preferably, the method further comprises synthesizing speech based on the input linguistic information and the F0 contours generated in the generating step.

[0021]

The predetermined tone language may be Chinese.

[0022]

Another aspect of the invention relates to a computer program product which, when run on a computer, causes the computer to perform all the steps of any of the methods described above.

[0023]

1.

First, this invention focuses on the role of pitch targets in displaying Chinese tone and intonation.

To investigate whether the pitch target is sufficient to represent Chinese tones and intonations, the pitch target was measured specifically as peaks and valleys of F0 with respect to time variation.

[0024]

Analysis and perceptual experiments were performed on 72 sentences, each uttered twice as declarative and interrogative sentences by female native speakers with nearly identical tone mappings.

The tonal and intonation patterns observed from the F0 contours were quantitatively analyzed using functional models and then resynthesized using model parameters predicted from the measured pitch targets.

Two cognitive experiments were conducted.

On the one hand, we evaluated the similarity of the predicted tone and intonation patterns to the pitch target and the original text.

On the other hand, we tested the human perception of tone and intonation when systematically varying the pitch targets of the final tones (2nd and 4th tones) in two declarative sentences.

[0025]

Experimental results consistently show that pitch targets play an important role in defining Chinese tone and intonation patterns.

Given a pitch target, the exact shape of the F0 contour is predictable.

Based on this result, a Chinese speech synthesis method can be constructed.

An experiment will be described first, and an embodiment of the invention will be described later in the specification.

[0026]

2.

Audio samples and analysis methods 2.1.

Speech Samples The speech data used here contains 72 Chinese sentences, almost all of which are taken from Non-Patent Document 2.

IPR2026-00272
Krisp Technologies EX1002 Page 344

These sentences were divided into 6 groups.

Each contained 12 basic sentences, which were further subdivided into three types.

Each type contains four sentences with the same number of syllables and the same grammatical structure characterized by the same tonal mapping to the whole sentence, as shown in Table 1.

T1, T2, T3, and T4 in the table indicate the first, second, third, and fourth tones, respectively.

<Table 1>

[0027]

Type 1 is a subject-verb (SV) structure and contains four syllables.

Type 2 contains five syllables in a subject-verb-object (SVO) structure.

Type 3 is a combination of Types 1 and 2.

That is, type 2 is added to type 1 as the object of the sentence, resulting in a 9-syllable SVO structure.

We have grouped these statements into the following categories:

[0028]

These 72 sentences were recorded twice in a soundproof room without emotional expressions by female speakers.

[0029]

２．２．

Functional Model of the F0 Contour In this application, a functional model is used (see Non-Patent Document 5) to represent the F0 contour in the form of parameters.

According to this model, the speaker's vocal register (speech frequency range) is first transformed to the so-called RONDO scale (similar to the logarithmic scale).

The RONDO-F0 contour is then represented as a continuous chevron-shaped pattern aligned in series along the time axis.

The F0 contour F0(t) is given below.

[0030]

Min(z1,z2) means the smaller of z1 and z2.

Equations (1) and (2) together represent the vocal register transformation.

Equation (3) represents the RONDO-F0 contour Λ(t), where Λri(t) and Λfi(t) denote the ascending and descending components of the i-th chevron pattern, respectively.

すなわち

[0031]

The parameters ζ, λt and λb can both be fixed at 0.237, 1 and 2 respectively.

(See Non-Patent Document 5.

) This results in the following two speaker-dependent but utterance-independent parameters in the frequency domain. [f0b, f0t]: the highest and lowest frequencies of the vocal register, and RONDO-in time space , five parameters that are utterance-dependent but speaker-independent, n: the number of chevron patterns Δtxi: the response time of the i-th rising/falling component ΔAxi: the amplitude of the i-th rising/falling component, xε {r, f} (tpi, Λpi): i-th peak of the mountain-shaped pattern, i=1, . . . n.

[0032]

IPR2026-00272
Krisp Technologies EX1002 Page 345

2. 3.

Method The 144 observed F0 contours were first automatically analyzed using the method of [6].

After that, the peaks and valleys of the F0 were judged manually by visually inspecting the F0 contour while considering the original tone.

The number of F0 peaks for a given tone was determined according to tone modeling [6].

The contours between adjacent peaks were then used to determine the F0 valleys.

We resynthesized these utterances with F0 contours generated by the model and performed perceptual experiments using a tool called STRAIGHT [7].

Three analytical and perceptual experiments were performed.

Experiment 1 analyzed the effectiveness of F0 contour resynthesis based on F0 peaks and valleys.

In Experiment 2, we investigated the correlation between changes in F0 peaks and troughs and interactions with tone and intonation.

Experiment 3 shows that tone and intonation can be changed by changing the pitch target.

Based on these experimental results, we discuss that pitch targets define tone and intonation.

[0033]

3.

Results 3.1.

Tone and intonation pattern resynthesis Experiment 1 was conducted to examine the effectiveness of resynthesizing F0 contours based on F0 peaks and valleys.

Let (tvi, λvi) denote the valley between the i-th and i+1-th peaks.

Given a mountain, the other model parameters required for the generation of the F0 contour are calculated as described above.

<Table 2>

[0034]

Table 2 shows the sample statistical results for tones measured from the speech samples, where μc and σc denote the manually checked mean and variance of these model parameters (checked parameters), μp and σp indicate what was predicted by the peaks and valleys of F0 (prediction parameters).

The μe and σe columns show the mean and variance of the error between the checked and predicted parameters.

[0035]

To test the similarity of the resynthesized tone and intonation patterns to the original text, we perceived 288 stimulus pairs, including 144 resynthesized utterances with checked parameters and 144 utterances with predicted parameters. I did an experiment.

Stimuli were presented to two native speakers using headphones in a silent room.

After listening to pairs of stimuli, listeners rated their tone and intonation similarity on a 3-point scale: 0 (very different), 1 (similar), 2 (no difference).

Listeners were allowed to listen to the stimuli multiple times before judging.

The average scores for the checked and predicted parameters were 1.93 and 1.89, respectively, and no "very different" samples occurred.

IPR2026-00272
Krisp Technologies EX1002 Page 346

The experimental results show that the pitch target, ie the peaks and troughs of F0 over time, is sufficient to capture tonal and intonation pattern features.

[0036]

3. 2.

Tone-intonation interaction Experiment 2 examined the interaction between tone and intonation by analyzing peaks and valleys of F0 for each of the 12 categories.

The main results are explained below.

First, the F0 contours of utterances in interrogative sentences generally moved upwards more or less than those in declarative sentences.

This result agrees with the findings of Non-Patent Document 2 and Non-Patent Document 3.

For the same 1st and 4th tone mapping utterances, the peaks and valleys of the F0 rose higher register than for the same 2nd and 3rd tone mappings.

Figure 1 shows two sentences uttered in declarative and unmarked interrogative sentences.

[0037]

, whose meanings are respectively (a) "Red noses have no power" and (b) "Passers-by take pictures", and (c) " Do red noses have power? and (d) Do passers-by take pictures?

[0038]

Second, there is a way to rely on the tone at the end of the sentence to express intonation in both unmarked and marked questions.

Try grouping the final tones into two sets.

2nd and 3rd tones, and 1st and 4th tones.

In the former, the ascending F0 crest is raised to a higher vocal register, thus increasing the tonal range considerably.

However, in the latter, both peaks and troughs (if any) of F0 are raised to higher registers, thus narrowing the tonal range and moving the scale of F0 troughs up.

This phenomenon is clearly observed in the examples shown in FIGS. 1(c) and (d).

[0039]

Third, there is an ascending-descending pattern for yes/no questions of X-not-X structure, ie categories NO, N1, N2, and wh questions, ie categories W0, W1, W2.

The ascending-descending pattern is basically expressed by arranging the peaks and valleys of F0 in the tone structure of the function word, such as shi4-bu2-shi4.

FIG. 2 shows four examples.

(a) "bao1shen1gong1 shi4-bu2-shi4 ca1 che1?

(Do indentured workers clean their cars?

); (b) "lao3shou3zhang3 shi4-bu2-shi4 mai3 jiu3?

(Do old dignitaries buy liquor?

); (c) "bao1shen1gong1 ca1-mei2-ca1 che1?

(Did the indentured worker clean the car?

); (d) "lao3shou3zhang3 mai3-mei2-mai3 jiu3?

IPR2026-00272
Krisp Technologies EX1002 Page 347

(Did the old dignitary buy a drink?

) である。

From this figure it is clear that the tones contained in the function words are adjusted to fit the ascending-descending pattern.

If a tone collides with this, it loses its basic shape and follows an ascending-descending pattern trajectory, eg syllables ca1 and mei2 in FIG. 2(c).

[0040]

Fourth, in alternative questions (categories Q0, Q1, Q2), "transition patterns" are used for function words shi4 and haishi4.

The term transition pattern means that the tones in function words take a fairly narrow F0 range and are located in the middle vocal register.

In contrast, in phrases around function words, focus phenomena (Non-Patent Document 5) are usually observed.

These observations clearly show that the interaction between tone and intonation is well captured by the pitch target.

[0041]

３．３．

Perception of Artificial Tone and Intonation Patterns Experiment 3 investigated the perception of tone and intonation while systematically varying the pitch target.

Two utterances of declarative sentences shown in FIGS. 1(a) and 1(b) were used as carrier utterances, and peaks and troughs of the final tone F0 were changed in two/three ways.

[0042]

The first mode was to vary the peaks of F0 (ie, the model parameters λpi) with respect to the original with a step size of 0.1, while fixing the valleys of F0 unchanged.

FIG. 3 represents the observed F0 contours ("+" sequences) and their model-generated contours (referred to as solid lines, artificial tones and intonation patterns).

[0043]

A second modality was to vary both the peaks and valleys of F0 simply by raising or lowering its tone with the same step size of 0.1.

The model-generated F0 contours are indicated in FIG. 4 by symbols B1 to B6 and D1 to D6.

[0044]

The third mode, especially for the fourth voice, was to move the valley of F0 upward with a step size of 0.1 while fixing the peak of F0 to raise the valley.

The model-generated F0 contours are indicated by symbols E1 to E3 in FIG.

Two carrier utterances were resynthesized with all model-generated F0 contours.

[0045]

Perceptual testing of these utterances was performed on three native speakers.

Stimuli were presented to listeners twice in random order through headphones in a silent room.

After listening to the stimuli, listeners answered three questions.

(1) Is the utterance declarative or interrogative?

IPR2026-00272
Krisp Technologies EX1002 Page 348

(2) Was the last syllable stressed, normal, or neutral?

(3) Which tone did you hear on the last syllable?

The experimental results are summarized in Table 3.

Here, "Que" and "Sta" indicate "interrogative sentence" and "declarative sentence" respectively, and "Emp", "Nor" and "Wea" indicate "emphasized", "normal" and "weak stress" respectively.

<Table 3>

[0046]

Three findings can be obtained from this experiment.

First, contours can represent tone, stress and intonation [see Non-Patent Documents 1 and 2].

Syllables were consistently perceived to be emphasized when the F0 peak was raised.

When the F0 peak was lowered, the syllables were perceived with less stress.

Second, it proved that interrogative intonation is expressed dependent on the final tone.

In the case of the second voice, the higher the peak of F0, the more easily the utterance was judged to be an interrogative sentence.

For the fourth voice, utterances were recognized as interrogative only if the tone was in the high vocal register.

However, all listeners perceived the utterances as declarative when the troughs were in the lower register.

Experimental results also showed that final tone features were not sufficient to distinguish between interrogative and declarative sentences.

Other features such as tempo are also key to perception.

Finally, the tone is determined by its F0 crests, troughs and their alignment with syllables [see Non-Patent Document 1].

The first and third tones were perceived under certain conditions in this experiment.

In addition, from the results that the F0 contours marked B1, B2 and E1 were perceived as the first voice, the first voice exhibits rising contours in the high register as shown in FIG. 2(c), explain the phenomenon that, by using , the need to express intonation is not lost, but the perception of it is not lost.

[0047]

4 .

EXAMPLES To study the tone and intonation patterns of Chinese, several analyzes and perceptual experiments were performed on well-designed speech samples.

Experimental results have shown that pitch targets play an important role in defining tone and intonation patterns.

For example, a functional model can be used to predict the exact F0 contour from the peaks and valleys of the F0.

Based on this result, we hypothesized that the observed F0 contour could be skeletonized as a sequence of F0 peaks and troughs without losing the main linguistic and paralinguistic information it conveys.

The embodiments described below are based on this concept.

[0048]

4 . 1 .

Structure FIG. 5 is a block diagram of a speech synthesis system according to one embodiment of the present invention.

Referring to FIG. 5, the system 20 includes an F0 parameter extraction module 34 that extracts F0 crest and trough data from the training data 30, which data is the difference between the prosodic features and the underlying linguistic information. used to

IPR2026-00272
Krisp Technologies EX1002 Page 349

associate

System 20 further includes a machine learning module 36 that is used to learn the internal dependencies of the associated parameters on the underlying linguistic information, estimates F0 contours from linguistic information 32, and and a synthesis module 38 for synthesizing the Chinese speech 40 .

[0049]

Training data 30 includes text with linguistic information and accompanying F0 contour data.

[0050]

F0 parameter extraction module 34 includes analysis-by-synthesis (ABS)-based analysis module 50 to achieve optimal estimation of model parameters from observed F0 contours of training data 30, and F0 valleys (tvxi, λvxi), where the peak of F0 is obtained at the output of the ABS-based analysis module 50.

[0051]

The ABS-based analysis module 50 attempts to achieve optimal estimation of model parameters from observed F0 contours.

Non-Patent Document 6 proposes an algorithm for extracting tone crests and descending gradient features from F0 contours based on a functional model.

In order to reliably skeletonize the F0 contour to its underlying F0 peaks and valleys, the algorithm re-estimates the model parameters for the tonal downslope features, i.e., Δtx, ΔAxi. introduced some new constraints.

(^λvi, ^tvi) (where ^ is attached above each letter) denotes the observed F0 valley between the i-th peak (λpi, tpi) and the next peak, if any. shall be shown.

To reduce the impact of F0 extraction error on valley determination, we consider the likelihood of speech frames.

The constraints for reestimating Δtxi and ΔAxi are listed below.

[0052]

Target finder module 52 estimates the valley of F0.

Two kinds of pitch targets, peaks and valleys of F0, are used as association parameters to make associations between prosodic features and underlying linguistic information.

As described above, experimental results from acoustic analysis and perceptual testing consistently indicate that pitch targets are sufficient to represent Chinese tone and intonation patterns.

Given a pitch target, it is possible to predict the exact shape of the F0 contour from the functional model.

[0053]

Since the peaks of F0 are available from the set of estimated model parameters, the target search specifically focuses on the valleys of F0 (tvxi, λvxi) for either the i-th rising or falling component, and the i-th peak It is done for nearby RONDO contours.

An initial candidate for a valley, say (tvfi, λvfi), is the first set of lowest dips denoted by (/tvi,/λvi) in the RONDO contour between the ith peak and the next.

after that,

[0054]

Or search for valley candidates along the RONDO contour toward the i-th peak by decreasing tvfi by a very short step interval (eg 0.005 seconds) until tvfi ≈ tpi.

In equation (13), the constant 0.95 was determined considering the definition of Δtx as the response time required for unit decay from 1 to 0.05.

IPR2026-00272
Krisp Technologies EX1002 Page 350

すなわち

[0055]

This corresponds to the expression α=4.8/Δix used in equation (6).

If tvfi and tvri+1 (where "i+1" in "tvri+1" is the subscript of "r").

) is less than a threshold, the two valleys are always fixed to the average of the two valley candidates.

[0056]

A machine learning module 36 is used to learn the internal dependencies of the association parameters on the underlying linguistic information.

Several effective machine learning methods are available, as described in Non-Patent Document 8.

[0057]

Synthesis module 38 includes a machine prediction module 70 for predicting pitch targets from input linguistic information by using regression algorithms associated with the machine learning methods used in machine learning module 36, and machine prediction module 70, a parameter transformation module 72 for computing model parameters relating to the tone downslope features from the peaks and valleys of F0 given from modal-based synthesis module 74 for synthesizing the F0 contours, and for synthesizing speech according to the input linguistic information 32 and appropriate tones based on the F0 contours provided by the modal-based synthesis module 74, and a speech synthesis module 76 .

Each of these modules is described below.

[0058]

The parameter transformation module 72 computes model parameters associated with tonal downslope features from given F0 peaks and valleys.

Specifically, it is as follows.

[0059]

4．2．

Operating system 20 operates as follows.

Action has three phases.

F0 parameters are extracted from training data 30 by F0 parameter extraction module 34 .

Machine learning is performed with the parameters extracted by the F0 parameter extraction module 34 .

An F0 contour is then estimated for the linguistic information 30, and Chinese speech is synthesized based on the linguistic information 32 with tones following the estimated F0 contour.

[0060]

In a first aspect, the ABS-based analysis module 50 analyzes the input training data 30 containing F0 contours and linguistic information and outputs the input data tone crest and downslope feature parameters. .

[0061]

Target finder module 52 estimates the valley of F0 using equations (13) and (14).

The estimated F0 troughs are collected together with the F0 peaks for machine learning in the second phase.

[0062]

IPR2026-00272
Krisp Technologies EX1002 Page 351

In a second aspect, the collected F0 peaks and troughs along with the corresponding linguistic information in the training data 30 are provided to the machine learning module 36, and the machine learning module 36 derives the underlying language of the associated parameters. learn internal dependence on academic information.

Once this phase is over, the machine learning module 36 can be used to estimate the F0 peaks and valleys that are likely to occur in the input linguistic information.

[0063]

In a third aspect, synthesis module 38 uses machine learning module 36 to synthesize Chinese speech 40 from the linguistic information.

[0064]

Specifically, machine prediction module 70, given linguistic information 32, uses machine learning module 36 to predict pitch targets that are likely to occur in the input linguistic information.

[0065]

The parameter transformation module 72 computes model parameters related to the tone downslope feature from the peaks and valleys of F0 estimated by equations (15)-(18).

[0066]

The modal-based synthesis module 74 uses the functional model to synthesize the F0 contour given the model parameters and vocal register [f0b, f0t].

[0067]

A speech synthesis module 76 analyzes the linguistic information 32 and the F0 contours provided by the modal-based synthesis module 74 and produces Chinese speech that conveys the linguistic information in the appropriate tones.

[0068]

- Example Results - Figures 6 and 7 are examples of some of the proposed system and method, showing the aspects of skeletonizing the F0 contours into underlying F0 peaks and valleys, and the contours from the F0 peaks and valleys. and a phase of restoring the

[0069]

Given the measured F0 contours ("+" sequences) as shown in FIG. 6(a), these are first expressed in parameter form as shown in FIG. be.

The parameters of the peaks are available from the set of model parameters, while the valleys are searched from the ascending/descending components around these peaks in RONDO time space.

The top of FIG. 7 and FIG. 6(c) plot peaks (black circles) and valleys (white circles), which give the skeleton of the F0 contour.

FIG. 6 shows the F0 contour recovered by transforming the peaks and valleys of the F0 to the model parameters by which the contour is synthesized, and FIG. 7 shows the observed F0 contour and the recovered contour in solid lines. show.

[0070]

The system and method of this embodiment for synthesizing Chinese F0 contours and speech from linguistic information uses a functional model.

IPR2026-00272
Krisp Technologies EX1002 Page 352

Compared with the prior art, this embodiment introduces the peaks and valleys of F0 as the main parameters that relate between prosodic and linguistic features, and uses machine learning techniques to learn the internal dependencies between them. to introduce

As noted above, preliminary experiments have confirmed the effectiveness of this embodiment.

[0071]

Although the invention has been described in terms of embodiments relating to Chinese, it is readily understood that the invention is applicable to any language in which the smallest characteristic component is tone. You can.

[0072]

The system described above can be implemented in computer hardware and software (computer programs) running on an operating system (OS).

When synthesizing voice, a voice generating device such as a speaker is used.

Other than that, no special hardware is required.

[0073]

This program may be stored in a storage medium such as FD (flexible disk), CD-ROM (compact disk read-only memory), MO (magneto-optical disk), or DVD (digital versatile disk), or may be downloaded via the Internet or the like. It may be transmitted over some data communication network.

Any program that implements the system described above when executed on a computer falls within the scope of the present invention.

[0074]

The above-described embodiments are merely examples and should not be construed as restrictive.

The scope of the present invention is indicated by each claim in the scope of claims after taking into account the description of the detailed description of the invention, and all changes within the meaning and range of equivalents to the wording described therein include.

[0075]

FIG. 10 is a diagram showing examples of F0 contours of two Chinese sentences "hong2bi2tou2 mei2 quan2" and "guo4lu4ke4 zhao4 xiang4" spoken in declarative sentences;

FIG. 4 shows ascending-descending patterns in four example yes/no questions;

Fig. 10 shows the observed F0 contours ("+" sequence) and the contours generated from the model when the last F0 crest was moved upwards in one experiment;

FIG. 11 shows F0 contours generated by moving the tone upwards or downwards in an experiment;

1 is a block diagram of a speech synthesis system according to one embodiment of the present invention; FIG.

FIG. 10 illustrates an example of F0 contour skeletonization and reconstruction;

1 illustrates a portion of the system and method of one embodiment of the invention; FIG.

Code explanation

[0076]

20 Chinese speech synthesis system, 30 training data, 32 linguistic information, 34 F0 parameter extraction module, 36 machine learning module, 38 synthesis module, 50 ABS-based analysis module, 52 target search module, 70 machine prediction module, 72 Parameter Conversion Module, 74 Modal-Based Synthesis Module, 76 Speech Synthesis Module

IPR2026-00272
Krisp Technologies EX1002 Page 353

IPR2026-00272
Krisp Technologies EX1002 Page 354

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patent Translate
Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

CLAIMS JP2005234418A

1.

A method for synthesizing fundamental frequency (F0) contours from linguistic information of a given tonal language, the method comprising modeling the internal dependence of tonal fundamental frequency (F0) peak and valley parameters on the linguistic features of said tonal language. Preparing a machine learning model for doing; using the machine learning model to predict F peak and valley parameters corresponding to input linguistic information; generating F0 contours based on valley parameters.

2.

the preparing step includes preparing a training data set including linguistic information and associated speech data; extracting F0 contour model parameters from the speech data in the training data set; estimating F0 peak and valley parameters from the F0 contour model parameters extracted in the step of performing; the F0 peak and valley parameters estimated in the estimating step; and training the machine learning model so that it learns the internal dependencies of F0 peak and valley parameters on linguistic information using A method for synthesizing the described F0 contours.

3.

The step of estimating the peak and valley parameters of the F0 transforms the F0 contours of each utterance in the training data set into RONDO-F0 contours represented by consecutive chevron patterns serially aligned along the time axis. identifying the positions of F0 crests in the RONDO-F0 contour obtained in the transforming step; finding F0 valleys between F0 peaks such that adjacent F0 peaks and F0 valleys satisfy a predetermined condition. .

4.

the finding step finds an initial candidate for an F0 valley (tvfi, λvfi) between the i-th F0 peak and the next peak in the RONDO-F0 contour; searching for the valley of F0 on the RONDO-F0 contour by subtracting tvfi at predetermined time intervals until the valley of (tvfi, λvfi) satisfies a predetermined condition. A method for synthesizing the F0 contours of .

5.

The step of finding the initial candidate places the trough (tvfi, λvfi) of the F0 of the initial candidate in the lowest depression denoted by (/tvi,/λvi) 5. The method of synthesizing F0 contours of claim 4, comprising the step of setting .

6.

IPR2026-00272
Krisp Technologies EX1002 Page 355

At predetermined time intervals, starting from the initial candidate (/tvi, /λvi), λvfi - λpi ≤ (/λvi - λpi) x C, where C is a predetermined constant, or tvfi = tpi. 6. A method for synthesizing F0 contours according to claim 5, comprising searching for F0 valleys on said RONDO-F0 contour by subtracting tvfi.

7.

A method for synthesizing F0 contours according to claim 6, wherein said constant C is chosen to be approximately 0.95.

8.

An F0 contour according to any of claims 1 to 7, wherein the F0 contour F0(t) and the corresponding RONDO-F0 contour Λ(t) are defined as functions of time t: How to synthesize.

9.

An F0 contour according to any preceding claim, further comprising synthesizing speech based on said input linguistic information and said F0 contour generated in said generating step. method of synthesizing

10.

A method for synthesizing F0 contours according to any one of claims 1 to 9, wherein the predetermined tone language is Chinese.

11.

A computer program product which, when run on a computer, causes the computer to perform all the steps of any of claims 1-10.

IPR2026-00272
Krisp Technologies EX1002 Page 356

(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-234418
(P2005-234418A)
(43)公開日　平成17年9月2日(2005.9.2)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G10L 13/08	G10L 3/00	H	5D045

審査請求 未請求　請求項の数 11　OL　外国語出願　（全 20 頁）

(21)出願番号　特願2004-45855(P2004-45855)
(22)出願日　平成16年2月23日(2004.2.23)

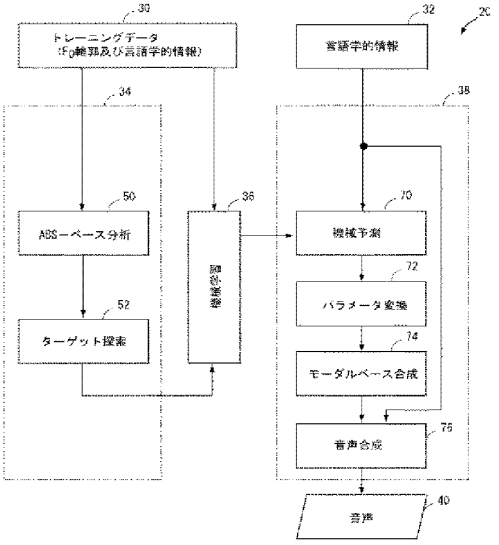
（出願人による申告）平成１５年度通信・放送機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(71)出願人　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　ジンフ・ニ
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　河井　恒
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

Ｆターム(参考)　5D045 AA09

(54)【発明の名称】ＦＯ輪郭を合成する方法及びコンピュータプログラム

(57)【要約】
【課題】中国語等の声調言語において、高い信頼性をもって自然な声調の音声を合成できるようにする。
【解決手段】　中国語の言語学的情報３２からF_0輪郭を合成する方法は、声調のF_0の山及び谷パラメータの、中国語の言語学的情報に対する内部依存性をモデル化するための機械学習モジュール３６を準備するステップと、機械学習モジュール３６を用いて、入力された言語学的情報に対応するF_0の山及び谷パラメータを予測するステップ７０と、予測されたF_0の山及び谷パラメータに基づき、F_0輪郭を生成するステップ７２、７４とを含む。
【選択図】　図５



IPR2026-00272
Krisp Technologies EX1002 Page 357

【特許請求の範囲】
【請求項1】
所定の声調言語の言語学的情報から基本周波数（F_0）輪郭を合成する方法であって、
声調基本周波数（F_0）の山及び谷パラメータの前記声調言語の言語学的特徴に対する内部依存性をモデリングするための機械学習モデルを準備するステップと、
前記機械学習モデルを用いて、入力された言語学的情報に対応するF_0の山及び谷パラメータを予測するステップと、
前記予測されたF_0の山及び谷パラメータに基づいてF_0輪郭を生成するステップとを含む、F_0輪郭を合成する方法。
【請求項2】
前記準備するステップが、
言語学的情報とそれに関連する発話データとを含むトレーニングデータセットを準備するステップと、
前記トレーニングデータセット内の前記発話データからF_0輪郭モデルパラメータを抽出するステップと、
前記抽出するステップで抽出された前記F_0輪郭モデルパラメータからF_0の山及び谷パラメータを推定するステップと、
前記推定するステップで推定された前記F_0の山及び谷パラメータと、前記トレーニングデータセット内の前記言語学的情報とを用いて、前記機械学習モデルがF_0の山及び谷パラメータの言語学的情報に対する前記内部依存性を学習するように、前記機械学習モデルをトレーニングするステップとを含む、請求項１に記載のF_0輪郭を合成する方法。
【請求項3】
前記F_0の山及び谷パラメータを推定するステップが、
前記トレーニングデータセット内の各発話のF_0輪郭を、時間軸に沿って直列に並んだ連続した山型パターンで表されるＲＯＮＤＯ－F_0輪郭に変換するステップと、
前記変換するステップで得られた前記ＲＯＮＤＯ－F_0輪郭内におけるF_0の山の位置を特定するステップと、
前記変換するステップで得られた前記ＲＯＮＤＯ－F_0輪郭内で、隣接する全てのF_0の山の間に、隣接するF_0の山とF_0の谷とが予め定められた条件を満たすように、F_0の谷を見出すステップとを含む、請求項２に記載のF_0輪郭を合成する方法。
【請求項4】
前記見出すステップが、
前記ＲＯＮＤＯ－F_0輪郭内のｉ番目のF_0の山と次の山との間にF_0の谷（$t_{v\,fi}$，$\lambda_{v\,fi}$）の初期候補を見出すステップと、
前記初期候補から始めて、前記F_0の谷（$t_{v\,fi}$，$\lambda_{v\,fi}$）が予め定められた条件を満たすまで$t_{v\,fi}$を所定の時間間隔で減じることにより、前記ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップとを含む、請求項３に記載のF_0輪郭を合成する方法。
【請求項5】
前記初期候補を見出すステップが、（$/t_{v\,i}$，$/\lambda_{v\,i}$）（以下本文中の「/」は上付きバーを示す）で表される最も低い窪みに、初期候補のF_0の谷（$t_{v\,fi}$，$\lambda_{v\,fi}$）を設定するステップを含む、請求項４に記載のF_0輪郭を合成する方法。
【請求項6】
前記探索するステップが、前記初期候補（$/t_{v\,i}$，$/\lambda_{v\,i}$）から始めて
$\lambda_{v\,fi}-\lambda_{p\,i} \leqq (/\lambda_{v\,i}-\lambda_{p\,i})\times C$、Ｃは所定の定数、
または
$t_{v\,fi}=t_{p\,i}$
となるまで、所定の時間間隔で$t_{v\,fi}$を減じることにより、前記ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップを含む、請求項５に記載のF_0輪郭を合成する方法。
【請求項7】

IPR2026-00272
Krisp Technologies EX1002 Page 358

前記定数Ｃがほぼ０．９５に選ばれる、請求項６に記載のＦ０輪郭を合成する方法。
【請求項8】
前記Ｆ０輪郭Ｆ０（ｔ）及び前記対応するＲＯＮＤＯ－Ｆ０輪郭Λ（ｔ）とが、以下の、時間ｔの関数

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, for\ t \geq 0,$$

ただし

$$A(\lambda) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \lambda \geq 1,$$

かつ

$$\Lambda(t) = A_{r_1}(t) + \sum_{i=1}^{n-1} Min(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t)$$

として定義される、請求項１～請求項７のいずれかに記載のＦ０輪郭を合成する方法。
【請求項9】
前記入力された言語学的情報と、前記生成するステップで生成された前記Ｆ０輪郭とに基づいて、音声を合成するステップをさらに含む、請求項１～請求項８のいずれかに記載のＦ０輪郭を合成する方法。
【請求項10】
所定の声調言語が中国語である、請求項１～請求項９のいずれかに記載のＦ０輪郭を合成する方法。
【請求項11】
コンピュータ上で実行されると、コンピュータに請求項１～請求項１０のいずれかに記載のすべてのステップを行なわせる、コンピュータプログラム。

【発明の詳細な説明】
【技術分野】
【0001】
この発明は音声合成方法に関し、特に、声調言語の音声合成方法及びシステムに関する。
【背景技術】
【0002】
高品質の合成音声を達成するために、テキストを分析して得られた言語学的情報を信頼性をもって伝えるよう、韻律生成の性能を改善するための努力がなされてきた。韻律生成の最も困難な点は、いかにして音声を適切な声調とイントネーションで合成するか、ということである。この韻律要素は、声調言語では特に重要である。
【0003】
中国語は典型的な声調言語の一つである。中国語の韻律的構成体系（complex）において、特に基本周波数（Ｆ０）輪郭に焦点をあてると、最も小さな特徴的な構成要素は声調である。声調が基本となるのは、韻律的構成体系それ自体を含むより大きな構造が、一つまたは二つ以上の声調を特定の形で変形したものだからである。この観点から、中国語のＦ０輪郭の合成では声調と、文脈によるその変動とに焦点を当てることができる。
【0004】
【非特許文献1】チャオ、Ｙ．Ｒ．１９６８、「中国語話し言葉の文法」、バークレー、ＣＡ、カリフォルニア大学出版局（Chao, Y. R., 1968. A Grammar of Spoken Chinese. Berkeley, CA. University of California Press.）
【非特許文献2】シェン、Ｘ．Ｓ．１９９０、「標準中国語の韻律」、カリフォルニア大

IPR2026-00272
Krisp Technologies EX1002 Page 359

学出版局。（Shen, X. S., 1990. The Prosody of Mandarin Chinese. University of California Publications.）
【非特許文献3】シェン、「漢語語調和語調類型」、中国　語文、１９９４、３、２２１－２２８
【0005】
【数1】

(Hànyǔyǔdiàohéyǔdiàolèixíng. *Zhōngguc yǔwén*)

【非特許文献4】シュイ、Ｙ．１９９９、「Ｆ０輪郭の形成及びアライメントに対する声調と焦点の効果」、音声学ジャーナル、２７、５５－１０５。（Xu, Y., 1999. Effects of Tone and Focus on the Formation and Alignment of F0 Contours. Journal of Phonetics, 27, 55-105.）
【非特許文献5】ニ、Ｊ及びヒロセ、Ｋ.２０００、「標準中国語文の基本周波数輪郭の機能的モデリングに対する実験的評価」、ＩＳＣＳＬＰ２０００、北京、３１９－３２２。（Ni, J. and Hirose, K., 2000. Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences. ISCSLP2000. Beijing, 319-322.）
【非特許文献6】ニ、Ｊ及びカワイ、Ｈ.、「パラメータ的モデリングによる声調特徴の抽出と合成ベース分析によるパターンマッチング」、ＩＣＡＳＳＰ２００３、ｐｐ.７２－７５、２００３。（Ni, J. and Kawai, H., "Tone Feature Extraction through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching," ICASSP2003, pp. 72-75, 2003.）
【非特許文献7】カワハラ、Ｈ.、イクヨ、Ｍ．Ｋ．、チェイニー、Ａ．１９９９、「ピッチ適応時間周波数平滑化及び瞬時周波数ベースのＦ０抽出を用いた音声表現の再構築：音声における反復構造の果たし得る役割」、音声コミュニケーション、２７、１８７－２０７。（Kawahara, H., Ikuyo, M. K., Cheneigne, A., 1999. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication, 27, 187-207.）
【非特許文献8】コロベール、Ｒ．、ベンジオ、Ｓ．及びマリソン、Ｊ.、「トーチ：モジュラー機械学習ソフトウェアライブラリ」、技術報告ＩＤＩＡＰ、ｐｐ.１－９、２００２。（Collobert, R., Bengio, S., and Mariethon, J., "Torch: a Modular Machine Learning Software Library," Technical Report IDIAP, pp. 1-9, 2002.）
【発明の開示】
【発明が解決しようとする課題】
【0006】
ピッチターゲットは、中国語の声調及びイントネーションの表出において重要な役割を果たしていると考えられる。ピッチターゲットは基本的には、高低を含むが、これは英語や日本語のようなアクセント言語のイントネーションを表すのに通常用いられるものである。中国語には、第一声から第四声と呼ばれる四声があり、さらに、第０声と呼ばれる中間的な声調がある。もし話者の音声を、１．低、２．半低、３．中間、４．半高、５．高という点数で表した４個の等しい間隔に分割するとすれば、第一声から第四声はそれぞれ、５５、３５、２１４、及び５１と表される。実際の間隔と絶対ピッチとは共に個々の音声と話すときの気分（ｍｏｏｄ）とに対し相対的なものであるから、この明細書で用いる「ピッチターゲット」という用語は、時間変化に対するＦ$_0$（基本周波数）の山と谷を意味する。
【0007】
他方で、声調とイントネーションとのパターンは、特にＦ$_0$輪郭に関連する。声調パターンの時間範囲は音節のサイズに限定されるのに対し、イントネーションパターンの時間

IPR2026-00272
Krisp Technologies EX1002 Page 360

範囲は音節以上のものをカバーし、発話全体に及ぶ場合もある。

【0008】

声調及びイントネーションについては、多くの研究があり、例えば非特許文献１から非特許文献４等の文献がそうである。過去の知覚試験及び器具を用いた分析から、発話のF_0輪郭は声調とイントネーションを複合的に表し得る、という一致した見解が得られている。

【0009】

しかし、いくつかの基本的な問題について、少なくとも実務的には、明確な解答は得られていない。例えば、中国語の声調とイントネーションとを表すのにピッチターゲットで十分であるのか、またはテキスト―音声変換において自然な音を達成するために、声調とイントネーションとの合成に必要な必須の特徴は何か、といったことは明確でない。このため、自然な中国語音声を合成する信頼のおける方法はなかった。

【0010】

従って、この発明の目的の一つは、高い信頼性をもって自然な音声を合成する方法とコンピュータプログラムとを提供することである。

【課題を解決するための手段】

【0011】

この発明の一つの局面に従えば、所定の声調言語の言語学的情報から基本周波数（F_0）輪郭を合成する方法は、声調基本周波数の山及び谷パラメータの、声調言語の言語学的特徴に対する内部依存性をモデリングするための機械学習モデルを準備するステップと、機械学習モデルを用いて、入力された言語学的情報に対応するF_0の山及び谷パラメータを予測するステップと、予測されたF_0の山及び谷パラメータに基づいてF_0輪郭を生成するステップとを含む。

【0012】

好ましくは、準備するステップは、言語学的情報とそれに関連する発話データとを含むトレーニングデータセットを準備するステップと、トレーニングデータセット内の発話データからF_0輪郭モデルパラメータを抽出するステップと、抽出するステップで抽出されたF_0輪郭モデルパラメータからF_0の山及び谷パラメータを推定するステップと、推定するステップで推定されたF_0の山及び谷パラメータと、トレーニングデータセット内の言語学的情報とを用いて、機械学習モデルがF_0の山及び谷パラメータの言語学的情報に対する内部依存性を学習するように、機械学習モデルをトレーニングするステップとを含む。

【0013】

より好ましくは、F_0の山及び谷パラメータを推定するステップが、トレーニングデータセット内の各発話のF_0輪郭を、時間軸に沿って直列に並んだ連続した山型パターンで表されるＲＯＮＤＯ－F_0輪郭に変換するステップと、変換するステップで得られたＲＯＮＤＯ－F_0輪郭内におけるF_0の山を特定するステップと、変換するステップで得られたＲＯＮＤＯ－F_0輪郭内で、隣接する全てのF_0の山の間に、隣接するF_0の山とF_0の谷とが予め定められた条件を満たすように、F_0の谷を見出すステップとを含む。

【0014】

見出すステップは、ＲＯＮＤＯ－F_0輪郭内のｉ番目のF_0の山と次の山との間にF_0の谷（$t_{v\,fi}$，　$\lambda_{v\,fi}$）の初期候補を見出すステップと、初期候補から始めて、F_0の谷（$t_{v\,fi}$，　$\lambda_{v\,fi}$）が予め定められた条件を満たすまで$t_{v\,fi}$を所定の時間間隔で減じることにより、ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップとを含んでも良い。

【0015】

好ましくは、初期候補を見出すステップが、（$/t_{v\,i}$，$/\lambda_{v\,i}$）で表される最も低い窪みに、初期候補のF_0の谷（$t_{v\,fi}$，$\lambda_{v\,fi}$）を設定するステップを含む。

【0016】

より好ましくは、探索するステップが、初期候補（$/t_{v\,i}$，$/\lambda_{v\,i}$）から始めて$\lambda_{v\,fi}$

IPR2026-00272
Krisp Technologies EX1002 Page 361

$-\lambda_{pi} \leqq (/\lambda_{vi}-\lambda_{pi}) \times C$、Cは所定の定数、または$t_{vfi}=t_{pi}$
となるまで、所定の時間間隔でt_{vfi}を減じることにより、ＲＯＮＤＯ－F_0輪郭上でF_0の谷を探索するステップを含む。
【0017】
定数Ｃがほぼ０．９５となるように選択されても良い。
【0018】
さらに好ましくは、F_0輪郭F_0（ｔ）が、以下の、時間ｔの関数
【0019】
【数2】

$$\frac{\ln F_0(t)-\ln f_{0_b}}{\ln f_{0_t}-\ln f_{0_b}}=\frac{A(\Lambda(t))-A(\lambda_b)}{A(\lambda_t)-A(\lambda_b)}, for\ t\geq 0,$$

ただし

$$A(\lambda)=\frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2+4\zeta^2(1-2\zeta^2)\lambda}}, \lambda\geq 1,$$

かつ

$$\Lambda(t)=A_{r_1}(t)+\sum_{i=1}^{n-1} Min(\Lambda_{f_i}(t),\Lambda_{r_{i+1}}(t))+\Lambda_{f_n}(t)$$

として定義される。
【0020】
好ましくは、この方法は入力された言語学的情報と、生成するステップで生成されたF_0輪郭とに基づいて、音声を合成するステップをさらに含む。
【0021】
所定の声調言語は中国語であっても良い。
【0022】
この発明の別の局面は、コンピュータ上で実行されると、コンピュータに上述の方法のいずれかのすべてのステップを行なわせる、コンピュータプログラムに関する。
【発明を実施するための最良の形態】
【0023】
１．はじめに
この発明は中国語の声調及びイントネーションの表示における、ピッチターゲットの役割に焦点をあてたものである。中国語の声調及びイントネーションを表すのにピッチターゲットで十分であるか否かを調べるために、ピッチターゲットを特に時間変化に関するF_0の山及び谷として測定した。
【0024】
各々がほぼ同一の声調マッピングで女性の母語話者により平叙文と疑問文として２回発話された７２の文に対し、分析及び知覚実験が行なわれた。F_0輪郭から観察された声調及びイントネーションのパターンが、関数モデルを用いて定量的に分析され、その後測定されたピッチターゲットから予測されたモデルパラメータを用いて再合成された。二つの認知実験が行なわれた。一方では、予測された声調及びイントネーションのパターンとピッチターゲット及び原文との類似性を評価した。他方では、２つの平叙文での最終的な声調（第２声及び第４声）のピッチターゲットを体系的に変化させた場合の人間による声調及びイントネーションの知覚を試験した。
【0025】
実験結果は一貫して、ピッチターゲットが中国語の声調及びイントネーションパターンの規定に重要な役割を果たすことを示した。ピッチターゲットが与えられれば、F_0輪郭の正確な形状が予測可能である。この結果に基づき、中国語音声合成方法を構築できる。

IPR2026-00272
Krisp Technologies EX1002 Page 362

まず始めに実験について説明し、明細書の後半でこの発明の実施例を説明する。

【0026】

２．音声試料及び分析方法

２．１．音声試料

ここで用いられた音声データは、７２の中国語文を含み、そのほとんどすべてが非特許文献２から採用されたものである。これらの文を６個のグループに分けた。各々は１２の基本文を含み、これをさらに３つのタイプに細分した。各タイプは４つの文を含み、それらは音節数が等しくさらに全文に対し同一の声調のマッピングで特徴付けられる同じ文法構造となっており、これは表１に示すとおりである。表においてＴ１、Ｔ２、Ｔ３、Ｔ４はそれぞれ、第１声、第２声、第３声及び第４声を示す。

＜表１＞

【0027】

【表1】

タイプ1	タイプ2	タイプ3
T1 T1 T1 T1	T1 T1 T1 T1 T1	T1 T1 T1 T1 T1 T1 T1 T1 T1
T3 T2 T2 T2	T2 T2 T2 T2 T2	T3 T2 T2 T2 T2 T2 T2 T2 T2
T3 T3 T3 T3	T3 T3 T3 T3 T3	T3 T3 T3 T3 T3 T3 T3 T3 T3
T4 T4 T4 T4	T4 T4 T4 T4 T4	T4 T4 T4 T4 T4 T4 T4 T4 T4

タイプ１は主語―動詞（ＳＶ）構造で４つの音節を含む。タイプ２は主語―動詞―目的語（ＳＶＯ）構造で５個の音節を含む。タイプ３はタイプ１及び２の組合せである。すなわち、タイプ２がタイプ１にその文の目的語として付加され、９音節のＳＶＯ構造となっている。これらの文を以下の範疇にグループ分けした。

【0028】

【数3】

Ｓ：平叙文におけるタイプ１、２、３

Ｕ：語彙的にも文法的にもマークされていないイエス／ノーの質問文（以下マーク無し質問文）でのタイプ１、２、３

Ｐ：文の最後の位置に疑問小辞*ma0*（吗）のあるイエス／ノーの質問文(マーク付疑問文)におけるタイプ１、２、３

Ｎ０：*shi4-bu2-shi4*（是-不-是）構造のイエス／ノーの質問文におけるタイプ２

Ｎ１：*X-mei2-X*（X-没-X）構造のイエス／ノーの質問文におけるタイプ２

Ｎ２：*X-le0mei2-X*（X-了没- X）構造のイエス／ノーの質問文におけるタイプ２

Ｑ０：*X-hai2shi4-Y*（X-还是-X)構造の択一的質問文におけるタイプ２

Ｑ１：*shi4-X-hai2shi4-Y*（是-X-还是-Y)構造の質問文におけるタイプ２

Ｑ２：*hai2shi4-X-hai2shi4-Y*（还是-X-还是-Y)構造の質問文におけるタイプ２

Ｗ０：なぜ(*wei4she2me0*)（为什么）の質問文におけるタイプ２

Ｗ１：いつ(*she2me0shi2hou4*)（什么时候)の質問文におけるタイプ２

Ｗ２：何(*she2me0*)（什么）の質問文におけるタイプ２

これら７２の文を、女性話者によって感情表現なしで防音室で２回録音した。

【0029】

２．２．F_0輪郭の関数モデル

この応用では、関数モデルを用いて（非特許文献５を参照）、F_0輪郭をパラメータの形で表す。このモデルによれば、話者の声区（発話の周波数区）はまず、いわゆるＲＯＮＤＯスケール（対数スケールと同様）に変換される。その後ＲＯＮＤＯ－F_0輪郭を時間軸に直列に並んだ連続した山形状のパターンとして表す。F_0輪郭F_0（ｔ）は以下で与

IPR2026-00272
Krisp Technologies EX1002 Page 363

えられる。
【0030】
【数4】

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, \text{for } t \geq 0, \quad (1)$$

where

$$A(\lambda) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \lambda \geq 1, \quad (2)$$

and

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \text{Min}(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t). \quad (3)$$

Ｍｉｎ（ｚ１，ｚ２）はｚ１及びｚ２のうち、小さい方を意味する。式（１）及び（２）は合わせて声区の変換を示す。式（３）はＲＯＮＤＯ－Ｆ０輪郭Λ（ｔ）を表し、ここでΛｒｉ（ｔ）及びΛｆｉ（ｔ）はそれぞれｉ番目の山形状パターンの上昇及び下降成分を示す。すなわち
【0031】
【数5】

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t \leq t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for t} \geq t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

$$\text{where } D_{x_i}(t) = (1 + \frac{4.8t}{\Delta t_{x_i}})\, e^{-\frac{4.8t}{\Delta t_{x_i}}}, \text{for } t \geq 0. \quad (6)$$

パラメータζ、λｔ及びλｂはそれぞれともに０．２３７、１及び２に固定され得る。（非特許文献５を参照。）これにより、周波数ドメインに、話者に依存するが発話には依存しない次の２個のパラメータ、
〔ｆ０ｂ，ｆ０ｔ〕：声区の最高及び最低周波数、
が得られ、さらにＲＯＮＤＯ－時間空間内に、発話に依存するが話者に依存しない５個のパラメータ、
ｎ：山形状パターンの数
Δｔｘｉ：ｉ番目の上昇／下降成分の応答時間
Δλｘｉ：ｉ番目の上昇／下降成分の振幅、ｘ∈｛ｒ，ｆ｝
（ｔｐｉ，λｐｉ）：ｉ番目の山形状パターンの山（ピーク）、ｉ＝１，…ｎ
ができる。
【0032】
２．３．方法
観察された１４４個のＦ０輪郭は最初に、非特許文献６の方法を用いて自動的に分析された。その後、Ｆ０の山と谷とを、もとの声調を考慮しながらＦ０輪郭を目で見て調べながらマニュアルで判断した。ある声調に対するＦ０の山の数は声調モデリング（非特許文献６）に従って定められた。その後、隣接する山の間の輪郭を用いて、Ｆ０の谷を決定した。モデルにより生成されたＦ０輪郭により、これらの発話を再合成し、ＳＴＲＡＩＧＨ

IPR2026-00272
Krisp Technologies EX1002 Page 364

Ｔ（非特許文献７）と呼ばれるツールを用いて知覚実験を行なった。３つの分析及び知覚実験を行なった。実験１では、F_0の山及び谷に基づきF_0輪郭の再合成の有効性を分析した。実験２では、F_0の山及び谷の変化と声調及びイントネーションとの相互作用の相関を調査した。実験３はピッチターゲットの変化により声調及びイントネーションが変化し得ることを示す。これらの実験結果に基づき、ピッチターゲットが声調及びイントネーションを規定することについて論じる。

【0033】

３．結果

３．１．声調及びイントネーションパターンの再合成

実験１は、F_0の山及び谷に基づきF_0輪郭の再合成の有効性を調べるために行なわれた。（t_{vi}，λ_{vi}）がｉ番目とｉ＋１番目の山の間にある谷を示すこととする。山が与えられると、F_0輪郭の生成のために必要な他のモデルパラメータが上述のように計算される。

＜表２＞

【0034】

【表2】

	個数	μ_c	σ_c	μ_p	σ_p	μ_e	σ_e
Δt_r	366	0.140	0.047	0.122	0.043	0.022	0.022
$\Delta \lambda_r$	366	0.224	0.147	0.215	0.143	0.013	0.035
Δt_f	382	0.139	0.047	0.134	0.055	0.019	0.027
$\Delta \lambda_f$	382	0.196	0.129	0.188	0.122	0.007	0.015

表２は音声試料から測定された声調に関するサンプルの統計的結果を示すものであり、ここでμ_c及びσ_cはそれぞれマニュアルでチェックされたこれらのモデルパラメータの平均及び分散を示し（チェック済パラメータ）、μ_p及びσ_pはF_0の山及び谷により予測されたもの（予測パラメータ）を示す。μ_e及びσ_eの欄はチェック済パラメータと予測パラメータとの間の誤差の平均及び分散を示す。

【0035】

再合成された声調及びイントネーションパターンと原文との類似性を試験するため、チェック済パラメータを伴う１４４個の再合成発話と、予測パラメータを伴う１４４個の発話とを含む２８８個の刺激対で知覚実験を行なった。刺激は、無音室でヘッドフォンを用いて二人の母語話者に提示された。刺激対を聴いた後、聴者はそれらの声調及びイントネーションの類似度を３点スケール、すなわち０（非常に異なる）、１（似ている）、２（相違なし）で評価した。聴者は判断に先だって刺激を何回も聴くことが許された。チェック済パラメータと予測パラメータとの平均スコアはそれぞれ１．９３と１．８９であり、「非常に異なる」サンプルは生じなかった。この実験結果から、ピッチターゲット、すなわち時間変化に対するF_0の山及び谷は、声調及びイントネーションパターンの特徴を捕捉するのに十分であることが示された。

【0036】

３．２．声調とイントネーションとの相互作用

実験２では、１２個のカテゴリの各々について、F_0の山及び谷の分析により声調とイントネーションとの相互作用を検証した。主な結果を以下に説明する。まず、疑問文における発話のF_0輪郭は、平叙文におけるそれに比べて、多少とも全体に上向きに動いた。この結果は非特許文献２及び非特許文献３の知見と一致する。同一の第１声及び第４声マッピングの発話では、そのF_0の山及び谷は同一の第２声及び第３声のマッピングのものより高い声区に上昇した。図１は平叙文とマーク無しの疑問文で発話された２つの文

【0037】

IPR2026-00272
Krisp Technologies EX1002 Page 365

【数6】

“hong2bi2tou2 mei2 quan2”（红鼻头没权）及び
“guo4lu4ke4 zhao4 xiang4”（过路客照相）
のF_0輪郭の例を示しており、その意味は、それぞれ、（ａ）「赤い鼻は権力を持っていない」及び（ｂ）「通りががりの人が写真を撮る」と、（ｃ）「赤い鼻は権力を持っているか」及び（ｄ）「通りががりの人が写真を撮るか」である。

【0038】

第２に、マーク無しの疑問文とマーク付きの疑問文の両方でイントネーションを表す、文の最後の声調に依存するやり方がある。最後の声調を２つの組にグループ分けしてみる。第２声と第３声、及び第１声と第４声である。前者では上昇部分のF_0の山が高い声区に上げられ、そのため声調の範囲がかなり広がる。しかしながら後者では、F_0の山と谷（もしあれば）が共に高い声区まで上げられるため、声調範囲は狭くなり、F_0の谷のスケールが上に移動する。この現象は図１（ｃ）と（ｄ）に示される例で明らかに観察される。

【0039】

第３に、Ｘ－ｎｏｔ－Ｘ構造のイエス／ノー疑問文、すなわちカテゴリＮ０、Ｎ１、Ｎ２と、ｗｈの疑問文、すなわちカテゴリＷ０、Ｗ１、Ｗ２とでは、上昇―下降パターンが存在する。上昇―下降パターンは基本的には、shi4-bu2-shi4等のように機能語の声調構造にF_0の山と谷とを配することによって表される。図２は４つの例を示す。（ａ）“bao1shen1gong1 shi4-bu2-shi4 ca1 che1?”（年季奉公の労働者は車を掃除するか？）；（ｂ）“lao3shou3zhang3 shi4-bu2-shi4 mai3 jiu3?”（年取った高官は酒を買うか？）；（ｃ）“bao1shen1gong1 ca1-mei2-ca1 che1?”（年季奉公の労働者は車を掃除したか？）；（ｄ）“lao3shou3zhang3 mai3-mei2-mai3 jiu3?”（年取った高官は酒を買ったか？）である。この図から、機能語に含まれる声調が上昇―下降パターンに適合するように調整されることが明らかである。もし声調がこれと衝突する場合は、声調はその基本的形状を失い、例えば図２（ｃ）の音節ｃａ１及びｍｅｉ２等のように、上昇―下降パターンの軌跡に従う。

【0040】

第４に、択一的疑問文（カテゴリＱ０、Ｑ１、Ｑ２）では、機能語 shi4 及び haishi4 に対し「遷移パターン」が用いられる。遷移パターンという用語は、機能語中の声調がかなり狭いF_0範囲をとり、中間の声区に位置付けられることを意味する。これに対し、機能語周辺の句では通常、焦点現象（focus phenomena）（非特許文献５）が観察される。これらの観察から、声調とイントネーションとの相互作用がピッチターゲットにより良好に捕捉されることが明らかに示される。

【0041】

３．３．人工の声調及びイントネーションパターンの知覚

実験３ではピッチターゲットを体系的に変化させながら、声調及びイントネーションの知覚を調査した。図１（ａ）と１（ｂ）とで示された平叙文の２つの発話をキャリア発話として用い、最終的な声調のF_0の山及び谷を２つ／３つの態様で変化させた。

【0042】

第１の態様は、原文に対しF_0の山を（すなわち、モデルパラメータλ_{pi}）をステップサイズ０．１で変化させ、一方F_0の谷は変化無しで固定するというものであった。図３は観察されたF_0輪郭（“＋”シーケンス）とこれらのモデル生成輪郭（実線、人工声調及びイントネーションパターンと称する）とを表す。

【0043】

第２の様態は、F_0の山及び谷の双方を単純にその声調を同じステップサイズ０．１で上昇させるか下降させることによって変化させるというものであった。モデル生成F_0輪郭は図４で記号Ｂ１からＢ６とＤ１からＤ６で示される。

【0044】

第３の様態は、特に第４声について、F_0の山は固定しながらF_0の谷をステップサイ

IPR2026-00272
Krisp Technologies EX1002 Page 366

ズ０．１で上に移動させ、谷を上昇させるというものであった。モデル生成F_0輪郭は図４で記号Ｅ１からＥ３で示される。２つのキャリア発話はすべてのモデル生成F_0輪郭で再合成された。

【0045】

３人の母語話者で、これら発話の知覚試験を行なった。刺激は聴者に２回、ランダムな順序で、無音室でヘッドフォンを通して提示された。刺激を聴いた後、聴者は３つの質問に答えた。

（１）発話は平叙文か疑問文か？

（２）最後の音節は強調されていたか、普通か、ニュートラルか？

（３）最後の音節で聴いたのはどの声調か？

実験結果を表３にまとめた。ここで、“Ｑｕｅ”及び“Ｓｔａ”はそれぞれ「疑問文」と「平叙文」を示し、“Ｅｍｐ”、“Ｎｏｒ”及び“Ｗｅａ”はそれぞれ「強調」「普通」「弱いストレス」を示す。

＜表３＞

【0046】

【表3】

要因	Que(%)	Sta(%)	Emp(%)	Nor(%)	Wea(%)	T1(%)	T2(%)	T3(%)	T4(%)
A1	66.7	33.3	100				100		
A2	66.7	33.3	100				100		
A3	66.7	33.3	66.7	33.3			100		
A4	33.3	66.7	66.7	33.3			100		
A5	16.7	83.3	66.7	33.3			100		
A6		100	33.3	66.7			100		
A7		100		66.7	33.3		66.7	33.3	
B1	100		100			100			
B2	100		83.3	16.7		50	50		
B3	83.3	16.7	100				100		
B4	16.7	83.3	100				100		
B5		100	83.3	16.7			100		
B6		100	16.7	83.3			100		
B7		100	16.7	50	33.3		100		
C1		100	100						100
C2		100	100						100
C3		100	66.7	33.3					100
C4		100	33.3	66.7					100
C5		100	16.7	50	33.3				100
C6		100		66.7	33.3			33.3	66.7
D1	66.7	33.3	100						100
D2		100	100						100
D3		100	66.7	33.3					100
D4		100	33.3	66.7					100
D5		100	16.7	66.7	16.7				100
D6		100		50	50			16.6	83.3
E1	66.7	33.3	83.3	16.7		50			50
E2	66.7	33.3	100						100
E3	66.7	33.3	100						100

この実験から３つの知見を得ることができる。第１に輪郭は、声調、ストレス及びイントネーションを表すことができる〔非特許文献１及び２を参照〕。F_0の山を上昇させると、音節は一貫して強調されたと知覚された。F_0の山を下げると、音節は弱いストレスで知覚された。第２に、これは疑問イントネーションが最終声調に依存して表わされることを証明した。第２声の場合、F_0の山が高いほど、その発話は容易に疑問文と判断された

。第４声の場合、声調が高い声区にある場合のみ、発話は疑問文であると認識された。しかしながら、声調の谷が低い声区にあるときには、聴者は全て、発話を平叙文であると知覚した。実験結果はまた、最終的な声調の特徴は疑問文と平叙文を区別するのに十分でないことを示した。テンポ等の他の特徴もまた、知覚の鍵となる。最後に、声調はそのF_0の山、谷及びそれらの音節との整列によって決定される〔非特許文献１を参照〕。第１声と第３声とはこの実験である条件下で知覚された。加えて、Ｂ１、Ｂ２及びＥ１と印をつけたF_0輪郭が第１声と知覚されたという結果から、第１声は図２（ｃ）に示されるように高い声区で上昇する輪郭を示して、イントネーションを表す必要性を満たすが、その知覚は失われない、という現象を説明する。

【0047】

４．実施例

中国語の声調及びイントネーションパターンを研究するため、良好に設計された音声試料に対しいくつかの分析と知覚実験とを行なった。実験結果は、声調及びイントネーションパターンの規定においてピッチターゲットが重要な役割を果たすことを示した。例えば関数モデルを用いて、F_0の山と谷とから正確なF_0輪郭を予測することができる。この結果に基づき、観察されたF_0輪郭を、それが伝える主たる言語学的及びパラ言語学的情報を失うことなく、F_0の山と谷のシーケンスとして骨格化できると仮定した。以下で説明する実施例はこの思想に基づくものである。

【0048】

４．１．構造

図５はこの発明の一実施例に従った音声合成システムのブロック図である。図５を参照して、システム２０は、トレーニングデータ３０からF_0の山及び谷のデータを抽出するF_0パラメータ抽出モジュール３４を含み、このデータは韻律的特徴と基になる言語学的情報との間を関連付けるために用いられる。システム２０はさらに、関連付けされたパラメータの基になる言語学的情報に対する内部依存性を学習するのに用いられる機械学習モジュール３６と、言語学的情報３２からF_0輪郭を推定し、適切な声調で中国語音声４０を合成するための合成モジュール３８とを含む。

【0049】

トレーニングデータ３０は、言語学的情報と、付随するF_0輪郭データとを備えたテキストを含む。

【0050】

F_0パラメータ抽出モジュール３４は、トレーニングデータ３０の観察されたF_0輪郭からモデルパラメータの最適な推定を達成するための分析合成（ａｎａｌｙｓｉｓ－ｂｙ－ｓｙｎｔｈｅｓｉｓ：ＡＢＳ）ベースの分析モジュール５０と、F_0の谷（t_{vxi}，λ_{vxi}）を推定するためのターゲット探索モジュール５２とを含み、ここでF_0の山はＡＢＳ－ベースの分析モジュール５０の出力で得られる。

【0051】

ＡＢＳ－ベースの分析モジュール５０は観察されたF_0輪郭からモデルパラメータの最適な推定を達成しようとするものである。非特許文献６は、関数モデルに基づき、F_0輪郭から声調の山と下り勾配の特徴とを抽出するためのアルゴリズムを提案している。F_0輪郭を、その基礎になるF_0の山と谷とに信頼性を持って骨格化するために、声調の下り勾配の特徴に関するモデルパラメータ、すなわち、Δt_{xi}、$\Delta\lambda_{xi}$、を再推定するステップでこのアルゴリズムにいくつかの制約を新たに組み入れた。（$\hat{}\lambda_{vi}$，$\hat{}t_{vi}$）（式中では$\hat{}$は各文字の上に付す）が、ｉ番目の山（λ_{pi}，t_{pi}）と、もしあれば次の山との間に観察されたF_0の谷を示すものとする。谷の決定におけるF_0抽出誤差の影響を抑制するため、音声フレームの尤度を考慮する。Δt_{xi}及び$\Delta\lambda_{xi}$を再推定するための制約を以下に列挙する。

【0052】

【数7】

IPR2026-00272
Krisp Technologies EX1002 Page 369

【表3】

$$\lambda_{p_i} + \Delta\lambda_{f_i} \leq \hat{\lambda}_{v_i} \times 1.1, \quad (7)$$
$$\lambda_{p_i} + \Delta\lambda_{f_i} = \lambda_{p_{i+1}} + \Delta\lambda_{r_{i+1}}, \quad (8)$$
$$\Delta\lambda_{f_i} \geq 0.02, \quad (9)$$
$$\Delta\lambda_{r_{i+1}} \geq 0.02, \quad (10)$$
$$0.05 \leq \Delta t_{f_i} \leq (\hat{t}_{v_i} - t_{p_i}) \times 1.1, \quad (11)$$
$$0.05 \leq \Delta t_{r_{i+1}} \leq (t_{p_{i+1}} - \hat{t}_{v_i}) \times 1.1. \quad (12)$$

ターゲット探索モジュール５２はF_0の谷を推定する。２種類のピッチターゲット、F_0の山と谷とが、韻律的特徴と基になる言語学的情報との間の関連付けを行なうための関連付けパラメータとして用いられる。上述の通り、音響分析と知覚試験とから得られた実験結果は一貫して、中国語の声調及びイントネーションパターンを表すのにピッチターゲットで十分であることを示している。ピッチターゲットが与えられれば、関数モデルからF_0輪郭の正確な形状を予測可能である。

【0053】

F_0の山は推定されたモデルパラメータの組から入手可能なので、ターゲットの探索は特に、ｉ番目の上昇または下降成分のいずれかについてF_0の谷（t_{vxi}，λ_{vxi}）に焦点をあて、ｉ番目の山付近のＲＯＮＤＯ輪郭に対して行なわれる。谷の初期候補、例えば（t_{vfi}，λ_{vfi}）は、i番目の山と次のものとの間のＲＯＮＤＯ輪郭で（/t_{vi}，/λ_{vi}）で示される最も低い窪みの最初の組である。その後、

【0054】

【数8】

【表3】

$$\lambda_{v_{f_i}} - \lambda_{p_i} \leq (\bar{\lambda}_{v_i} - \lambda_{p_i}) \times 0.95, \quad (13)$$

または$t_{vfi} = t_{pi}$となるまで、ごく短いステップ間隔（たとえば０．００５秒）t_{vfi}を減少させることで、ＲＯＮＤＯ輪郭に沿ってｉ番目の山に向かって谷の候補を探索する。式（１３）において、定数０．９５は１から０．０５への単位減衰に必要とされる応答時間としてのΔ_{tx}の定義を考慮して決定された。すなわち

【0055】

【数9】

$$D_x(\Delta t_x) = (1 + \alpha\Delta t_x)e^{-\alpha\Delta t_x} = 1 - 0.95, \quad (14)$$

これは式（６）で用いられる$\alpha = 4.8/\Delta t_x$という表現に対応する。もしt_{vfi}とt_{vri+1}（ここで「t_{vri+1}」の「ｉ＋１」は「ｒ」の添え字である。）との差がしきい値より小さい場合、２つの谷は常に２つの谷候補の平均に固定される。

【0056】

機械学習モジュール３６は関連付けパラメータの基になる言語学的情報に対する内部依存性を学習するのに用いられる。非特許文献８に記載の通り、いくつかの有効な機械学習方法が利用可能である。

【0057】

合成モジュール３８は、機械学習モジュール３６で用いられる機械学習方法に関連する回帰アルゴリズムを用いることにより、入力された言語学的情報からピッチターゲットを予測するための機械予測モジュール７０と、機械予測モジュール７０から与えられるF_0の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算するためのパラメータ変換モジュール７２と、モデルパラメータと特定の声区[f_{0b}，f_{0t}]に関数モデルを適用することにより、F_0輪郭を合成するためのモーダルベースの合成モジュール７４と、入力された言語学的情報３２とモーダルベースの合成モジュール７４から与えられるF_0輪郭に基づく適切な声調とに従って音声を合成するための音声合成モジュール７６とを含む。これらのモジュールの各々を以下で説明する。

【0058】

IPR2026-00272
Krisp Technologies EX1002 Page 370

パラメータ変換モジュール７２は所与のＦ０の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算する。具体的には、以下の通りである。

【0059】

【数10】

$$\Delta t_{r_i} = \max(0.05, t_{p_i} - t_{v_{r_i}}), \quad (15)$$
$$\Delta \lambda_{r_i} = \max(0.02, (\lambda_{v_{r_i}} - \lambda_{p_i}) \times 1.05), \quad (16)$$
$$\Delta t_{f_i} = \max(0.05, t_{v_{f_i}} - t_{p_i}), \quad (17)$$
$$\Delta \lambda_{f_i} = \max(0.02, (\lambda_{v_{f_i}} - \lambda_{p_i}) \times 1.05). \quad (18)$$

４．２．動作

システム２０は以下のように動作する。動作には３局面がある。トレーニングデータ３０からＦ０パラメータ抽出モジュール３４によりＦ０パラメータを抽出する。Ｆ０パラメータ抽出モジュール３４により抽出されたパラメータで機械学習を行なう。その後言語学的情報３０についてＦ０輪郭を推定し、推定されたＦ０輪郭に従った声調で、言語学的情報３２に基づき中国語音声を合成する。

【0060】

第１の局面では、ＡＢＳ－ベースの分析モジュール５０がＦ０輪郭と言語学的情報とを含む入力されたトレーニングデータ３０を分析し、入力データの声調の山と下り勾配の特徴パラメータとを出力する。

【0061】

ターゲット探索モジュール５２は式（１３）及び（１４）を用いてＦ０の谷を推定する。推定されたＦ０の谷は、第２の局面での機械学習のために、Ｆ０の山とともに集められる。

【0062】

第２の局面では、集められたＦ０の山と谷とがトレーニングデータ３０内の対応する言語学的情報とともに機械学習モジュール３６に与えられ、機械学習モジュール３６は関連付けされたパラメータの基になる言語学的情報への内部依存性を学習する。この局面が終わると、機械学習モジュール３６を用い、入力された言語学的情報に起こる可能性の高いＦ０の山と谷とを推定することが可能になる。

【0063】

第３の局面で、合成モジュール３８は機械学習モジュール３６を用いて言語学的情報から中国語音声４０を合成する。

【0064】

具体的には、機械予測モジュール７０は、言語学的情報３２が与えられると、機械学習モジュール３６を用いて、入力された言語学的情報に起こる可能性の高いピッチターゲットを予測する。

【0065】

パラメータ変換モジュール７２は式（１５）から式（１８）により推定されたＦ０の山及び谷から声調の下り勾配の特徴に関連するモデルパラメータを計算する。

【0066】

モーダルベースの合成モジュール７４は、モデルパラメータ及び声区〔f_{0b}，f_{0t}〕が与えられると、関数モデルを用いてＦ０輪郭を合成する。

【0067】

音声合成モジュール７６は言語学的情報３２及びモーダルベースの合成モジュール７４から与えられるＦ０輪郭を分析し、適切な声調で言語学的情報を伝える中国語音声を生成する。

【0068】

―　結果例　―

IPR2026-00272
Krisp Technologies EX1002 Page 371

図６及び図７は提案されたシステム及び方法を一部示す例であって、F_0輪郭を基となるF_0の山及び谷に骨格化する局面と、F_0の山及び谷から輪郭を復元する局面とを含んでいる。

【0069】

図６（ａ）に示されるような測定されたF_0輪郭（“＋”シーケンス）が与えられるとすると、これらはまず関数モデルに基づき図６（ｂ）に示されるようなパラメータの形で表される。山のパラメータはモデルパラメータの組から利用可能であるが、谷は、ＲＯＮＤＯ時間空間におけるこれらの山の周囲の上昇／下降成分から探索される。図７の上部と図６（ｃ）は山（黒○）と谷（白○）とをプロットしており、これによりF_0輪郭の骨格が与えられる。図６は、それによって輪郭が合成されるモデルパラメータにF_0の山と谷とを変換することによって復元されたF_0輪郭を示し、図７は観察されたF_0輪郭と回復された輪郭とを実線で示す。

【0070】

言語学的情報から中国語のF_0輪郭と音声とを合成するこの実施例のシステム及び方法は関数モデルを使用する。先行技術と比較して、この実施例ではF_0の山及び谷を、韻律的及び言語学的特徴の間を関連付ける主たるパラメータとして導入し、それらの間の内部依存性を学習するのに機械学習技術を導入する。上述の通り、予備的実験によりこの実施例の有効性が確認された。

【0071】

この発明は中国語に関する実施例について説明されてきたが、この発明は、特徴を表す最も小さい構成要素が声調であるような言語であればどのようなものにも適用可能なことは容易に理解できるであろう。

【0072】

上述のシステムはコンピュータハードウェア及びオペレーティングシステム（ＯＳ）上で実行されるソフトウェア（コンピュータプログラム）で実現され得る。音声を合成する場合、スピーカ等の音声生成装置を用いることになる。それ以外に特別なハードウェアの必要はない。

【0073】

このプログラムはＦＤ（フレキシブルディスク）、ＣＤ－ＲＯＭ（コンパクトディスク読出専用メモリ）、ＭＯ（光磁気ディスク）、またはＤＶＤ（ディジタル多用途ディスク）等の記憶媒体に記憶されても良く、またはインターネット等の何らかのデータ通信ネットワーク上を送信しても良い。コンピュータ上で実行されると上述のシステムを実現するようなプログラムであれば、どのようなものでもこの発明の範囲内に含まれる。

【0074】

上述の実施の形態は単なる例示であって制限的なものと解してはならない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】

【0075】

【図1】平叙文で発話された２つの中国語文“hong2bi2tou2　mei2　quan2”　及び　“guo4lu4ke4　zhao4　xiang4”のF_0輪郭の例を示す図である。

【図2】４つの例のイエス／ノー疑問文での上昇―下降パターンを示す図である。

【図3】観察されたF_0輪郭（“＋”シーケンス）と、ある実験において最後のF_0の山を上向きに移動させた場合にモデルから生成された輪郭とを示す図である。

【図4】ある実験において声調を上向きまたは下向きに移動させて生成したF_0輪郭を示す図である。

【図5】この発明の一実施例の音声合成システムのブロック図である。

【図6】F_0輪郭骨格化及び復元の例を示す図である。

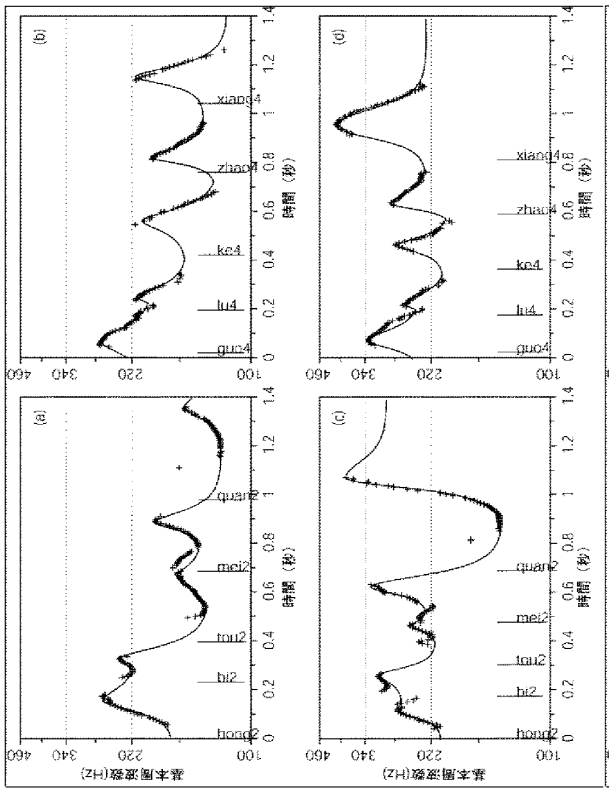
【図7】この発明の一実施例のシステム及び方法の一部を例示する図である。

【符号の説明】

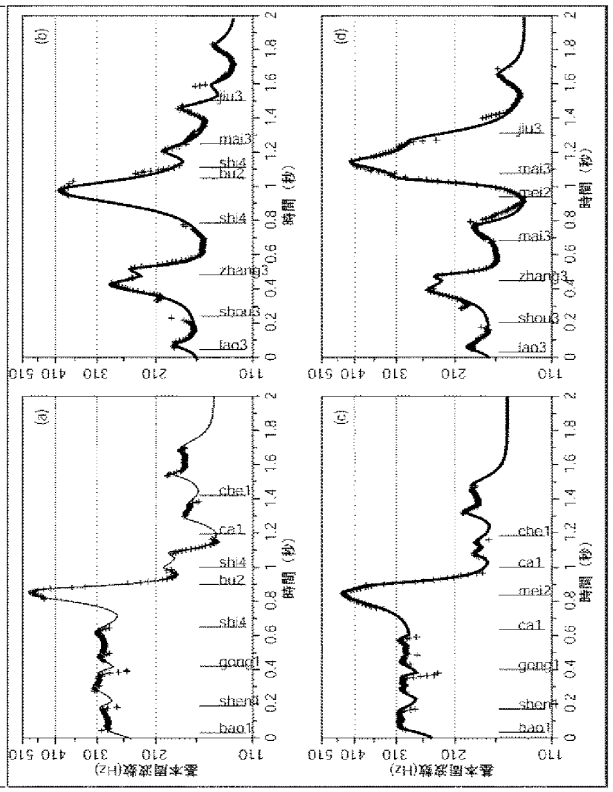
IPR2026-00272
Krisp Technologies EX1002 Page 372

【0076】
２０　中国語音声合成システム、３０　トレーニングデータ、３２　言語学的情報、３４　F_0 パラメータ抽出モジュール、３６　機械学習モジュール、３８　合成モジュール、５０　ＡＢＳ－ベース分析モジュール、５２　ターゲット探索モジュール、７０　機械予測モジュール、７２　パラメータ変換モジュール、７４　モーダルベース合成モジュール、７６　音声合成モジュール

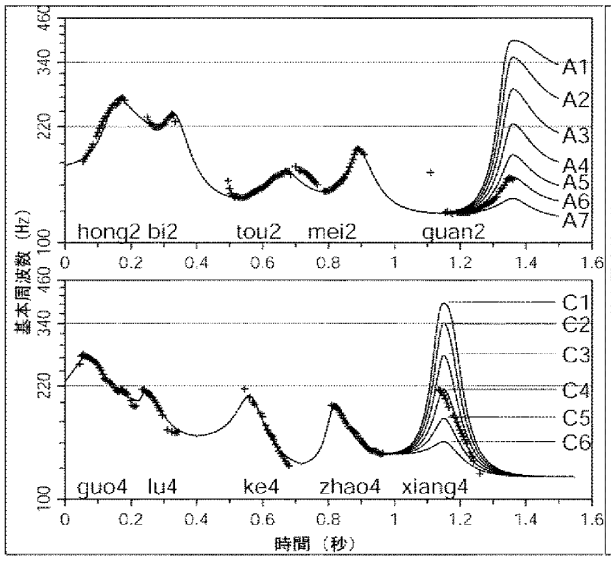
【図1】



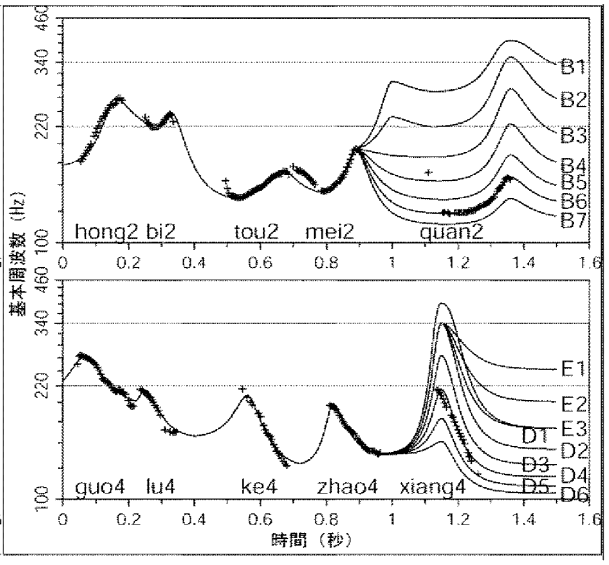
【図2】



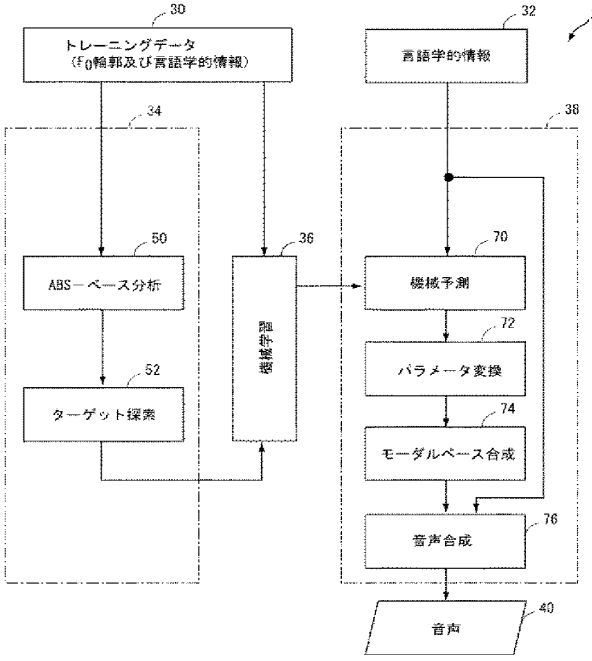
【図3】


基本周波数（Hz）
460
340
220
100
A1
A2
A3
A4
A5
A6
A7
hong2 bi2
tou2
mei2
guan2
C1
C2
C3
C4
C5
C6
guo4
lu4
ke4
zhao4
xiang4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
時間（秒）

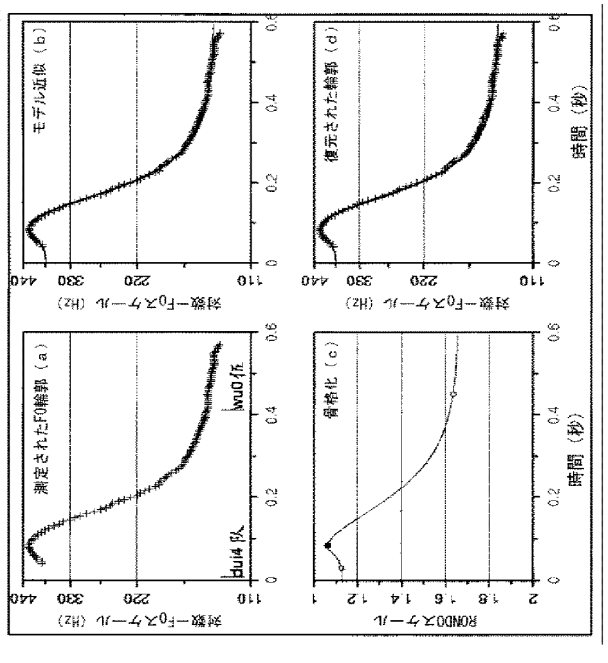
【図4】


基本周波数（Hz）
460
340
220
100
B1
B2
B3
B4
B5
B6
B7
hong2 bi2
tou2
mei2
quan2
E1
E2
E3
D1
D2
D3
D4
D5
D6
guo4
lu4
ke4
zhao4
xiang4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
時間（秒）

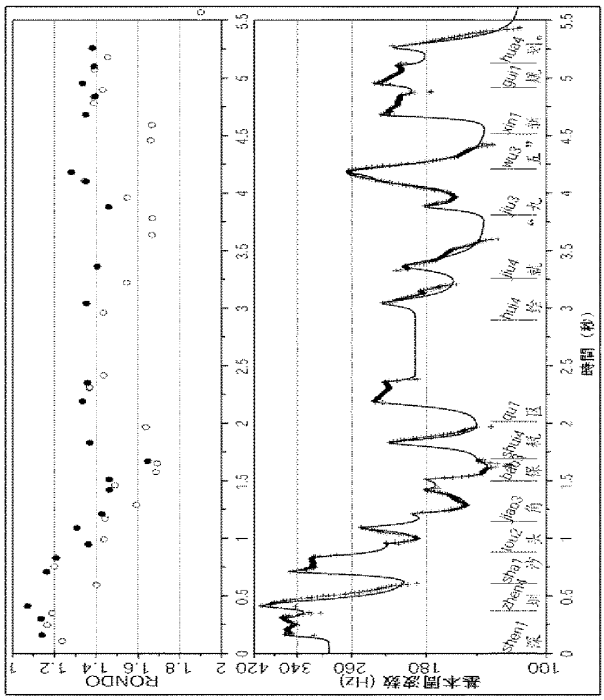
【図5】


30
トレーニングデータ
（F_0輪郭及び言語学的情報）
32
言語学的情報
20
34
38
50
ABS－ベース分析
36
機械学習
70
機械予測
52
ターゲット探索
72
パラメータ変換
74
モーダルベース合成
76
音声合成
40
音声

【図6】


モデル近似（b）
復元された輪郭（d）
測定されたF0輪郭（a）
骨格化（c）
対数－F_0スケール（Hz）
440
330
220
110
0
0.2
0.4
0.6
時間（秒）
RMDOスケール
1
1.2
1.4
1.6
1.8
2

【図7】



【外国語明細書】
2005234418000001.pdf
2005234418000002.pdf
2005234418000003.pdf
2005234418000004.pdf

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference		**FOR FURTHER ACTION** see Form PCT/ISA/220 as well as, where applicable, item 5 below.
Sanas-00007		
International application No.	International filing date *(day/month/year)*	(Earliest) Priority Date *(day/month/year)*
PCT/US2022/028156	**06 May 2022**	**06 May 2021**
Applicant	**SANAS.AI INC.**	

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of **4** sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

   a. With regard to the **language**, the international search was carried out on the basis of:

   ☑ the international application in the language in which it was filed.

   ☐ a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).

   b. ☐ This international search report has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43.6*bis*(a)).

   c. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.

2. ☐ **Certain claims were found unsearchable** (see Box No. II).

3. ☐ **Unity of invention is lacking** (see Box No. III).

4. With regard to the **title**,

   ☑ the text is approved as submitted by the applicant.

   ☐ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

   ☑ the text is approved as submitted by the applicant.

   ☐ the text has been established, according to Rule 38.2, by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the **drawings**,

   a. the figure of the **drawings** to be published with the abstract is Figure No. **2**

   ☑ as suggested by the applicant.

   ☐ as selected by this Authority, because the applicant failed to suggest a figure.

   ☐ as selected by this Authority, because this figure better characterizes the invention.

   b. ☐ none of the figures is to be published with the abstract.

IPR2026-00272
Krisp Technologies EX1002 Page 377

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2022/028156

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/10(2013.01)i; **G10L 15/06**(2006.01)i; **G10L 25/18**(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L 13/10(2013.01); G10L 13/02(2006.01); G10L 13/04(2006.01); G10L 13/08(2006.01); G10L 15/00(2006.01); G10L 19/00(2006.01); G10L 21/013(2013.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: microphone, accent, speech, machine learning, synthesize, mel spectrogram, select, monophone, triphone, linguistic, audio

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	US 2018-0174595 A1 (AMAZON TECHNOLOGIES, INC.) 21 June 2018 (2018-06-21)	
X	paragraphs [0034], [0036]-[0037]; and claim 1	1-2,4-6,10-11,13-15,19
Y		3,8,12,17
A		7,9,16,18,20
	CN 112382267 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 19 February 2021 (2021-02-19)	
Y	claim 7	3,8,12,17
	CN 111462769 A (SPEECHX LTD.) 28 July 2020 (2020-07-28)	
A	claims 1-6	1-20
	JP 2005-234418 A (ADVANCED TELECOMMUNICATION RESEARCH INSTITUTE INTERNATIONAL) 02 September 2005 (2005-09-02)	
A	claims 1-5	1-20

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 August 2022	**17 August 2022**
Name and mailing address of the ISA/KR **Korean Intellectual Property Office** **189 Cheongsa-ro, Seo-gu, Daejeon** **35208, Republic of Korea** Facsimile No. **+82-42-481-8578**	Authorized officer **YANG, Jeong Rok** Telephone No. **+82-42-481-5709**

Form PCT/ISA/210 (second sheet) (July 2019)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2022/028156

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 112382270 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 19 February 2021 (2021-02-19) claims 1-7	1-20

Form PCT/ISA/210 (second sheet) (July 2019)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2022/028156

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2018-0174595 A1	21 June 2018	CN 110199348 A	03 September 2019
		EP 3559942 A1	30 October 2019
		KR 10-2303917 B1	23 September 2021
		US 10163451 B2	25 December 2018
		WO 2018-119145 A1	28 June 2018
CN 112382267 A	19 February 2021	None	
CN 111462769 A	28 July 2020	None	
JP 2005-234418 A	02 September 2005	JP 4223416 B2	12 February 2009
CN 112382270 A	19 February 2021	None	

IPR2026-00272
Krisp Technologies EX1002 Page 380

PATENT COOPERATION TREATY

From the INTERNATIONAL SEARCHING AUTHORITY

To:
KENNEDY, Brandon J.

Lee Sullivan Shea & Smith LLP 656 W. Randolph Street Floor 5W Chicago, Illinois 60661 USA

PCT

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

(PCT Rule 43*bis*.1)

Date of mailing *(day/month/year)*	**17 August 2022**
Applicant's or agent's file reference **Sanas-00007**	**FOR FURTHER ACTION** See paragraph 2 below

International application No.	International filing date *(day/month/year)*	Priority date *(day/month/year)*
PCT/US2022/028156	**06 May 2022**	**06 May 2021**

International Patent Classification (IPC) or both national classification and IPC
G10L 13/10(2013.01)i; **G10L 15/06**(2006.01)i; **G10L 25/18**(2013.01)i

Applicant
SANAS.AI INC.

1. This opinion contains indications relating to the following items:
   - [✓] Box No. I Basis of the opinion
   - [ ] Box No. II Priority
   - [ ] Box No. III Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
   - [ ] Box No. IV Lack of unity of invention
   - [✓] Box No. V Reasoned statement under Rule 43*bis*.1(a)(i) with regard to novelty, inventive step and industrial applicability; citations and explanations supporting such statement
   - [ ] Box No. VI Certain documents cited
   - [ ] Box No. VII Certain defects in the international application
   - [ ] Box No. VIII Certain observations on the international application

2. **FURTHER ACTION**

If a demand for international preliminary examination is made, this opinion will be considered to be a written opinion of the International Preliminary Examining Authority ("IPEA") except that this does not apply where the applicant chooses an Authority other than this one to be the IPEA and the chosen IPEA has notified the International Bureau under Rule 66.1*bis*(b) that written opinions of this International Searching Authority will not be so considered.

If this opinion is, as provided above, considered to be a written opinion of the IPEA, the applicant is invited to submit to the IPEA a written reply together, where appropriate, with amendments, before the expiration of 3 months from the date of mailing of Form PCT/ISA/220 or before the expiration of 22 months from the priority date, whichever expires later.

For further options, see Form PCT/ISA/220.

Name and mailing address of the ISA/KR	Date of completion of this opinion	Authorized officer
Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon, 35208	**17 August 2022**	**YANG, Jeong Rok**
Facsimile No. **+82-42-481-8578**		Telephone No. **+82-42-481-5709**

Form PCT/ISA/237 (Cover sheet) (revised January 2019)

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

International application No.
PCT/US2022/028156

Box No. I Basis of the opinion

1. With regard to the **language**, this opinion has been established on the basis of:
   - [✓] the international application in the language in which it was filed.
   - [ ] a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).

2. [ ] This opinion has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43*bis*.1(b)).

3. [ ] With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, this opinion has been established on the basis of a sequence listing:
   - a. [ ] forming part of the international application as filed:
     - [ ] in the form of an Annex C/ST.25 text file.
     - [ ] on paper or in the form of an image file.
   - b. [ ] furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
   - c. [ ] furnished subsequent to the international filing date for the purposes of international search only:
     - [ ] in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).
     - [ ] on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

4. [ ] In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

5. Additional comments:

IPR2026-00272
Krisp Technologies EX1002 Page 382

WRITTEN OPINION OF THE INTERNATIONAL SEARCHING AUTHORITY

International application No.
PCT/US2022/028156

Box No. V **Reasoned statement under Rule 43*bis*.1(a)(i) with regard to novelty, inventive step and industrial applicability; citations and explanations supporting such statement**

1. Statement

Novelty (N)	Claims	1-20	YES
	Claims	NONE	NO
Inventive step (IS)	Claims	7,9,16,18,20	YES
	Claims	1-6,8,10-15,17,19	NO
Industrial applicability (IA)	Claims	1-20	YES
	Claims	NONE	NO

2. Citations and explanations :

Reference is made to the following documents:
D1: US 2018-0174595 A1 (AMAZON TECHNOLOGIES, INC.) 21 June 2018
D2: CN 112382267 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 19 February 2021
D3: CN 111462769 A (SPEECHX LTD.) 28 July 2020
D4: JP 2005-234418 A (ADVANCED TELECOMMUNICATION RESEARCH INSTITUTE INTERNATIONAL) 02 September 2005
D5: CN 112382270 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 19 February 2021

I. Novelty and Inventive Step (PCT Article 33(2) and (3))

1. Claims 1, 10, 19

D1, which is considered to be the closest prior art to the subject matter of claim 1, discloses a system comprising a microphone (see paragraph [0036]); one or more processors including one or more memories to store a set of instructions, which if executed by the one or more processors, cause the one or more processors to perform operations comprising: receiving a plurality of audio samples; associating each of the plurality of audio samples with at least one of a plurality of accents; determining first combined audio characteristics of a first accent of the plurality of accents by combining first sample audio characteristics of audio samples associated with the first accent; comparing the first combined audio characteristics of the first accent to other audio characteristics associated with at least one other accent of the plurality of accents; generating a translation model between the first accent and a second accent; receiving an input audio portion in a first spoken language; and determining whether the input audio portion is substantially associated with the first accent, and if so, outputting, based at least in part on the translation model, an output audio portion substantially associated with the second accent in the first spoken language (see claim 1).

Claim 1 differs from D1 in that it further comprises "derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a linguistic representation of the received speech content having the first accent".

However, the different feature would be easily derived from the disclosure of D1 (see paragraph [0034]: "the accent translation system 130 includes one or more artificial neural networks, such as a long short-term memory (LSTM) architecture or other systems that may efficiently parse and analyze large amounts of input data. In one embodiment, the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques").

Claim 10 relates to a non-transitory computer-readable medium and claim 19 relates to a method. Since the technical features of claims 10, 19 are substantially the same as those of claim 1, the same reasoning as in claim 1 can be applied to claims 10, 19, respectively.

Therefore, claims 1, 10, 19 lack an inventive step as being obvious to a person skilled in the art over D1.

2. Claims 2-9, 11-18, 20

IPR2026-00272
Krisp Technologies EX1002 Page 383

Box No. V **Reasoned statement under Rule 43*bis*.1(a)(i) with regard to novelty, inventive step and industrial applicability; citations and explanations supporting such statement**

The additional features of claims 2, 11 would be easily derived from the disclosure of D1 (see claim 1: "generating a translation model between the first accent and a second accent; receiving an input audio portion in a first spoken language; and determining whether the input audio portion is substantially associated with the first accent, and if so, outputting, based at least in part on the translation model, an output audio portion substantially associated with the second accent in the first spoken language").

The additional features of claims 4, 13 would be easily derived from the disclosure of D1 (see claim 1: "receiving a plurality of audio samples; associating each of the plurality of audio samples with at least one of a plurality of accents; determining first combined audio characteristics of a first accent of the plurality of accents by combining first sample audio characteristics of audio samples associated with the first accent").

The additional features of claims 5, 14 would be easily derived from the disclosure of D1 (see claim 1: "receiving an input audio portion in a first spoken language; and determining whether the input audio portion is substantially associated with the first accent, and if so, outputting, based at least in part on the translation model, an output audio portion substantially associated with the second accent in the first spoken language").

The additional features of claims 6, 15 would be easily derived from the disclosure of D1 (see paragraph [0037]: "first party and/or the second party that may provide user input that indicates the first party's accent, such as by manually selecting the first accent from a list of available accents or via other user input").

The additional features of claims 3, 8, 12, 17 would be easily derived from the disclosure of D2 (see claim 7: means for synthesizing an audio with a target accent according to the Mel frequency spectrum and a preset neural network vocoder).

The additional features of claims 7, 9, 16, 18, 20 are not disclosed in any of the documents, nor are they obvious to a person skilled in the art over the documents individually or in combination.

Therefore, claims 2, 4-6, 11, 13-15 lack an inventive step as being obvious over D1, and claims 3, 8, 12, 17 lack an inventive step as being obvious over D1 in view of D2, and claims 7, 9, 16, 18, 20 are novel and involve an inventive step.

II. Industrial Applicability (PCT Article 33(4))

Claims 1-20 are industrially applicable.

IPR2026-00272
Krisp Technologies EX1002 Page 384

Electronic Acknowledgement Receipt

EFS ID:	46656778
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	20-SEP-2022
Filing Date:	27-AUG-2021
Time Stamp:	21:11:37
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Information Disclosure Statement (IDS) Form (SB08)	Sanas-00003_IDS.pdf	1263627 c130a7a99589cb35de473884876e973fcc2a79c0	no	4

Warnings:

Information:					
A U.S. Patent Number Citation or a U.S. Publication Number Citation is required in the Information Disclosure Statement (IDS) form for autoloading of data into USPTO systems. You may remove the form to add the required data in order to correct the Informational Message if you are citing U.S. References. If you chose not to include U.S. References, the image of the form will be processed and be made available within the Image File Wrapper (IFW) system. However, no data will be extracted from this form. Any additional data such as Foreign Patent Documents or Non Patent Literature will be manually reviewed and keyed into USPTO systems.					
2	Foreign Reference	CN111462769A.pdf	3835032 14c9ed4c90d4863db621045c146c0396446703b9	no	19
Warnings:					
Information:					
3	Foreign Reference	CN112382267A-Beijing_Youzhuju_Network.pdf	7781813 63a85371ea3e5e156fc35e1a92aaf56a0e97b2f0	no	40
Warnings:					
Information:					
4	Foreign Reference	CN112382270A.pdf	9033035 666aa79b65bac58eab9f1d0ebb8caf409e9e98db	no	43
Warnings:					
Information:					
5	Foreign Reference	JP2005234418A.pdf	1354028 505d738f8422fdd701a5203ff81c7597b814a4b4	no	37
Warnings:					
Information:					
6	Non Patent Literature	2022-08-17_Sanas-00007_ISR_and_Written_Opinion.pdf	332630 ac24c11c27439790b71b068557f4986c2ff06f7c	no	8
Warnings:					
Information:					
		Total Files Size (in bytes):	23600165		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 09/21/2022

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
09/21/2022	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 388

Applicant-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J. KENNEDY	Attorney of Record	

Date of Interview: 15 September 2022

Issues Discussed:

Proposed Amendment(s)

Applicant and Examiner discussed the proposed amendments to the claims, but no agreement was reached.

☑ Attachment

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 389

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

AGENDA FOR TELEPHONE INTERVIEW

Thank you for agreeing to interview the above-referenced application on Thursday, September 15, 2022 at 12:30am ET. An agenda of topics is outlined below.

- Rejection of Claims 1, 10, and 19 under 35 U.S.C. § 103 in the Non-Final Office Action mailed August 10, 2022.

Without conceding the merits of the rejections, and in an attempt to advance prosecution, Applicant submits the following claim amendments and arguments for consideration.

Amendments for Consideration

1. (Proposed Amendment) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

IPR2026-00272
Krisp Technologies EX1002 Page 390

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content~~received speech content having the first accent~~;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content ~~received speech content having the first accent~~, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content ~~accent~~ to an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second accent.

Arguments for Consideration:

- Dirac does not teach "a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content"
- The 103 rejection in the Action fails to consider Applicant's claim as a whole. In particular, the Action points to Gruber for teaching "wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first accent to an updated set of phonemes associated with the second accent." Action at 5. However, Applicant's claim specifies that synthesizing the audio data is "based on the derived non-text linguistic representation..." Gruber involves text to speech synthesis. Gruber at [0045] ("The speech synthesis module 265 synthesizes speech outputs based on text provided by the digital assistant."). Accordingly, Gruber does not teach the "synthesizing" element of Applicant's claim when the claim is considered as a whole.

IPR2026-00272
Krisp Technologies EX1002 Page 391

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003

CONFIRMATION NO. 3314

130085
Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

PUBLICATION NOTICE

OC000000136811954

Title: Real-Time Accent Conversion Model

Publication No.US-2022-0358903-A1
Publication Date:11/10/2022

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publically available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently http://www.uspto.gov/patft/.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Public Records Division. The Public Records Division can be reached by telephone at (571) 272-3150 or (800) 972-6382, by facsimile at (571) 273-3250, by mail addressed to the United States Patent and Trademark Office, Public Records Division, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently https://portal.uspto.gov/pair/PublicPair. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

Office of Data Managment, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

IPR2026-00272
Krisp Technologies EX1002 Page 392

REMARKS

1. Summary of the Office Action

In the Non-Final Office action dated August 10, 2022, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17 and 19 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"), and further in view of U.S. Pub. 2022/0180762 ("Lurie"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interviews

A telephonic Examiner Interview took place on September 15, 2022. Participants included Examiner Bryan Blankenagel and Applicant's representative Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims. No agreement was reached. Applicant thanks the Examiner for his time in conducting the interview.

IPR2026-00272
Krisp Technologies EX1002 Page 393

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 2, 7, 10, 11, 16, and 19. Applicant has added new claim 21. Claims 1-21 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found generally throughout Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac, Gruber, and Lurie. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited combination of references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. Applicant submits that the combination of Dirac, Gruber, and Lurie does not teach at least "wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation . . ." in combination with the other elements of Applicant's amended independent claims 1, 10, and 19.

In the Action at p. 4-5, the Examiner conceded that Dirac does not teach several limitations of Applicant's claims as previously presented, including "wherein synthesizing the audio data comprises . . .," and pointed to Gruber for this teaching. As an initial matter, Applicant submits that Dirac also does not teach "wherein synthesizing the audio data comprises . . ." as amended above. Further, Applicant submits that Gruber does not compensate for the shortcomings of Dirac

IPR2026-00272
Krisp Technologies EX1002 Page 394

in this regard. In particular, Gruber discusses mapping one set of text-based phonemes to another set of text-based phonemes. *See* Gruber at [0125], FIGS. 4, 5A. Accordingly, there is nothing in Gruber that can amount to "mapping the non-text linguistic representation of the set of phonemes . . . to a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

In the Action, the Examiner noted in the response to Applicant's prior arguments that "the linguistic representation as taught in Dirac col. 5 lines 9-27 are accent audio characteristics of the accent sample sets, and are not text." Applicant submits that Dirac's teaching of "accent audio characteristic analysis" in this regard is focused entirely on analyzing (e.g., assigning values to) non-phonetic characteristics of speech, such as pitch characteristics, tone characteristics, stress characteristics, etc., and is silent regarding phonetic characteristics within a given audio sample that could be used as a basis to synthesize speech having a different pronunciation. To the extent that the "other audio characteristics 205A-D" discussed at col. 5 lines 9-27 and shown in the figures of Dirac might contain such a "non-text linguistic representation of a set of phonemes," Dirac provides no teaching at all around using a non-text linguistic representation of a set of phonemes to synthesize speech having a different pronunciation. Moreover, Gruber's squarely text-based techniques clearly do not teach someone of ordinary skill in the art who is starting from Dirac how to use a non-text linguistic representation to synthesize speech having a different pronunciation.

Applicant has also reviewed Lurie and submits that it does not compensate for the combined shortcomings of Dirac and Gruber. In particular, Lurie discloses a server computer that includes "an Accent Conversion Model 118 that uses known audio conversation techniques to convert user input into accent-corrected output." Lurie at [0028]. Lurie is otherwise silent regarding how this conversion is accomplished. Accordingly, nowhere does Lurie teach or suggest

IPR2026-00272
Krisp Technologies EX1002 Page 395

"mapping the non-text linguistic representation of the set of phonemes . . . to a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

Because the combination of Dirac, Gruber, and Lurie does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of claims 1, 10, and 19, over the combination of Dirac, Gruber, and Lurie, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 21 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

5. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: January 2, 2022

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 396

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	01/04/2023	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 21	Minus	** 21	= 0		x $40 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $192 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	SLIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/JULIET MCMILLAN/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of: Sanas.ai Serial No.: 17/460,145 Filing Date: August 27, 2021 Title: Real-Time Accent Conversion Model	Confirmation No.: 3314 Examiner: Blankenagel, Bryan S Group Art Unit: 2658

RESPONSE TO FINAL OFFICE ACTION
MAILED AUGUST 10, 2022

In response to the Non-Final Office Action mailed August 10, 2022, Applicant submits the following amendments and remarks.

Claim Amendments begin at page 2.

Remarks begin at page 12.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 398

CLAIM AMENDMENTS

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content~~received speech content having the first accent~~;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content ~~received speech content having the first accent~~, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the

IPR2026-00272
Krisp Technologies EX1002 Page 399

speech content ~~accent~~ to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content~~accent~~.

2. (Currently Amended) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between [[the]] audio data comprising the first accent and [[the]] audio data comprising the second accent.

3. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 400

4. (Original) The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. (Original) The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Previously Presented) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent.

7. (Currently Amended) The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content ~~having the first~~

IPR2026-00272
Krisp Technologies EX1002 Page 401

~~accent~~ comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. (Original) The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech

IPR2026-00272
Krisp Technologies EX1002 Page 402

content having the second accent between 50-700 ms after receiving the speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content ~~received speech content having the first accent~~;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content ~~received speech content having the first accent~~, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content ~~accent~~ to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

IPR2026-00272
Krisp Technologies EX1002 Page 403

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content ~~accent~~.

11. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between [[the]] audio data comprising the first accent and [[the]] audio data comprising the second accent.

12. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

13. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 404

14. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent.

16. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content ~~having the first accent~~ comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 405

17. (Original) The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

receiving an indication of a first accent;

receiving an indication of a second accent;

IPR2026-00272
Krisp Technologies EX1002 Page 406

receiving, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content ~~received speech content having the first accent~~;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content ~~received speech content having the first accent~~, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content ~~accent~~ to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content ~~accent~~.

20. (Original) The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

IPR2026-00272
Krisp Technologies EX1002 Page 407

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

21. (New) The computing device of claim 1, wherein the speech content having the first accent further comprises a set of prosodic features, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent and the set of prosodic features, and wherein the synthesized version of the received speech content having the second accent has the set of prosodic features.

IPR2026-00272
Krisp Technologies EX1002 Page 408

REMARKS

1. Summary of the Office Action

In the Non-Final Office action dated August 10, 2022, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17 and 19 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"), and further in view of U.S. Pub. 2022/0180762 ("Lurie"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interviews

A telephonic Examiner Interview took place on September 15, 2022. Participants included Examiner Bryan Blankenagel and Applicant's representative Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims. No agreement was reached. Applicant thanks the Examiner for his time in conducting the interview.

IPR2026-00272
Krisp Technologies EX1002 Page 409

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 2, 7, 10, 11, 16, and 19. Applicant has added new claim 21. Claims 1-21 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found generally throughout Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac, Gruber, and Lurie. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited combination of references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. Applicant submits that the combination of Dirac, Gruber, and Lurie does not teach at least "wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation . . ." in combination with the other elements of Applicant's amended independent claims 1, 10, and 19.

In the Action at p. 4-5, the Examiner conceded that Dirac does not teach several limitations of Applicant's claims as previously presented, including "wherein synthesizing the audio data comprises . . .," and pointed to Gruber for this teaching. As an initial matter, Applicant submits that Dirac also does not teach "wherein synthesizing the audio data comprises . . ." as amended above. Further, Applicant submits that Gruber does not compensate for the shortcomings of Dirac

IPR2026-00272
Krisp Technologies EX1002 Page 410

in this regard. In particular, Gruber discusses mapping one set of text-based phonemes to another set of text-based phonemes. *See* Gruber at [0125], FIGS. 4, 5A. Accordingly, there is nothing in Gruber that can amount to "mapping the non-text linguistic representation of the set of phonemes . . . to a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

In the Action, the Examiner noted in the response to Applicant's prior arguments that "the linguistic representation as taught in Dirac col. 5 lines 9-27 are accent audio characteristics of the accent sample sets, and are not text." Applicant submits that Dirac's teaching of "accent audio characteristic analysis" in this regard is focused entirely on analyzing (e.g., assigning values to) non-phonetic characteristics of speech, such as pitch characteristics, tone characteristics, stress characteristics, etc., and is silent regarding phonetic characteristics within a given audio sample that could be used as a basis to synthesize speech having a different pronunciation. To the extent that the "other audio characteristics 205A-D" discussed at col. 5 lines 9-27 and shown in the figures of Dirac might contain such a "non-text linguistic representation of a set of phonemes," Dirac provides no teaching at all around using a non-text linguistic representation of a set of phonemes to synthesize speech having a different pronunciation. Moreover, Gruber's squarely text-based techniques clearly do not teach someone of ordinary skill in the art who is starting from Dirac how to use a non-text linguistic representation to synthesize speech having a different pronunciation.

Applicant has also reviewed Lurie and submits that it does not compensate for the combined shortcomings of Dirac and Gruber. In particular, Lurie discloses a server computer that includes "an Accent Conversion Model 118 that uses known audio conversation techniques to convert user input into accent-corrected output." Lurie at [0028]. Lurie is otherwise silent regarding how this conversion is accomplished. Accordingly, nowhere does Lurie teach or suggest

IPR2026-00272
Krisp Technologies EX1002 Page 411

"mapping the non-text linguistic representation of the set of phonemes . . . to a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

Because the combination of Dirac, Gruber, and Lurie does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of claims 1, 10, and 19, over the combination of Dirac, Gruber, and Lurie, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 21 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

5. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: January 2, 2022

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 412

Electronic Patent Application Fee Transmittal

Application Number:	17460145
Filing Date:	27-Aug-2021
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Filer:	Brandon Jacob Kennedy
Attorney Docket Number:	Sanas-00003

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
CLAIMS IN EXCESS OF 20	2202	1	40	40
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Extension - 2 months with $0 paid	2252	1	256	256
Miscellaneous:				
		Total in USD ($)		**296**

Electronic Acknowledgement Receipt	
EFS ID:	47315642
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Brandon Jacob Kennedy
Filer Authorized By:	
Attorney Docket Number:	Sanas-00003
Receipt Date:	04-JAN-2023
Filing Date:	27-AUG-2021
Time Stamp:	14:24:36
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	$296
RAM confirmation Number	E202314E25113848
Deposit Account	506632
Authorized User	Brandon Kennedy
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: 37 CFR 1.16 (National application filing, search, and examination fees) 37 CFR 1.17 (Patent application and reexamination processing fees)	

37 CFR 1.19 (Document supply fees)

37 CFR 1.20 (Post Issuance fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		Sanas-00003_Response_to_NFOA.pdf	126533 cd7265c6654629c83bc572c0614459b332311ab6	yes	15

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment/Request for Reconsideration-After Non-Final Rejection	1	1
	Claims	2	11
	Applicant Arguments/Remarks Made in an Amendment	12	15

Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	40434 34440091f8c717ad2b5ddcd1309ae631ee052d95	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	166967		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 01/23/2023

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
01/23/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 418

Office Action Summary	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 01/04/2023.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☑ This action is **FINAL.** 2b) ☐ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-21 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-21 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☑ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Response to Arguments

2. Applicant's arguments filed 01/04/2023 have been fully considered but they are not persuasive. Regarding arguments on pages 13-14 of the Remarks, Examiner notes that while Dirac did not teach mapping phonemes from one pronunciation to another, Dirac does teach mapping non-text linguistic representations of the phonemes, as cited in the rejection below.

Claim Rejections - 35 USC § 103

3. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

4. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

5. Claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac, in view of Gruber et

al. (US 2014/0365216 A1), hereinafter referred to as Gruber, and further in view of Lurie et al. (US 2022/0180762 A1), hereinafter referred to as Lurie.

Regarding claim 1, Dirac teaches:

A computing device comprising:

at least one microphone (Fig. 6 element 621, col. 8 lines 16-32, where a microphone is used);

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used);

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used); and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) such that the computing device is configured to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via the at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the audio data comprises mapping the non-linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content to an updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125]) to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Regarding claim 2, Dirac in view of Gruber and Lurie teaches:

The computing device of claim I, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between audio data comprising the first accent and audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines

33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 8, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used), wherein the non-transitory computer- readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receive an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receive, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics

of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content to an updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125]) to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Regarding claim 11, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between audio data comprising the first accent and audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that,

when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 17, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

receiving an indication of a first accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving an indication of a second accent (col. 4 lines 50-62, where the participant identifies their accent);

receiving, via at least one microphone, speech content having the first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics

of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the audio data comprises mapping the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Gruber teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0123], where a phonetic representation is determined);

wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content to an updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the first set of phonemes is mapped to a second set of phonemes);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0125], where the synthesized speech uses the second set of phonemes).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phoneme mapping of Gruber (Gruber para [0125]) on the phonemes of Dirac (Dirac col. 6 line 56 – col. 7 line 6) so that the pronunciation or accent is capable of being used by both speech recognizer and synthesizer (Gruber para [0125]).

Gruber does not explicitly teach that the first and second accents have different pronunciations.

Lurie teaches:

an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation (para [0026-27], where a user learning a new accent encounters some phonemes that are harder to pronounce than others).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber by using the different pronunciations of Lurie (Lurie para [0026]) in the accents of Dirac in view of Gruber (Gruber para [0125]) to identify currently appropriate phrases to keep up with changer user pronunciation skills (Lurie para [0026]).

Regarding claim 21, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the speech content having the first accent further comprises a set of prosodic features, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent and the set of prosodic features, and wherein the synthesized version of the received speech content having the second accent has the set of prosodic features (Dirac col. 12 lines 52-62, where at least one of the characteristics is adjusted, where the one adjusted characteristic is chosen to be pitch, while the stress or prosody remains unadjusted).

6. Claims 3 and 12 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Zhao et al. (Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *INTERSPEECH* (pp. 2843-2847).), hereinafter referred to as Zhao.

Regarding claim 3, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one

processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

Regarding claim 12, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

Zhao teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent (Page 2844 section 3.2, where mel-spectrograms are created to represent the speech).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the mel-spectrograms of Zhao (Zhao page 2844 section 3.2) to represent the speech of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), in order to generate speech waveforms in real-time (Zhao page 2844 section 3).

7. Claims 5 and 14 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

Regarding claim 14, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

8. Claims 7 and 16 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Sajjan et al. (Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.), hereinafter referred to as Sajjan.

Regarding claim 7, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

determine, for each frame in the received speech content, a respective (i) monophone (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent) sound detected in the frame based on the non-text learned linguistic representation for the first accent.

Dirac in view of Gruber and Lurie does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

(ii) triphone (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used) sound detected in the frame

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber and Lurie (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

Regarding claim 16, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received), and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Dirac in view of Gruber and Lurie does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

ii) triphone sound detected in the frame (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used)

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Gruber and Lurie (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

9. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Gruber, and Lurie, and further in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac in view of Gruber and Lurie teaches:

The computing device of claim 1,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac in view of Gruber and Lurie teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Gruber and Lurie does not teach:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that,

when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col.

13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac in view of Gruber and Lurie teaches:

The method of claim 19,

Dirac in view of Gruber and Lurie does not teach:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

Leff teaches:

wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Gruber and Lurie by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Gruber and Lurie (Dirac col. 12 line 52 - col.

13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

10. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 11,361,753 B2 col. 4 line 60 – col. 5 line 9 teaches changing the voice timbre for a target speaker while maintaining the prosody style of a source speaker.

11. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

12. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-11361753-B2	06-2022	Pan; Shifeng	G10L13/10	1/1
	B					
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

IPR2026-00272
Krisp Technologies EX1002 Page 444

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022	08/04/2022	01/17/2023					
	1	✓	✓	✓	✓					
	2	✓	✓	✓	✓					
	3	✓	✓	✓	✓					
	4	✓	✓	✓	✓					
	5	✓	✓	✓	✓					
	6	✓	✓	✓	✓					
	7	✓	✓	✓	✓					
	8	✓	✓	✓	✓					
	9	✓	✓	✓	✓					
	10	✓	✓	✓	✓					
	11	✓	✓	✓	✓					
	12	✓	✓	✓	✓					
	13	✓	✓	✓	✓					
	14	✓	✓	✓	✓					
	15	✓	✓	✓	✓					
	16	✓	✓	✓	✓					
	17	✓	✓	✓	✓					
	18	✓	✓	✓	✓					
	19	✓	✓	✓	✓					
	20	✓	✓	✓	✓					
	21				✓					

IPR2026-00272
Krisp Technologies EX1002 Page 445

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	Date	Examiner
G10L13/02	01/17/2023	BB
G10L15/02	01/17/2023	BB

CPC Combination Sets - Searched*		
Symbol	Date	Examiner

US Classification - Searched*			
Class	Subclass	Date	Examiner

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	Date	Examiner
Performed inventor search in PE2E, EAST	01/17/2023	BB
Performed text search	01/17/2023	BB
Performed text limited CPC search	01/17/2023	BB
Updated search	01/17/2023	BB

Interference Search			
US Class/CPC Symbol	US Subclass/CPC Group	Date	Examiner

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 446

Doc code: IDS
Doc description: Information Disclosure Statement (IDS) Filed

PTO/SB/08a (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT
(Not for submission under 37 CFR 1.99)

Application Number	17460145
Filing Date	2021-08-27
First Named Inventor	Maxim Serebryakov
Art Unit	2658
Examiner Name	Bryan S. Blankenagel
Attorney Docket Number	Sanas-00003

U.S.PATENTS Remove

Examiner Initial*	Cite No	Patent Number	Kind Code[1]	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Patent citation information please click the Add button. Add

U.S.PATENT APPLICATION PUBLICATIONS Remove

Examiner Initial*	Cite No	Publication Number	Kind Code[1]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button. Add

FOREIGN PATENT DOCUMENTS Remove

Examiner Initial*	Cite No	Foreign Document Number[3]	Country Code[2] i	Kind Code[4]	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T[5]
/B.B./	1	111462769	CN	A	2020-07-28	SPEECHX LTD		☒
/B.B./	2	112382267	CN	A	2021-02-19	BEIJING YOUZHUJU NETWORK TECH CO LTD		☒
/B.B./	3	112382270	CN	A	2021-02-19	BEIJING YOUZHUJU NETWORK TECH CO LTD		☒

EFS Web 2.1.18

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	Bryan S. Blankenagel
	Attorney Docket Number	Sanas-00003

/B.B./	4	2005234418	JP	A	2005-09-02	ATR ADVANCED TELECOMM RES INST		☒

If you wish to add additional Foreign Patent Document citation information please click the Add button Add

NON-PATENT LITERATURE DOCUMENTS Remove

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T[5]
/B.B./	1	INTERNATIONAL SEARCHING AUTHORITY, PCT International Search Report and Written Opinion, PCT International Application No. PCT/US2022/028156, dated 17 August 2022, 8 pages.	

If you wish to add additional non-patent literature document citation information please click the Add button Add

EXAMINER SIGNATURE

Examiner Signature	/BRYAN S BLANKENAGEL/	Date Considered	01/17/2023

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. [2] Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). [3] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [4] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [5] Applicant is to place a check mark here if English language translation is attached.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)		
	Application Number	17460145
	Filing Date	2021-08-27
	First Named Inventor	Maxim Serebryakov
	Art Unit	2658
	Examiner Name	Bryan S. Blankenagel
	Attorney Docket Number	Sanas-00003

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☒ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

See attached certification statement.

The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	/Brandon J. Kennedy/	Date (YYYY-MM-DD)	2022-09-20
Name/Print	Brandon J. Kennedy	Registration Number	67,894

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these record s.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.B./

IPR2026-00272
Krisp Technologies EX1002 Page 450

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L33	153	(accent pronounc$3 pronunciation) near5 (chang$3 conver$4 morph$3 shift$3) AND (prosody stress rhythm) near7 (original source)	(US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 05:59 PM
L34	50	(accent pronounc$3 pronunciation) near5 (chang$3 conver$4 morph$3 shift$3) AND (prosody stress rhythm) near7 (original source) SAME (second target)	(US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 05:59 PM
L35	25	(accent pronounc$3 pronunciation) near5 (chang$3 conver$4 morph$3 shift$3) AND (prosody stress rhythm) near7 (original source) SAME (second target) AND phone$2 SAME (accent pronounc$3 pronunciation)	(US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 06:42 PM
L36	25	(accent pronounc$3 pronunciation) near5 (chang$3 conver$4 morph$3 shift$3) AND (prosody stress rhythm) near7 (original source) SAME (second target) AND (phone$2 monophone triphone) SAME (accent pronounc$3 pronunciation)	(US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 06:44 PM
L37	10	36 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 06:44 PM
L38	3	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO;	OR	ON	ON	2023/01/17 06:44 PM

IPR2026-00272
Krisp Technologies EX1002 Page 451

L39	1	"20180174595".pn.	JPO; DERWENT; IBM_TDB) (US-PGPUB; USPAT)	OR	ON	ON	2023/01/17 06:50 PM

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

Web Search History

date, time	web site	search string
1/17/2023 6:47:12 PM	Google Scholar	(accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm)
1/17/2023 6:47:16 PM	Google Scholar	(accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm) [before:2021]
1/17/2023 6:52:07 PM	Google Scholar	[before:2021] (accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm) AND (phone OR phoneme OR triphone OR monophone)
1/17/2023 6:52:21 PM	Google Scholar	[before:2021] (accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm) AND (phone OR phoneme OR triphone OR monophone) AND (source OR original) AND target
1/17/2023 6:53:33 PM	Google Scholar	[before:2021] (accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm) AND (phone OR phoneme OR triphone OR monophone) AND (source OR original) AND (target OR second)
1/17/2023 6:53:47 PM	Google Scholar	[before:2021] (accent OR pronounce OR pronunciation) AND (change OR convert OR morph OR shift) AND (prosody OR stress OR rhythm) AND (phone OR phoneme OR triphone OR monophone) AND (source OR original) AND (target OR second) AND synthesize

IPR2026-00272
Krisp Technologies EX1002 Page 453

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 02/28/2023

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
02/28/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 454

Applicant-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J. KENNEDY	Attorney of Record	
ANKITA JHA	Inventor	

Date of Interview: 23 February 2023

Issues Discussed:

Proposed Amendment(s)

Applicant and Examiner discussed the prior art rejection as well as proposed amendments, and Examiner agreed that the proposed amendments would overcome the prior art rejection.

☑ Attachment

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 455

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

AGENDA FOR TELEPHONE INTERVIEW

Thank you for agreeing to interview the above-referenced application on Thursday, February 23, 2022 at 9:30am ET. An agenda of topics is outlined below.

- Rejection of Claims 1, 10, and 19 under 35 U.S.C. § 103 in the Final Office Action mailed January 23, 2023.

Without conceding the merits of the rejections, and in an attempt to advance prosecution, Applicant submits the following claim amendments and arguments for consideration.

Amendments for Consideration

1. (Proposed Amendment) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

IPR2026-00272
Krisp Technologies EX1002 Page 456

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Remarks for Consideration:

- Applicant submits that the cited references do not teach "mapping each audio frame in the non-text linguistic representation of the set of phonemes . . . to a corresponding frame in a non-text linguistic representation of an updated set of phonemes" in Applicant's claims.
- The Action at p. 4 states that Dirac teaches "wherein synthesizing the audio data comprises mapping the non-linguistic [sic] representation of the set of phonemes associated with the first pronunciation of the speech content to a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation [citation]." However, nothing in Dirac teaches or suggests synthesizing output audio, or any other audio data, that is associated with "a second pronunciation of the speech content that is different from the first pronunciation." Applicant would like to better understand the basis for the Examiner's contention above that Dirac teaches of the "synthesizing" element in Applicant's claims.

IPR2026-00272
Krisp Technologies EX1002 Page 457

REMARKS

1. Summary of the Office Action

In the Final Office action dated January 23, 2023, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"), and further in view of U.S. Pub. 2022/0180762 ("Lurie"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interview

A telephonic Examiner Interview took place on February 23, 2023. Participants included Examiner Bryan Blankenagel, Applicant's representative Ankita Jha, and attorney of record Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims, and the Examiner agreed that the suggested amendments would overcome the current rejections of record. Applicant thanks the Examiner for his time in conducting the interview.

IPR2026-00272
Krisp Technologies EX1002 Page 458

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 10, and 19. Claims 1-21 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found in at least paragraphs [43]-[44] of Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

a. Agreement as to Amended Claims

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac, Gruber, and Lurie. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited combination of references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. Applicant submits, and the Examiner agreed, that the combination of Dirac, Gruber, and Lurie does not teach at least "mapping each audio frame in the non-text linguistic representation of the set of phonemes . . . to a corresponding frame in a non-text linguistic representation of an updated set of phonemes" in combination with the other elements of Applicant's amended independent claims.

Applicant has also reviewed the Zhao, Huffman, Sajjan, and Leff references cited by the Examiner in the Action and respectfully submits that they do not compensate for the shortcomings of Dirac, Gruber, and Lurie.

Because the combination of Dirac, Gruber, and Lurie does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of

IPR2026-00272
Krisp Technologies EX1002 Page 459

claims 1, 10, and 19, over the combination of Dirac, Gruber, and Lurie, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 21 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

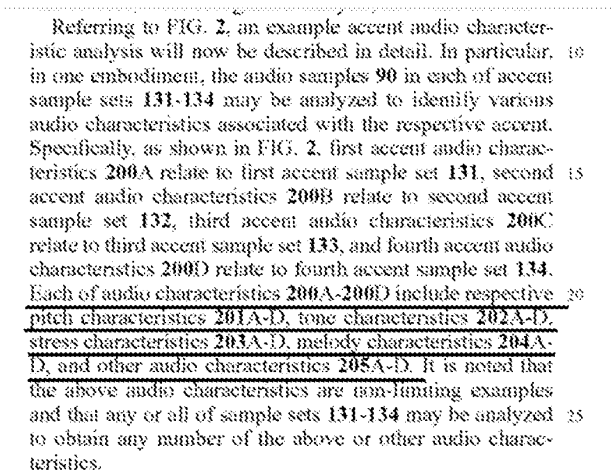
b. Applicant's Additional Arguments Regarding Rejections of Record

Notwithstanding the agreement noted above, Applicant provides for the record at least some of the reasons why the claims as previously presented are not unpatentable over the cited combination of references. The Action at p. 4 states that Dirac teaches "wherein synthesizing the audio data comprises mapping the non-linguistic [sic] representation of the set of phonemes ***associated with the first pronunciation*** of the speech content to a non-text linguistic representation of an updated set of phonemes ***associated with a second pronunciation*** of the speech content that is different from the first pronunciation [citation]." (emphasis added). However, nothing in Dirac teaches or suggests synthesizing output audio, or any other audio data, that is associated with "a second pronunciation of the speech content that is different from the first pronunciation." Applicant has pointed out this deficiency of Dirac in at least its two most recent responses and cannot understand how Examiner is now pointing to Dirac for teaching different pronunciations.

Further, the Examiner has asserted that Dirac discloses "derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (**col. 5 lines 9-27**, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and **col. 7 lines 52-60**, where machine learning is used to update when new samples are received);

IPR2026-00272
Krisp Technologies EX1002 Page 460

Dirac reads as follows at Column 5:

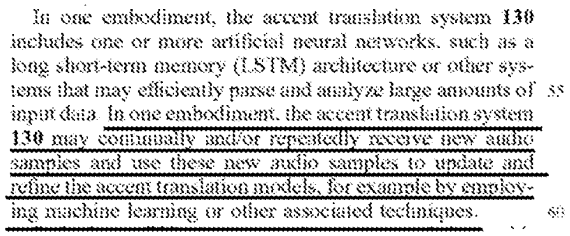


Referring to FIG. **2**, an example accent audio characteristic analysis will now be described in detail. In particular, in one embodiment, the audio samples **90** in each of accent sample sets **131-134** may be analyzed to identify various audio characteristics associated with the respective accent. Specifically, as shown in FIG. **2**, first accent audio characteristics **200**A relate to first accent sample set **131**, second accent audio characteristics **200**B relate to second accent sample set **132**, third accent audio characteristics **200**C relate to third accent sample set **133**, and fourth accent audio characteristics **200**D relate to fourth accent sample set **134**. Each of audio characteristics **200**A-**200**D include respective pitch characteristics **201**A-D, tone characteristics **202**A-D, stress characteristics **203**A-D, melody characteristics **204**A-D, and other audio characteristics **205**A-D. It is noted that the above audio characteristics are non-limiting examples and that any or all of sample sets **131-134** may be analyzed to obtain any number of the above or other audio characteristics.

Here it is clear that Dirac teaches performing Voice Recognition Analysis upon the audio samples in the sample sets to determine audio characteristics for various letters, phonemes, words and other units of speech.

However, Applicant's claims involve using **machine-learning algorithm trained with audio data** comprising the first accent to derive **a non-text linguistic representation of the received speech content having the first accent.**

Dirac reads, at Column 7 – lines 52-60:



In one embodiment, the accent translation system **130** includes one or more artificial neural networks, such as a long short-term memory (LSTM) architecture or other systems that may efficiently parse and analyze large amounts of input data. In one embodiment, the accent translation system **130** may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques.

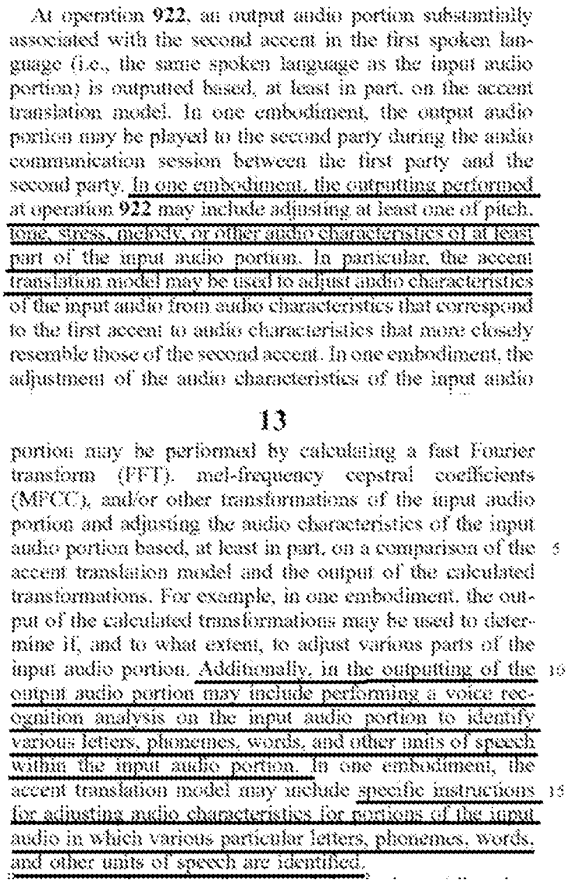
In the above paragraph Dirac talks about using machine learning algorithm to update and refine accent translation models when new audio samples are received. This step comes into picture at a later stage.

IPR2026-00272
Krisp Technologies EX1002 Page 461

However, Applicant's claims involve using machine learning algorithm trained with audio data to derive a non-text linguistic representation of the set of phonemes. Thus, Dirac is not using any machine learning algorithm trained with audio data—Dirac is just performing Voice Recognition Analysis upon the audio samples in the sample sets to determine audio characteristics.

The Examiner has also asserted that Dirac discloses "based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning)

IPR2026-00272
Krisp Technologies EX1002 Page 462



At operation **922**, an output audio portion substantially associated with the second accent in the first spoken language (i.e., the same spoken language as the input audio portion) is outputted based, at least in part, on the accent translation model. In one embodiment, the output audio portion may be played to the second party during the audio communication session between the first party and the second party. In one embodiment, the outputting performed at operation **922** may include adjusting at least one of pitch, tone, stress, melody, or other audio characteristics of at least part of the input audio portion. In particular, the accent translation model may be used to adjust audio characteristics of the input audio from audio characteristics that correspond to the first accent to audio characteristics that more closely resemble those of the second accent. In one embodiment, the adjustment of the audio characteristics of the input audio

13

portion may be performed by calculating a fast Fourier transform (FFT), mel-frequency cepstral coefficients (MFCC), and/or other transformations of the input audio portion and adjusting the audio characteristics of the input audio portion based, at least in part, on a comparison of the accent translation model and the output of the calculated transformations. For example, in one embodiment, the output of the calculated transformations may be used to determine if, and to what extent, to adjust various parts of the input audio portion. Additionally, in the outputting of the output audio portion may include performing a voice recognition analysis on the input audio portion to identify various letters, phonemes, words, and other units of speech within the input audio portion. In one embodiment, the accent translation model may include specific instructions for adjusting audio characteristics for portions of the input audio in which various particular letters, phonemes, words, and other units of speech are identified.

Applicant's claims involve a second machine learning-algorithm trained with audio data comprising the first accent and second accent. However, the portion from Dirac's specification just talks about adjusting the pitch, tone, stress, melody or other audio characteristics.

The Examiner cites to Gruber for the teaching of "wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content to an updated set of phonemes associated with the second pronunciation of the speech content." Action at p. 5. Although this is inconsistent with the Examiner's later statement that "Gruber does not explicitly teach that the first and second accents have different pronunciations" (*id.* at p. 5), Applicant understands the Examiner to be pointing to Gruber for teaching the mapping of phoneme sounds from input speech to phoneme sounds in synthesized output speech, as shown

IPR2026-00272
Krisp Technologies EX1002 Page 463

in FIG. 4 of Gruber. However, Gruber only teaches phoneme mapping techniques that are text-based. Consequently, Gruber provides no insight whatsoever as to how a person of skill in the art, at the time of Applicant's invention, could map "non-text linguistic representations" of phonemes to convert input speech having a first pronunciation into synthesized output speech having a second pronunciation.

Finally, the Examiner points to Lurie for teaching "an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation." *Id.* at p. 5. However, Lurie itself states that it is not innovative in this regard, as the Accent Conversion Model 118 of Lurie "uses known audio conversation techniques to convert user input into accent-corrected output." Lurie at [0028]. Accordingly, Lurie spends no time describing how it performs such conversion and as a result, Lurie does not teach mapping "the non-text linguistic representation of the set of phonemes" to ". . . a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

5. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: March 3, 2023

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 464

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	03/03/2023	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 21	Minus	** 21	= 0		x $40 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $192 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	SLIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/JULIET MCMILLAN/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Doc Code: A.NE.AFCP
Document Description: After Final Consideration Program Request

PTO/SB/434 (3-22)

CERTIFICATION AND REQUEST FOR CONSIDERATION UNDER THE AFTER FINAL CONSIDERATION PILOT PROGRAM 2.0

Practitioner Docket No.:	Application No.:	Filing Date:
Sanas-00003	17/460,145	August 27, 2021
First Named Inventor:	**Title:**	
Maxim Serebryakov	Real-Time Accent Conversion Model	

APPLICANT HEREBY CERTIFIES THE FOLLOWING AND REQUESTS CONSIDERATION UNDER THE AFTER FINAL CONSIDERATION PILOT PROGRAM 2.0 (AFCP 2.0) OF THE ACCOMPANYING RESPONSE UNDER 37 CFR 1.116.

1. The above-identified application is (i) an original utility, plant, or design nonprovisional application filed under 35 U.S.C. 111(a) [a continuing application (*e.g.*, a continuation or divisional application) is filed under 35 U.S.C. 111(a) and is eligible under (i)], or (ii) an international application that has entered the national stage in compliance with 35 U.S.C. 371(c).
2. The above-identified application contains an outstanding final rejection.
3. Submitted herewith is a response under 37 CFR 1.116 to the outstanding final rejection. The response includes an amendment to at least one independent claim, and the amendment does not broaden the scope of the independent claim in any aspect.
4. This certification and request for consideration under AFCP 2.0 is the only AFCP 2.0 certification and request filed in response to the outstanding final rejection.
5. Applicant is willing and available to participate in any interview requested by the examiner concerning the present response.
6. This certification and request is being filed electronically using the USPTO's patent electronic filing system.
7. Any fees that would be necessary consistent with current practice concerning responses after final rejection under 37 CFR 1.116, *e.g.*, extension of time fees, are being concurrently filed herewith. [There is no additional fee required to request consideration under AFCP 2.0.]
8. By filing this certification and request, applicant acknowledges the following:

- Reissue applications and reexamination proceedings are not eligible to participate in AFCP 2.0.
- The examiner will verify that the AFCP 2.0 submission is compliant, *i.e.*, that the requirements of the program have been met (see items 1 to 7 above). For compliant submissions:
  - The examiner will review the response under 37 CFR 1.116 to determine if additional search and/or consideration (i) is necessitated by the amendment and (ii) could be completed within the time allotted under AFCP 2.0. If additional search and/or consideration is required but cannot be completed within the allotted time, the examiner will process the submission consistent with current practice concerning responses after final rejection under 37 CFR 1.116, *e.g.*, by mailing an advisory action.
  - If the examiner determines that the amendment does not necessitate additional search and/or consideration, or if the examiner determines that additional search and/or consideration is required and could be completed within the allotted time, then the examiner will consider whether the amendment places the application in condition for allowance (after completing the additional search and/or consideration, if required). If the examiner determines that the amendment does not place the application in condition for allowance, then the examiner will contact the applicant and request an interview.
    - The interview will be conducted by the examiner, and if the examiner does not have negotiation authority, a primary examiner and/or supervisory patent examiner will also participate.
    - If the applicant declines the interview, or if the interview cannot be scheduled within ten (10) calendar days from the date that the examiner first contacts the applicant, then the examiner will proceed consistent with current practice concerning responses after final rejection under 37 CFR 1.116.

Signature	Date	Practitioner Registration No.
/Brandon J. Kennedy/	March 3, 2023	67894
Name (Print/Typed)	**Telephone Number**	
Brandon J. Kennedy	312-754-9315	

***Note**: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. Submit multiple forms if more than one signature is required, see below*.*

☑ * Total of 1 forms are submitted.

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

CLAIM AMENDMENTS

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via the at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in a non-text linguistic representation of an updated set of phonemes associated

IPR2026-00272
Krisp Technologies EX1002 Page 468

with a second pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

2. (Previously Presented) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between audio data comprising the first accent and audio data comprising the second accent.

3. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 469

4. (Original) The computing device of claim 1, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. (Original) The computing device of claim 1, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Previously Presented) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive a second user input indicating a selection of the second accent.

7. (Previously Presented) The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that are executable by the at least one processor such that the computing device is configured to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that are executable by the at least one processor such that the computing device is configured to:

IPR2026-00272
Krisp Technologies EX1002 Page 470

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. (Original) The computing device of claim 1, further comprising program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Original) The computing device of claim 1, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 471

10. (Currently Amended) A non-transitory computer-readable medium, wherein the non-transitory computer-readable medium is provisioned with program instructions that, when executed by at least one processor, cause a computing device to:

receive an indication of a first accent;

receive an indication of a second accent;

receive, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

convert the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent

IPR2026-00272
Krisp Technologies EX1002 Page 472

comprises the updated set of phonemes associated with the second pronunciation of the speech content.

11. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between audio data comprising the first accent and audio data comprising the second accent.

12. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent.

13. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

14. (Original) The non-transitory computer-readable medium of claim 10, wherein the audio data comprising the second accent corresponds to a single speaker having the second accent.

IPR2026-00272
Krisp Technologies EX1002 Page 473

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a first user input indicating a selection of the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive a second user input indicating a selection of the second accent.

16. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent, and wherein the program instructions that, when executed by at least one processor, cause the computing device to derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that, when executed by at least one processor, cause the computing device to:

determine, for each frame in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 474

17. (Original) The non-transitory computer-readable medium of claim 10, wherein the non-transitory computer-readable medium is also provisioned with program instructions that, when executed by at least one processor, cause the computing device to:

transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Original) The non-transitory computer-readable medium of claim 10, wherein the program instructions that, when executed by at least one processor, cause the computing device to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive, in real time, continuous speech content having the first accent; and

wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to continuously convert the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

receiving an indication of a first accent;

receiving an indication of a second accent;

IPR2026-00272
Krisp Technologies EX1002 Page 475

receiving, via at least one microphone, speech content having the first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

deriving, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent, wherein synthesizing the audio data comprises mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in a non-text linguistic representation of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation; and

converting the synthesized audio data into a synthesized version of the received speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

20. (Original) The method of claim 19, wherein receiving speech content having the first accent comprises receiving, in real time, continuous speech content having the first accent; and

IPR2026-00272
Krisp Technologies EX1002 Page 476

wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises continuously converting the synthesized audio data into a synthesized version of the received speech content having the second accent between 50-700 ms after receiving the speech content having the first accent.

21. (Previously Presented) The computing device of claim 1, wherein the speech content having the first accent further comprises a set of prosodic features, wherein the program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize the audio data representative of the received speech content having the second accent and the set of prosodic features, and wherein the synthesized version of the received speech content having the second accent has the set of prosodic features.

IPR2026-00272
Krisp Technologies EX1002 Page 477

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of: Sanas.ai Serial No.: 17/460,145 Filing Date: August 27, 2021 Title: Real-Time Accent Conversion Model	Confirmation No.: 3314 Examiner: Blankenagel, Bryan S Group Art Unit: 2658

RESPONSE TO FINAL OFFICE ACTION
MAILED JANUARY 23, 2023

In response to the Final Office Action mailed January 23, 2022, Applicant submits the following amendments and remarks pursuant to 37 C.F.R. § 1.116 and the After Final Consideration Pilot Program 2.0.

Claim Amendments begin at page 2.

Remarks begin at page 12.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 478

REMARKS

1. Summary of the Office Action

In the Final Office action dated January 23, 2023, ("the Action") the Examiner (i) rejected claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 under 35 U.S.C. § 103 as allegedly unpatentable over U.S. Pat. 10,163,451 ("Dirac") in view of U.S. Pub. 2014/0365216 ("Gruber"), and further in view of U.S. Pub. 2022/0180762 ("Lurie"); (ii) rejected claims 3 and 12 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" ("Zhao"); (iii) rejected claims 5 and 14 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pat. 10,614,826 ("Huffman"); (iv) rejected claims 7 and 16 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of "Continuous Speech Recognition of Kannada language using triphone modeling" ("Sajjan"); and (v) rejected claims 9, 18, and 20 under 35 U.S.C. § 103 as allegedly unpatentable over Dirac in view of Gruber, and Lurie, and further in view of U.S. Pub. 2016/0203827 ("Leff").

2. Summary of Examiner Interview

A telephonic Examiner Interview took place on February 23, 2023. Participants included Examiner Bryan Blankenagel, Applicant's representative Ankita Jha, and attorney of record Brandon Kennedy. During the interview, the participants discussed the rejections of the claims as well as suggested amendments to the claims, and the Examiner agreed that the suggested amendments would overcome the current rejections of record. Applicant thanks the Examiner for his time in conducting the interview.

IPR2026-00272
Krisp Technologies EX1002 Page 479

3. Status of the Claims

Without conceding the merits of the claim rejections, Applicant has amended claims 1, 10, and 19. Claims 1-21 are pending, of which claims 1, 10, and 19 are independent and the remainder are dependent. Support for the claim amendments can be found in at least paragraphs [43]-[44] of Applicant's specification. No new matter has been introduced by way of these amendments.

4. Response to Rejections of Claims 1, 10, and 19 under 35 U.S.C. § 103

a. Agreement as to Amended Claims

As noted above, the Examiner rejected independent claims 1, 10, and 19 under § 103 as allegedly unpatentable over the combination of Dirac, Gruber, and Lurie. Applicant respectfully disagrees and submits that Applicant's claims as previously presented are not unpatentable over the cited combination of references. Nonetheless, without conceding the merits of the rejections and in an effort to advance prosecution, Applicant has amended the independent claims as detailed above. Applicant submits, and the Examiner agreed, that the combination of Dirac, Gruber, and Lurie does not teach at least "mapping each audio frame in the non-text linguistic representation of the set of phonemes . . . to a corresponding frame in a non-text linguistic representation of an updated set of phonemes" in combination with the other elements of Applicant's amended independent claims.

Applicant has also reviewed the Zhao, Huffman, Sajjan, and Leff references cited by the Examiner in the Action and respectfully submits that they do not compensate for the shortcomings of Dirac, Gruber, and Lurie.

Because the combination of Dirac, Gruber, and Lurie does not teach each and every element of independent claims 1, 10, and 19, the cited combination does not render the independent claims unpatentable. Consequently, Applicant requests withdrawal of the § 103 rejections of

IPR2026-00272
Krisp Technologies EX1002 Page 480

claims 1, 10, and 19, over the combination of Dirac, Gruber, and Lurie, and respectfully submits that the independent claims should be allowed. Further, Applicant submits that dependent claims 2-9, 11-18, and 21 should be allowed as well for at least the reason that they each depend from an allowable independent claim.

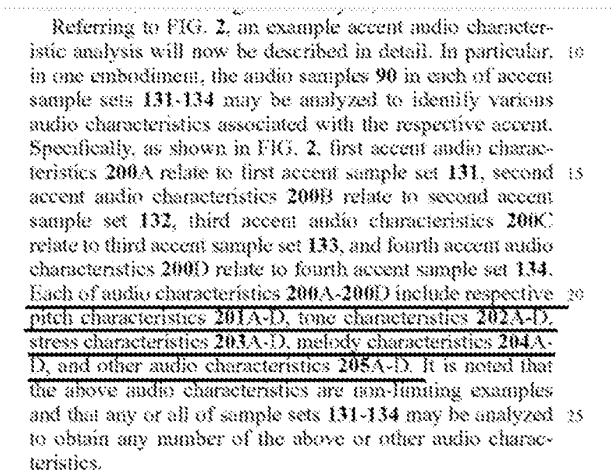
b. Applicant's Additional Arguments Regarding Rejections of Record

Notwithstanding the agreement noted above, Applicant provides for the record at least some of the reasons why the claims as previously presented are not unpatentable over the cited combination of references. The Action at p. 4 states that Dirac teaches "wherein synthesizing the audio data comprises mapping the non-linguistic [sic] representation of the set of phonemes ***associated with the first pronunciation*** of the speech content to a non-text linguistic representation of an updated set of phonemes ***associated with a second pronunciation*** of the speech content that is different from the first pronunciation [citation]." (emphasis added). However, nothing in Dirac teaches or suggests synthesizing output audio, or any other audio data, that is associated with "a second pronunciation of the speech content that is different from the first pronunciation." Applicant has pointed out this deficiency of Dirac in at least its two most recent responses and cannot understand how Examiner is now pointing to Dirac for teaching different pronunciations.

Further, the Examiner has asserted that Dirac discloses "derive, using a first machine-learning algorithm trained with audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (**col. 5 lines 9-27**, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and **col. 7 lines 52-60**, where machine learning is used to update when new samples are received);

IPR2026-00272
Krisp Technologies EX1002 Page 481

Dirac reads as follows at Column 5:

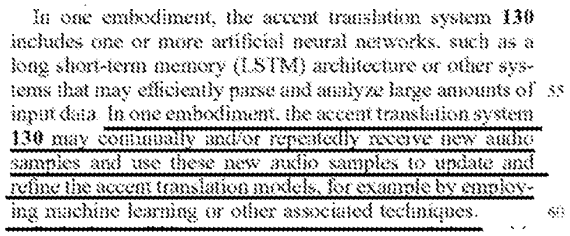


Referring to FIG. **2**, an example accent audio characteristic analysis will now be described in detail. In particular, in one embodiment, the audio samples **90** in each of accent sample sets **131-134** may be analyzed to identify various audio characteristics associated with the respective accent. Specifically, as shown in FIG. **2**, first accent audio characteristics **200**A relate to first accent sample set **131**, second accent audio characteristics **200**B relate to second accent sample set **132**, third accent audio characteristics **200**C relate to third accent sample set **133**, and fourth accent audio characteristics **200**D relate to fourth accent sample set **134**. Each of audio characteristics **200**A-**200**D include respective pitch characteristics **201**A-D, tone characteristics **202**A-D, stress characteristics **203**A-D, melody characteristics **204**A-D, and other audio characteristics **205**A-D. It is noted that the above audio characteristics are non-limiting examples and that any or all of sample sets **131-134** may be analyzed to obtain any number of the above or other audio characteristics.

Here it is clear that Dirac teaches performing Voice Recognition Analysis upon the audio samples in the sample sets to determine audio characteristics for various letters, phonemes, words and other units of speech.

However, Applicant's claims involve using **machine-learning algorithm trained with audio data** comprising the first accent to derive **a non-text linguistic representation of the received speech content having the first accent.**

Dirac reads, at Column 7 – lines 52-60:



In one embodiment, the accent translation system **130** includes one or more artificial neural networks, such as a long short-term memory (LSTM) architecture or other systems that may efficiently parse and analyze large amounts of input data. In one embodiment, the accent translation system **130** may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques.

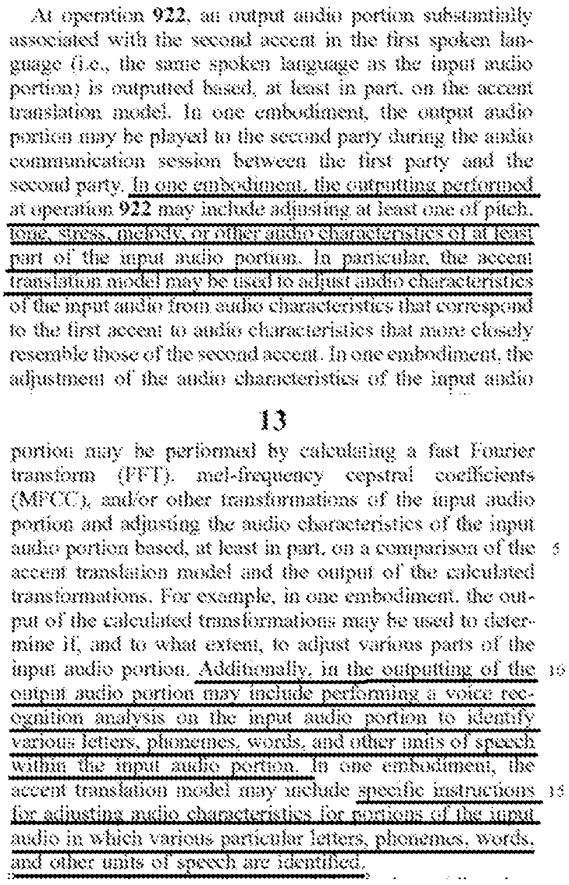
In the above paragraph Dirac talks about using machine learning algorithm to update and refine accent translation models when new audio samples are received. This step comes into picture at a later stage.

IPR2026-00272
Krisp Technologies EX1002 Page 482

However, Applicant's claims involve using machine learning algorithm trained with audio data to derive a non-text linguistic representation of the set of phonemes. Thus, Dirac is not using any machine learning algorithm trained with audio data—Dirac is just performing Voice Recognition Analysis upon the audio samples in the sample sets to determine audio characteristics.

The Examiner has also asserted that Dirac discloses "based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) audio data comprising the first accent and (ii) audio data comprising the second accent, audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning)

IPR2026-00272
Krisp Technologies EX1002 Page 483



At operation **922**, an output audio portion substantially associated with the second accent in the first spoken language (i.e., the same spoken language as the input audio portion) is outputted based, at least in part, on the accent translation model. In one embodiment, the output audio portion may be played to the second party during the audio communication session between the first party and the second party. In one embodiment, the outputting performed at operation **922** may include adjusting at least one of pitch, tone, stress, melody, or other audio characteristics of at least part of the input audio portion. In particular, the accent translation model may be used to adjust audio characteristics of the input audio from audio characteristics that correspond to the first accent to audio characteristics that more closely resemble those of the second accent. In one embodiment, the adjustment of the audio characteristics of the input audio

13

portion may be performed by calculating a fast Fourier transform (FFT), mel-frequency cepstral coefficients (MFCC), and/or other transformations of the input audio portion and adjusting the audio characteristics of the input audio portion based, at least in part, on a comparison of the accent translation model and the output of the calculated transformations. For example, in one embodiment, the output of the calculated transformations may be used to determine if, and to what extent, to adjust various parts of the input audio portion. Additionally, in the outputting of the output audio portion may include performing a voice recognition analysis on the input audio portion to identify various letters, phonemes, words, and other units of speech within the input audio portion. In one embodiment, the accent translation model may include specific instructions for adjusting audio characteristics for portions of the input audio in which various particular letters, phonemes, words, and other units of speech are identified.

Applicant's claims involve a second machine learning-algorithm trained with audio data comprising the first accent and second accent. However, the portion from Dirac's specification just talks about adjusting the pitch, tone, stress, melody or other audio characteristics.

The Examiner cites to Gruber for the teaching of "wherein synthesizing the audio data comprises mapping the set of phonemes associated with the first pronunciation of the speech content to an updated set of phonemes associated with the second pronunciation of the speech content." Action at p. 5. Although this is inconsistent with the Examiner's later statement that "Gruber does not explicitly teach that the first and second accents have different pronunciations" (*id.* at p. 5), Applicant understands the Examiner to be pointing to Gruber for teaching the mapping of phoneme sounds from input speech to phoneme sounds in synthesized output speech, as shown

IPR2026-00272
Krisp Technologies EX1002 Page 484

in FIG. 4 of Gruber. However, Gruber only teaches phoneme mapping techniques that are text-based. Consequently, Gruber provides no insight whatsoever as to how a person of skill in the art, at the time of Applicant's invention, could map "non-text linguistic representations" of phonemes to convert input speech having a first pronunciation into synthesized output speech having a second pronunciation.

Finally, the Examiner points to Lurie for teaching "an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation." *Id.* at p. 5. However, Lurie itself states that it is not innovative in this regard, as the Accent Conversion Model 118 of Lurie "uses known audio conversation techniques to convert user input into accent-corrected output." Lurie at [0028]. Accordingly, Lurie spends no time describing how it performs such conversion and as a result, Lurie does not teach mapping "the non-text linguistic representation of the set of phonemes" to ". . . a non-text linguistic representation of an updated set of phonemes" as found in Applicant's amended claims.

5. Conclusion

For at least the foregoing reasons, Applicant submits that the claims are in condition for allowance. Applicant thus respectfully requests favorable reconsideration and allowance of the claims. Applicant does not acquiesce in any assertion by the Examiner that is not expressly addressed by these remarks. Should the Examiner wish to discuss this case, the Examiner is encouraged to call the undersigned at (312) 754-9315.

Respectfully submitted,

LEE SULLIVAN SHEA & SMITH LLP

Date: March 3, 2023

By: /Brandon J. Kennedy/
Brandon J. Kennedy
Reg. No. 67,894

IPR2026-00272
Krisp Technologies EX1002 Page 485

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/460,145	**03/03/2023 05:55:38 PM ET**	**Sanas-00003**

Title of Invention

Real-Time Accent Conversion Model

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	3314	FILED BY	Holly Mueller
PATENT CENTER #	61687824	FILING DATE	08/27/2021
CUSTOMER #	130085	FIRST NAMED INVENTOR	Maxim Serebryakov
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	Brandon Kennedy

Documents

TOTAL DOCUMENTS: 4

DOCUMENT		PAGES	DESCRIPTION	SIZE (KB)
Sanas_00003_AFCP.pdf		2	After Final Consideration Program Request	268 KB
Sanas-00003_Response_to_FOA.pdf		18	-	991 KB
Sanas-00003_Response_to_FOA-A.NE.pdf	(1-1)	1	Response After Final Action	83 KB
Sanas-00003_Response_to_FOA-CLM.pdf	(2-11)	10	Claims	97 KB

Sanas-00003_Response_to_FOA-REM.pdf	(12-18)	7	Applicant Arguments/Remarks Made in an Amendment	971 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)
Sanas_00003_AFCP.pdf	73342C22ED6D40FF3D524B7A236757EF8A44DB093CA14350DC68E60FC8EF6754DD9E9D8CD85B1928556C103947CD8B9704B30F4C97CE034C62FC88F4BD03EF8A
Sanas-00003_Response_to_FOA.pdf	5F3D497A8BA74396BC7CEE372919B30D049F415AE7F20AB56226C9A184002B0B6BDC544E752901EBB6D30D260E0A1675557391E6DF7CC20FB03F69946CDE1947
Sanas-00003_Response_to_FOA-A.NE.pdf	61765DFA14970D973F691382EC8995EA9EF8A4415914A1D01A4966FDCC91365AB5697BC083B7EC3213C5042F63B59AF62BAA5B7CB36901015A22E85A540F131B
Sanas-00003_Response_to_FOA-CLM.pdf	5EF18E92FD0EFEAE063EFD4764559B16787BA1822D0A0A3A8A51A834AC69A04F493D95BE8CF64B079B80D5F12CDC2FA285BEAFF7A4F6C305A6CD15BA24249A65
Sanas-00003_Response_to_FOA-REM.pdf	71682359A082263E7944978871F26C6D5E40506062891CF37B94DE248D1F66DA164D47AADC98B6D799B830502835209D006ED3CFEEDEF146B60FBAEEC260DB43

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003	3314

130085 7590 03/20/2023

Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
03/20/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

ls3_docketing@cardinal-ip.com
york@ls3ip.com

IPR2026-00272
Krisp Technologies EX1002 Page 489

Advisory Action Before the Filing of an Appeal Brief	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (FITF) Status** Yes

--The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

THE REPLY FILED 03 March 2023 FAILS TO PLACE THIS APPLICATION IN CONDITION FOR ALLOWANCE.

NO NOTICE OF APPEAL FILED

1. ☑ The reply was filed after a final rejection. No Notice of Appeal has been filed. To avoid abandonment of this application, applicant must timely file one of the following replies: (1) an amendment, affidavit, or other evidence, which places the application in condition for allowance; (2) a Notice of Appeal (with appeal fee) in compliance with 37 CFR 41.31; or (3) a Request for Continued Examination (RCE) in compliance with 37 CFR 1.114 if this is a utility or plant application. Note that RCEs are not permitted in design applications. The reply must be filed within one of the following time periods:
   a) ☐ The period for reply expires ____ months from the mailing date of the final rejection.
   b) ☑ The period for reply expires on: (1) the mailing date of this Advisory Action; or (2) the date set forth in the final rejection, whichever is later. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the mailing date of the final rejection.
   c) ☐ A prior Advisory Action was mailed more than 3 months after the mailing date of the final rejection in response to a first after-final reply filed within 2 months of the mailing date of the final rejection.The current period for reply expires ____ months from the mailing date of *the prior Advisory Action* or SIX MONTHS from the mailing date of the final rejection, whichever is earlier.
   Examiner Note: If box 1 is checked, check either box (a), (b) or (c). ONLY CHECK BOX (b) WHEN THIS ADVISORY ACTION IS THE FIRST RESPONSE TO APPLICANTS FIRST AFTER-FINAL REPLY WHICH WAS FILED WITHIN TWO MONTHS OF THE FINAL REJECTION. ONLY CHECK BOX (c) IN THE LIMITED SITUATION SET FORTH UNDER BOX (c). See MPEP 706.07(f).

Extensions of time may be obtained under 37 CFR 1.136(a). The date on which the petition under 37 CFR 1.136(a) and the appropriate extension fee have been filed is the date for purposes of determining the period of extension and the corresponding amount of the fee. The appropriate extension fee under 37 CFR 1.17(a) is calculated from: (1) the expiration date of the shortened statutory period for reply originally set in the final Office action; or (2) as set forth in (b) or (c) above, if checked. Any reply received by the Office later than three months after the mailing date of the final rejection, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

NOTICE OF APPEAL

2. ☐ The Notice of Appeal was filed on _____. A brief in compliance with 37 CFR 41.37 must be filed within two months of the date of filing the Notice of Appeal (37 CFR 41.37(a)), or any extension thereof (37 CFR 41.37(e)), to avoid dismissal of the appeal. Since a Notice of Appeal has been filed, any reply must be filed within the time period set forth in 37CFR 41.37(a).

AMENDMENTS

3. ☑ The proposed amendments filed after a final rejection, but prior to the date of filing a brief, will not be entered because
   a) ☑ They raise new issues that would require further consideration and/or search (see NOTE below);
   b) ☐ They raise the issue of new matter (see NOTE below);
   c) ☐ They are not deemed to place the application in better form for appeal by materially reducing or simplifying the issues for appeal; and/or
   d) ☐ They present additional claims without canceling a corresponding number of finally rejected claims.
   NOTE: See below (See 37 CFR 1.116 and 41.33(a)).
4. ☐ The amendments are not in compliance with 37 CFR 1.121. See attached Notice of Non-Compliant Amendment (PTOL-324).
5. ☐ Applicant's reply has overcome the following rejection(s): _____
6. ☐ Newly proposed or amended claim(s) _____ would be allowable if submitted in a separate, timely filed amendment canceling the non-allowable claim(s).
7. ☑ For purposes of appeal, the proposed amendment(s):(a) ☑ will not be entered, or (b) ☐ will be entered, and an explanation of how the new or amended claims would be rejected is provided below or appended.

AFFIDAVIT OR OTHER EVIDENCE

8. ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____
9. ☐ The affidavit or other evidence filed after final action, but before or on the date of filing a Notice of Appeal will not be entered because applicant failed to provide a showing of good and sufficient reasons why the affidavit or other evidence is necessary and was not earlier presented. See 37 CFR 1.116(e).
10. ☐ The affidavit or other evidence filed after the date of filing the Notice of Appeal, but prior to the date of filing a brief, will not be entered because the affidavit or other evidence failed to overcome all rejections under appeal and/or appellant fails to provide a showing of good and sufficient reasons why it is necessary and was not earlier presented. See 37 CFR 41.33(d)(1).
11. ☐ The affidavit or other evidence is entered. An explanation of the status of the claims after entry is below or attached.

REQUEST FOR RECONSIDERATION/OTHER

12. ☑ The request for reconsideration has been considered but does NOT place the application in condition for allowance because: See Continuation Sheet.
13. ☐ Note the attached Information *Disclosure Statement*(s). (PTO/SB/08) Paper No(s). _____
14. ☑ Other: See attached PTO-413, PTO-892, PTO-2323.

STATUS OF CLAIMS

15. The status of the claim(s) is (or will be) as follows:
    Claim(s) allowed:_____.
    Claim(s) objected to:_____.
    Claim(s) rejected:1-21.
    Claim(s) withdrawn from consideration:_____.

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 490

Continuation of REQUEST FOR RECONSIDERATION/OTHER 12. The request for reconsideration has been considered but does NOT place the application in condition for allowance because: The amended claims would require further search and consideration. Regarding arguments on page 14 of the Remarks, Examiner notes that Dirac teaches mapping a first accent to a second accent, including locations of differing stress, tone/pitch, and melody. While Dirac does not teach the set of phonemes associated with the pronunciation, Examiner notes that Gruber is relied upon to teach the particular limitation, and thus the combination of references is relied upon.

Regarding arguments on pages 15-16 of the Remarks, Examiner notes that col. 7 lines 52-60 of Dirac teach that the system 130 employs neural networks, which repeatedly receive new audio samples and update the accent translation models. This same system 130 includes the sample sets 131-134 that are used to derive the accent audio characteristics in col. 5 lines 9-27, which are interpreted as the non-text linguistic representation. These accent audio characteristics are then used in the accent translation models. The algorithm could then be interpreted as the process of receiving the audio, deriving the accent audio characteristics, and forming the accent translation model, and the algorithm uses machine learning to update. Therefore, the claimed limitation is taught.

Regarding arguments on pages 16-18, Examiner notes that Dirac teaches synthesis of audio in a second accent from a first accent. Gruber teaches mapping a first set of phonemes to a second set of phonemes. Lurie teaches that the two sets of phonemes correspond to different pronunciations. Dirac teaches mapping of one accent to another, including pitch, tone, stress, melody, etc. One of ordinary skill in the art would be able to combine the three references such that the mapping of one accent to another includes phonemes in addition to the characteristics taught by Dirac.

A brief search conducted under the AFCP2.0 guidelines revealed the following references, which appear to be relevant to the amended claims. US 2022/0122579 A1 para [0024] teaches converting speech from a British to an American accent without using speech recognition by converting frames of audio to vectors and using a spectrogram decoder to generate synthesized audio; US 2022/0148562 A1 para [0091] teaches a mel-spectrogram as an acoustic feature in a frame of speech.

Examiner-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
BRANDON J. KENNEDY	Attorney of Record	

Date of Interview: 14 March 2023

Issues Discussed:

Proposed Amendment(s)

Examiner informed Applicant of the results of the brief search conducted under the AFCP2.0 guidelines that revealed the following references, which appear to be relevant to the amended claims. US 2022/0122579 A1 para [0024] teaches converting speech from a British to an American accent without using speech recognition by converting frames of audio to vectors and using a spectrogram decoder to generate synthesized audio; US 2022/0148562 A1 para [0091] teaches a mel-spectrogram as an acoustic feature in a frame of speech.

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: It is not necessary for applicant to provide a separate record of the substance of interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 492

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20220122579-A1	04-2022	Biadsy; Fadi	G10L21/003	1/1
*	B	US-20220148562-A1	05-2022	PARK; Sangjun	G10L13/047	1/1
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

IPR2026-00272
Krisp Technologies EX1002 Page 493

AFCP 2.0 Decision	Application No.	Applicant(s)	
	17/460,145	Serebryakov et al.	
	Examiner	Art Unit	AIA (FITF) Status
	BRYAN S BLANKENAGEL	2658	Yes

This is in response to the After Final Consideration Pilot request filed 03 March 2023.

1. **Improper Request** – The AFCP 2.0 request is improper for the following reason(s) and the after final amendment submitted with the request will be treated under pre-pilot procedure.

   ☐ An AFCP 2.0 request form PTO/SB/434 (or equivalent document) was not submitted.

   ☐ A non-broadening amendment to at least one independent claim was not submitted.

   ☐ The request is not the first proper AFCP 2.0 request submitted in response to the most recent final rejection.

   ☐ Other: _____

2. **Proper Request**

   A. After final amendment submitted with the request will not be treated under AFCP 2.0.
   The after final amendment cannot be reviewed and a search conducted within the guidelines of the pilot program.

   ☐ The after final amendment will be treated under pre-pilot procedure.

   B. Updated search and/or completed additional consideration.
   The examiner performed an updated search and/or completed additional consideration of the after final amendment within the time authorized for the pilot program. The result(s) of the updated search and/or completed additional consideration are:

   ☐ 1. All of the rejections in the most recent final Office action are overcome and a Notice of Allowance is issued herewith.

   ☐ 2. The after final amendment would not overcome all of the rejections in the most recent final Office action . See attached interview summary for further details.

   ☐ 3. The after final amendment was reviewed, and it raises a new issue(s). See attached interview summary for further details.

   ☑ 4. The after final amendment raises new issues, but would overcome all of the rejections in the most recent final Office action. A decision on determining allowability could not be made within the guidelines of the pilot. See attached interview summary for further details, including any newly discovered prior art.

   ☐ 5. Other: _____

   Examiner Note: Please attach an interview summary when necessary as described above.

IPR2026-00272
Krisp Technologies EX1002 Page 494

Do Not Enter /B.B./ 03/13/2023

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
Attorney Docket No. Sanas-00003

In re Application of:	
Sanas.ai	Confirmation No.: 3314
Serial No.: 17/460,145	Examiner: Blankenagel, Bryan S
Filing Date: August 27, 2021	Group Art Unit: 2658
Title: Real-Time Accent Conversion Model	

RESPONSE TO FINAL OFFICE ACTION
MAILED JANUARY 23, 2023

In response to the Final Office Action mailed January 23, 2022, Applicant submits the following amendments and remarks pursuant to 37 C.F.R. § 1.116 and the After Final Consideration Pilot Program 2.0.

Claim Amendments begin at page 2.

Remarks begin at page 12.

Applicant believes that no fees required for the present response. Nonetheless, Applicant authorizes the office to charge any underpayment or credit any overpayment to Deposit Account No. 506632, and to treat any filing in this matter that requires an extension of time as incorporating a request for the extension.

IPR2026-00272
Krisp Technologies EX1002 Page 495

Doc Code: PA..
Document Description: Power of Attorney

PTO/AIA/82A (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

TRANSMITTAL FOR POWER OF ATTORNEY TO ONE OR MORE REGISTERED PRACTITIONERS

NOTE: This form is to be submitted with the Power of Attorney by Applicant form (PTO/AIA/82B) to identify the application to which the Power of Attorney is directed, in accordance with 37 CFR 1.5, unless the application number and filing date are identified in the Power of Attorney by Applicant form. If neither form PTO/AIA/82A nor form PTO/AIA82B identifies the application to which the Power of Attorney is directed, the Power of Attorney will not be recognized in the application.

Application Number	17/460,145
Filing Date	27-Aug-2021
First Named Inventor	Maxim Serebryakov
Title	Real-Time Accent Conversion Model
Art Unit	2658
Examiner Name	BRYAN S BLANKENAGEL
Attorney Docket Number	264248.000100

SIGNATURE of Applicant or Patent Practitioner

Signature	/Nicholas J. Gallo/	Date (Optional)	2023-04-04
Name	Nicholas J. Gallo	Registration Number	60,758
Title (if Applicant is a juristic entity)	Attorney for Applicant		
Applicant Name (if Applicant is a juristic entity)	Sanas.ai Inc.		

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4(d) for signature requirements and certifications. If more than one applicant, use multiple forms.

☑ *Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

DocuSign Envelope ID: D12A9D17-8E8C-44C5-91BE-2C24987FB888

Document Description: Power of Attorney

PTO/AIA/82B (07-13)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY BY APPLICANT

I hereby revoke all previous powers of attorney given in the application identified in either the attached transmittal letter or the boxes below.

Application Number	Filing Date

(Note: The boxes above may be left blank if information is provided on form PTO/AIA/82A.)

☑ I hereby appoint the Patent Practitioner(s) associated with the following Customer Number as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above: 11951

OR

☐ I hereby appoint Practitioner(s) named in the attached list (form PTO/AIA/82C) as my/our attorney(s) or agent(s), and to transact all business in the United States Patent and Trademark Office connected therewith for the patent application referenced in the attached transmittal letter (form PTO/AIA/82A) or identified above. (Note: Complete form PTO/AIA/82C.)

Please recognize or change the correspondence address for the application identified in the attached transmittal letter or the boxes above to:

☑ The address associated with the above-mentioned Customer Number

OR

☐ The address associated with Customer Number:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the Applicant (if the Applicant is a juristic entity, list the Applicant name in the box):

Sanas.ai Inc.

☐ Inventor or Joint Inventor (title not required below)

☐ Legal Representative of a Deceased or Legally Incapacitated Inventor (title not required below)

☑ Assignee or Person to Whom the Inventor is Under an Obligation to Assign (provide signer's title if applicant is a juristic entity)

☐ Person Who Otherwise Shows Sufficient Proprietary Interest (e.g., a petition under 37 CFR 1.46(b)(2) was granted in the application or is concurrently being filed with this document) (provide signer's title if applicant is a juristic entity)

SIGNATURE of Applicant for Patent

The undersigned (whose title is supplied below) is authorized to act on behalf of the applicant (e.g., where the applicant is a juristic entity).

Signature	Maxim Serebryakov	Date (Optional)	
Name	Maxim Serebryakov		
Title	Chief Executive Officer		

NOTE: Signature - This form must be signed by the applicant in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. If more than one applicant, use multiple forms.

☑ Total of 1 forms are submitted.

This collection of information is required by 37 CFR 1.131, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	47791264
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	130085
Filer:	Nicholas Joseph Gallo/Mary Frances Heeks
Filer Authorized By:	Nicholas Joseph Gallo
Attorney Docket Number:	Sanas-00003
Receipt Date:	04-APR-2023
Filing Date:	27-AUG-2021
Time Stamp:	16:42:03
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Power of Attorney	20230404_Gen_POA_and_Transmittal_264248_000100.pdf	337887 6d2766545f339cda297e94ee8688ed453b1719ee	no	2

Warnings:

Information:	
Total Files Size (in bytes):	337887

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
17/460,145	08/27/2021	Maxim Serebryakov	Sanas-00003

CONFIRMATION NO. 3314

130085
Lee Sullivan Shea & Smith LLP
656 W. Randolph St.
Floor 5W
Chicago, IL 60661

POWER OF ATTORNEY NOTICE

OC000000137738268

Date Mailed: 04/14/2023

NOTICE REGARDING CHANGE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 04/04/2023.

- The Power of Attorney to you in this application has been revoked by the applicant. Future correspondence will be mailed to the new address of record(37 CFR 1.33).

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/qnguyen/

IPR2026-00272
Krisp Technologies EX1002 Page 500

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100

CONFIRMATION NO. 3314

11951
Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

POA ACCEPTANCE LETTER

OC000000137738271

Date Mailed: 04/14/2023

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 04/04/2023.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/qnguyen/

IPR2026-00272
Krisp Technologies EX1002 Page 501

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	04/17/2023	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 21	Minus	** 21	= 0		x $40 =	0
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $192 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	0

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/ALLYSON PURNELL/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PTO/SB/30EFS (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

REQUEST FOR CONTINUED EXAMINATION(RCE)TRANSMITTAL
(Submitted Only via EFS-Web)

Application Number	17/460,145	Filing Date	2021-04-27	Docket Number (if applicable)	264248.000100	Art Unit	2658
First Named Inventor	Maxim Serebryakov			Examiner Name	Bryan S. Blankenagel		

This is a Request for Continued Examination (RCE) under 37 CFR 1.114 of the above-identified application.
Request for Continued Examination (RCE) practice under 37 CFR 1.114 does not apply to any utility or plant application filed prior to June 8, 1995, or to any design application. The Instruction Sheet for this form is located at WWW.USPTO.GOV

SUBMISSION REQUIRED UNDER 37 CFR 1.114

Note: If the RCE is proper, any previously filed unentered amendments and amendments enclosed with the RCE will be entered in the order in which they were filed unless applicant instructs otherwise. If applicant does not wish to have any previously filed unentered amendment(s) entered, applicant must request non-entry of such amendment(s).

☐ Previously submitted. If a final Office action is outstanding, any amendments filed after the final Office action may be considered as a submission even if this box is not checked.

- ☐ Consider the arguments in the Appeal Brief or Reply Brief previously filed on _____
- ☐ Other _____

☒ Enclosed

- ☒ Amendment/Reply
- ☐ Information Disclosure Statement (IDS)
- ☐ Affidavit(s)/ Declaration(s)
- ☐ Other _____

MISCELLANEOUS

☐ Suspension of action on the above-identified application is requested under 37 CFR 1.103(c) for a period of months _____
(Period of suspension shall not exceed 3 months; Fee under 37 CFR 1.17(i) required)

☐ Other _____

FEES

☒ **The RCE fee under 37 CFR 1.17(e) is required by 37 CFR 1.114 when the RCE is filed.**
The Director is hereby authorized to charge any underpayment of fees, or credit any overpayments, to Deposit Account No 500436

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT REQUIRED

☒ Patent Practitioner Signature

☐ Applicant Signature

PTO/SB/30EFS (01-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Signature of Registered U.S. Patent Practitioner			
Signature	/Nicholas J. Gallo/	Date (YYYY-MM-DD)	2023-04-17
Name	Nicholas J. Gallo	Registration Number	60,758

This collection of information is required by 37 CFR 1.114. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450.

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

IPR2026-00272
Krisp Technologies EX1002 Page 504

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.

2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.

3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.

4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).

5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.

6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).

7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.

8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.

9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Patent Application Fee Transmittal

Application Number:	17460145
Filing Date:	27-Aug-2021
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Filer:	Nicholas Joseph Gallo/Mary Frances Heeks
Attorney Docket Number:	264248.000100

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
RCE- 2ND AND SUBSEQUENT REQUEST	2820	1	800	800
		Total in USD ($)		**800**

Electronic Acknowledgement Receipt	
EFS ID:	47850553
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	11951
Filer:	Nicholas Joseph Gallo
Filer Authorized By:	
Attorney Docket Number:	264248.000100
Receipt Date:	17-APR-2023
Filing Date:	27-AUG-2021
Time Stamp:	12:17:35
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$800
RAM confirmation Number	E20234GC18206993
Deposit Account	500436
Authorized User	Nicholas Gallo
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: 37 CFR 1.16 (National application filing, search, and examination fees) 37 CFR 1.17 (Patent application and reexamination processing fees)	

37 CFR 1.19 (Document supply fees)

37 CFR 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		20230417_Amendment_Response_Final_Office_Action_264248_000100.pdf	193617 8ce1cafa38b6d773772f3547d7f062ccfde15262	yes	16

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Response After Final Action	1	1
	Claims	2	9
	Applicant Arguments/Remarks Made in an Amendment	10	16

Warnings:

Information:

2	Request for Continued Examination (RCE)	20230417_RCE_264248_000100.pdf	368543 669c36f299ed35db8e8a2671d6fe03001b0d9b05	no	3

Warnings:

This is not a USPTO supplied RCE SB30 form.

Information:

3	Fee Worksheet (SB06)	fee-info.pdf	38124 47d8f915ae036821092adfbf504168266499fe34	no	2

Warnings:

Information:

Total Files Size (in bytes):	600284

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PATENT
Docket No. **264248.000100**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventors	:	Maxim Serebryakov, et al.	Examiner: Bryan S. Blankenagel
Applicant	:	Sanas.ai Inc.	
Serial No.	:	17/460,145	Art Unit: 2658
Cnfrm. No.	:	3314	
Filed	:	April 27, 2021	
For	:	REAL-TIME ACCENT CONVERSION MODEL	

AMENDMENT IN RESPONSE TO FINAL OFFICE ACTION

Mail Stop RCE
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Commissioner:

In response to the January 23, 2023, final Office action, please amend the above-identified patent application as follows:

Amendments to the Claims are reflected in the listing of claims which begins on page 2 of this paper.

Remarks begin on page 11 of this paper.

IPR2026-00272
Krisp Technologies EX1002 Page 511

Amendments to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

1. (Currently Amended) A computing device comprising:

at least one microphone;

at least one processor;

non-transitory computer-readable medium; and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor such that the computing device is configured to:

~~receive an indication of a first accent;~~

~~receive an indication of a second accent;~~

receive, via the at least one microphone, speech content having [[the]] a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the ~~received~~ speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first [[the]] non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content, wherein the first and second phonemes are different phonemes; and

IPR2026-00272
Krisp Technologies EX1002 Page 512

convert the synthesized fourth audio data into a synthesized version of the ~~received~~ speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

2. (Currently Amended) The computing device of claim 1, wherein the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured to ~~synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to~~ apply, to the derived non-text linguistic representation of the ~~received~~ speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

3. (Currently Amended) The computing device of claim 1, wherein the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content ~~synthesize audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to synthesize a set of mel spectrograms representative of the received speech content having the second accent~~.

4. (Currently Amended) The computing device of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. (Currently Amended) The computing device of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Currently Amended) The computing device of claim 1, wherein the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured to ~~receive the indication of the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to~~ receive a first user input indicating a selection of the first accent[[;]] and ~~wherein the program instructions that are executable by the at least one processor such that the computing device is configured to receive the indication of the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to receive~~ a second user input indicating a selection of the second accent.

7. (Currently Amended) The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent[[,]] and ~~wherein~~ the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured to ~~derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that are executable by the at least one processor such that the computing device is configured to:~~ determine, for each frame in the ~~received~~ speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. (Currently Amended) The computing device of claim 1, ~~further comprising~~ wherein the program instructions ~~stored on the non-transitory computer-readable medium that~~ are executable by the at least one processor such that the computing device is further configured to[[:]] transmit the synthesized version of the ~~received~~ speech content having the second accent to a second computing device.

9. (Currently Amended) The computing device of claim 1, wherein the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured ~~to receive speech content having the first accent comprise program instructions that are executable by the at least one processor such that the computing device is configured~~ to:

receive, in real time, continuous speech content having the first accent; and

~~wherein the program instructions that are executable by the at least one processor such that the computing device is configured to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to~~ continuously convert the synthesized fourth audio data into a synthesized version of the ~~received~~ continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium~~, wherein the non-transitory computer-readable medium is~~ provisioned with program instructions that, when executed by at least one processor of a computing device, cause [[a]] the computing device to:

~~receive an indication of a first accent;~~

~~receive an indication of a second accent;~~

receive, via at least one microphone, speech content having [[the]] a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

derive, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the ~~received~~ speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first [[the]] non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the ~~received~~ speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

11. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions [[that]], when executed by the at least one processor, further cause the computing device to ~~synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to~~ apply, to the derived non-text linguistic representation of the ~~received~~ speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

12. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions [[that]], when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content ~~synthesize audio data representative of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to synthesize a set of mel spectrograms representative of the received speech content having the second accent~~.

13. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.

14. (Canceled)

15. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions [[that]], when executed by the at least one processor, further cause the computing device to ~~receive the indication of the first accent comprise program instructions that, when executed by at least one processor, cause the computing device to~~ receive a first user input indicating a selection of the first accent[[;]] and ~~wherein the program instructions that, when executed by at least one processor, cause the computing device to receive the indication of the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to receive~~ a second user input indicating a selection of the second accent.

16. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent[[,]] and ~~wherein~~ the program instructions [[that]], when executed by the at least one processor, further cause the computing device to ~~derive the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content comprise program instructions that, when executed by at least one processor, cause the computing device to:~~ determine, for each frame in the ~~received~~ speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

17. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the ~~non- transitory computer-readable medium is also provisioned with~~ program instructions [[that]], when executed by the at least one processor, further cause the computing device to[[:]] transmit the synthesized version of the ~~received~~ speech content having the second accent to a second computing device.

18. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions [[that]], when executed by the at least one processor, further cause the computing device ~~to receive speech content having the first accent comprise program instructions that, when executed by at least one processor, cause the computing device~~ to:

receive, in real time, continuous speech content having the first accent; and

~~wherein the program instructions that, when executed by at least one processor, cause the computing device to convert the synthesized audio data into a synthesized version of the received speech content having the second accent comprise program instructions that, when executed by at least one processor, cause the computing device to~~ continuously convert the synthesized fourth audio data into a synthesized version of the ~~received~~ continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

19. (Currently Amended) A method carried out by a computing device, the method comprising:

~~receiving an indication of a first accent;~~

~~receiving an indication of a second accent;~~

receiving, via at least one microphone, speech content having [[the]] a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

deriving, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the ~~received~~ speech content, synthesizing, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the ~~received~~ speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first [[the]] non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content, wherein the first and second phonemes are different phonemes; and

IPR2026-00272
Krisp Technologies EX1002 Page 518

converting the synthesized fourth audio data into a synthesized version of the ~~received~~ speech content having the second accent, wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

20. (Currently Amended) The method of claim 19, ~~wherein receiving speech content having the first accent comprises~~ further comprising:

receiving, in real time, continuous speech content having the first accent; and

~~wherein converting the synthesized audio data into a synthesized version of the received speech content having the second accent comprises~~ continuously converting the synthesized fourth audio data into a synthesized version of the ~~received~~ continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

21. (Currently Amended) The computing device of claim 1, wherein the speech content having the first accent further comprises a set of prosodic features, ~~wherein~~ the program instructions [[that]] are executable by the at least one processor such that the computing device is further configured to ~~synthesize the audio data representative of the received speech content having the second accent comprise program instructions that are executable by the at least one processor such that the computing device is configured to~~ synthesize the fourth audio data representative of the ~~received~~ speech content having the second accent and the set of prosodic features, and ~~wherein~~ the synthesized version of the ~~received~~ speech content having the second accent has the set of prosodic features.

22. (New) The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content.

REMARKS

Applicant has canceled claim 14, without prejudice, amended claims 1-13 and 15-21, and added new claim 22, as set forth above. No new matter has been added by way of these amendments. In view of the above amendments and the following remarks, reconsideration of the pending Office action is respectfully requested.

The Office has rejected, under 35 U.S.C. § 103, claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 as being obvious in view of U.S. Patent No. 10,163,451 to Dirac et al. ("Dirac"), U.S. Patent Application Publication No. 2014/0365216 to Gruber et al. ("Gruber"), and U.S. Patent Application Publication No. 2022/0180762 to Lurie et al. ("Lurie"); claims 3 and 12 as being obvious in view of Dirac, Gruber, and Lurie in further view of "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams" to Zhao et al. ("Zhao"); claims 5 and 14 as being obvious in view of Dirac, Gruber, and Lurie in further view of U.S. Patent No. 10,614,826 to Huffman et al. ("Huffman"); claims 7 and 16 as being obvious in view of Dirac, Gruber, and Lurie in further view of "Continuous Speech Recognition of Kannada Language Using Triphone Modeling" to Sajjan et al. ("Sajjan"); and claims 9, 18, and 20 as being obvious in view of Dirac, Gruber, and Lurie in further view of U.S. Patent Application Publication No. 2016/0203827 to Leff et al. ("Leff"). Applicant respectfully disagrees with these rejections, as explained in detail below.

Dirac, Gruber, Lurie, Zhao, Huffman, Sajjan, and Leff, alone or in combination, do not disclose or suggest, "mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content, wherein the first and second phonemes are different phonemes," as recited by independent claims 1, 10, and 19.

The Office has acknowledged on page 4 of the Office action that Dirac does not disclose a set of phonemes associated with a first pronunciation of speech content having a first accent or an updated set of phonemes associated with a second pronunciation of the speech content having a second accent. Dirac discloses in col. 10, lines 7-12 that, "it may be determined

IPR2026-00272
Krisp Technologies EX1002 Page 520

that speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to [speak] that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." (Emphasis added).

Thus, Dirac discloses adjusting audio characteristics for the same phonemes originating with speakers having different accents, and does not disclose or suggest multiple sets of phonemes, let alone associated with different pronunciations of speech content. Therefore, Dirac inherently cannot disclose mapping non-text linguistic representations of such different phonemes from multiple phoneme sets, as required by the above-identified claim limitations. Dirac does not disclose mapping between multiple sets of non-text linguistic representations of different phonemes because it is not concerned with, nor capable of compensating for, different pronunciations. In fact, "pronunciation" is not found anywhere in Dirac's disclosure.

The Office has asserted that Gruber addresses the deficiencies of Dirac in paragraphs [0123] and [0125]. However, Gruber also fails to disclose or suggest multiple sets of non-text linguistic representations of different phonemes that could possibly be mapped. Gruber discloses only text-based linguistic representations of phonemes. By way of example, Gruber discloses in paragraph [0013]:

> A first phonetic representation of the at least one word is determined, the first phonetic representation comprising a first set of phonemes selected from a speech recognition phonetic alphabet. The first set of phonemes is mapped to a second set of phonemes to generate a second phonetic representation, where the second set of phonemes are selected from a speech synthesis phonetic alphabet. (Emphasis added).

Gruber also discloses in paragraph [0125] that, "[f]or example, the first set of phonemes in the speech recognition alphabet 'f-ill-ee-p-ay' may be converted to the second set of phonemes 'fill-eep-ay.'"

Even assuming, *arguendo*, that Gruber teaches mapping different text-based phonemes between phoneme sets (i.e., a speech recognition phonetic alphabet and a speech synthesis phonetic alphabet), one of ordinary skill in the art in possession of Dirac would have no reason to look to Gruber because those references have opposing teachings. Specifically,

Dirac teaches adjusting audio characteristics of the same phonemes and, as alleged by the Office, Gruber maps text-based representations of different phonemes. In fact, to the extent Dirac teaches any phoneme mapping, Dirac expressly teaches away from the above-identified claim limitations that require mapping different phonemes.

Even if one of ordinary skill in the art in possession of Dirac would be motivated to look to Gruber, the combination would not provide any teaching as to how non-text linguistic representations of different phonemes could possibly be mapped. In particular, Gruber allegedly maps text-based linguistic representations of different phonemes and Dirac, at best, maps non-text linguistic representations of the same phoneme. Thus, the combination does not provide any teaching for one of ordinary skill in the art to map non-text linguistic representations of different phonemes. In other words, the text-based phoneme mapping of Gruber will not work on the non-text phonemes of Dirac, contrary to the Office's assertion on page 9 of the Office action, even ignoring that the phonemes in Dirac are not different.

Dirac and Gruber are simply silent with respect to different sets of non-text linguistic representations of different phonemes associated with different pronunciations of speech content, let alone mapping such different phonemes. Lurie merely discloses that some phonemes are harder to pronounce than others, which is irrelevant to the particular above-identified claim limitations. Zhao, Huffman, Sajjan, and Leff also do not address the deficiencies of Dirac and Gruber, nor disclose or suggest the above-identified claim limitations.

Dirac, Gruber, Lurie, Zhao, Huffman, Sajjan, and Leff, alone or in combination, also do not disclose or suggest, "derive, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content," as recited by independent claims 1 and 10 or "deriving, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content," as recited by independent claim 19.

The Office has cited col. 7, lines 52-60 of Dirac as disclosing the first machine-learning algorithm of the above-identified claim limitations. In col. 7, lines 52-60, Dirac discloses:

IPR2026-00272
Krisp Technologies EX1002 Page 522

> [T]he accent translation system 130 includes one or more artificial neural networks, such as a long short-term memory (LSTM) architecture or other systems that may efficiently parse and analyze large amounts of input data. In one embodiment, the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques. (Emphasis added).

As illustrated above, Dirac generally discloses that machine learning can be used to refine accent translation models.

However, in the above-identified claim limitations there is no accent translation associated with the first machine-learning algorithm, which instead is used to derive a non-text linguistic representation of a set of phonemes associated with a first pronunciation of speech content. In col. 5, lines 28-62, Dirac discloses methods for determining the audio characteristics the Office asserts teach the non-text linguistic representation of the set of phonemes of the above-identified claim limitations, including performing a fast Fourier transform and determining frequency and amplitude in relation to time, for example, none of which discloses or suggests a machine-learning algorithm, let alone trained with audio data comprising a particular accent.

Moreover, on page 4 of the Office action, the Office has cited the same disclosure of Dirac (i.e., col. 7, lines 52-60) as disclosing the second machine-learning algorithm of the independent claims, which is used to synthesize audio data representative of speech content having a second accent. The independent claims recite two different machine-learning algorithms that are each trained in a particular way to facilitate performance of a particular function. The Office has ignored the distinction between the first machine-learning algorithm of the above-identified claim limitations and the second machine-learning algorithm of the independent claims, and Dirac is silent with respect to the first machine-learning algorithm. Gruber, Lurie, Zhao, Huffman, Sajjan, and Leff also do not address the deficiencies of Dirac.

As explained in paragraph [0005] of the above-identified application, a voice conversion model that only adjusts audio characteristics of input speech (e.g., as disclosed by Dirac) does not address particular differences in pronunciation. The above-identified application also discloses in paragraph [0044] that, "because the VC engine 304 works at the

level of encoded linguistic data and does not need to predict and generate output speech as a midpoint for the conversion, it can function more quickly than alternatives," such as disclosed by Gruber, which generates text-based linguistic representations of phonemes as a midpoint for the conversion. Thus, the claimed technology advantageously facilitates accent conversion that accounts for pronunciation with extremely low latency, which is not possible based on the disclosures of the cited references.

In view of the foregoing amendments and remarks, the Office is respectfully requested to reconsider and withdraw the rejection of independent claims 1, 10, and 19. Since claims 2-9 and 21, 11-13 and 15-18, and 20 and 22 depend from and contain the limitations of claims 1, 10, and 19, respectively, they are distinguishable over the cited references and are patentable at least in the same manner as claims 1, 10, and 19.

Further, Dirac, Gruber, Lurie, Zhao, Huffman, Sajjan, and Leff, alone or in combination, do not disclose or suggest, "map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content," as recited by dependent claims 3 and 12 or "mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content," as recited by dependent claim 22.

The Office has cited col. 6, line 56 – col. 7, line 6 and col. 9, line 66 – col. 10, line 12 of Dirac as disclosing determining a respective monophone for each frame in received speech content. However, Dirac's disclosure in the portions cited by the Office that phonemes are spoken differently depending on the accent simply does not suggest any mapping of each frame in a non-text linguistic representation of a set of phonemes to a corresponding frame in a non-text linguistic representation of an updated set of phonemes. In fact, "frame" is not recited anywhere in Dirac's disclosure.

IPR2026-00272
Krisp Technologies EX1002 Page 524

Applicant understands that the Office's citation to paragraph [0024] of U.S. Patent Application Publication No. 2022/0122579 to Biadsy et al. ("Biadsy") and paragraph [0091] of U.S. Patent Application Publication No. 2022/0148562 to Park et al. ("Park") in the Advisory Action mailed March 20, 2023 is an acknowledgement of the deficiencies of Dirac (and the remainder of the prior art references of record) with respect to the above-identified claim limitations. However, Biadsy and Park, alone or in combination, also do not disclose or suggest the above-identified claim limitations.

In paragraph [0024], Biadsy discloses:

> The encoder 128 may be configured to convert the audio data 102 of the utterance 108 to an internal representation, such as a series of vectors. For example, as the encoder 128 receives the audio data 102 of the utterance 108, the encoder 128 may process five frames of audio and convert those five frames of audio to ten vectors….[T]he spectrogram decoder 130 may receive the ten vectors from the encoder 128 that represent the five frames of audio. The spectrogram decoder 130 generates five frames of audio data 106 of the synthesized utterance 114 that includes the same words or parts of words as the five frames of audio data, but with a different voice than the user 104. (Emphasis added).

Accordingly, Biadsy teaches converting frames of audio data to vectors, which are provided to a spectrogram decoder for synthesis into other frames of audio data with a different voice or accent.

However, the frames in Biadsy are in audio data generally, and not a non-text linguistic representation of a set of phonemes, as required by the above-identified claim limitations. Biadsy simply does not disclose or suggest any frames in non-text linguistic representations of sets of phonemes that are mapped. Additionally, the vectors in Biadsy are merely converted to other frames of audio data without any explicit mapping of frames, let alone frames in two different non-text linguistic representations of sets of phonemes associated with different pronunciations.

Park discloses in paragraph [0091], "a unique feature (for example, a tone, a pitch, an intensity, and a pronunciation) of a sound in a speech section (for example, a frame unit or a sentence unit)." Park discloses a pronunciation in a frame unit can be unique, but this

disclosure is irrelevant to the particular above-identified claim limitations relating to mapping frames of non-text linguistic representations of sets of phonemes. Park does not address the deficiencies of Biadsy and also does not disclose any mapping of frames in non-text linguistic representations of different sets of phonemes associated with different pronunciations of speech content.

As explained in paragraphs [0043]-[0044] of the above-identified application, for example, "the VC engine 304 may map each incoming frame in the derived linguistic representation to an outgoing target speech frame. In this way, the VC engine 304 functions to reconstruct acoustic features from the derived linguistic representation into audio data that is representative of speech by a different speaker having the second accent, all with very low latency." Thus, the claimed technology leverages a mapping of incoming frames in derived non-text linguistic representations to outgoing target speech frames to facilitate more accurate accent conversion with relatively low latency than possible based on the disclosures of the prior art of record. In view of the foregoing amendments and remarks, the Office is also respectfully requested to reconsider and withdraw the rejection of claims 3, 12, and 22.

In view of all of the foregoing, Applicant submits that this case is in condition for allowance and such allowance is earnestly solicited. The Office is encouraged to contact the undersigned to discuss any deficiencies the Office believes prevent allowability of the above-identified application.

Respectfully submitted,

Date: April 17, 2023

/Nicholas J. Gallo/
Nicholas J. Gallo
Registration No. 60,758

Troutman Pepper Hamilton Sanders LLP
70 Linden Oaks, Suite 210
Rochester, NY 14625
Telephone: (585) 270-2133
Facsimile: (585) 270-2179

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20060129399-A1	06-2006	Turk; Oytun	G10L21/00	704/E21.001
*	B	US-20090204395-A1	08-2009	Kato; Yumiko	G10L13/033	704/E11.001
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

IPR2026-00272
Krisp Technologies EX1002 Page 527

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022	08/04/2022	01/17/2023	07/11/2023				
	1	✓	✓	✓	✓	✓				
	2	✓	✓	✓	✓	✓				
	3	✓	✓	✓	✓	✓				
	4	✓	✓	✓	✓	✓				
	5	✓	✓	✓	✓	✓				
	6	✓	✓	✓	✓	✓				
	7	✓	✓	✓	✓	✓				
	8	✓	✓	✓	✓	✓				
	9	✓	✓	✓	✓	✓				
	10	✓	✓	✓	✓	✓				
	11	✓	✓	✓	✓	✓				
	12	✓	✓	✓	✓	✓				
	13	✓	✓	✓	✓	✓				
	14	✓	✓	✓	✓	-				
	15	✓	✓	✓	✓	✓				
	16	✓	✓	✓	✓	✓				
	17	✓	✓	✓	✓	✓				
	18	✓	✓	✓	✓	✓				
	19	✓	✓	✓	✓	✓				
	20	✓	✓	✓	✓	✓				
	21				✓	✓				
	22					✓				

IPR2026-00272
Krisp Technologies EX1002 Page 528

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	**Date**	**Examiner**
G10L13/02	07/11/2023	BB
G10L15/02	07/11/2023	BB

CPC Combination Sets - Searched*		
Symbol	**Date**	**Examiner**

US Classification - Searched*			
Class	**Subclass**	**Date**	**Examiner**

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	**Date**	**Examiner**
Performed inventor search in PE2E, EAST	07/11/2023	BB
Performed text search	07/11/2023	BB
Performed text limited CPC search	07/11/2023	BB
Updated search	07/11/2023	BB
Consulted with Primary Examiner Paras Shah	07/11/2023	BB

Interference Search			
US Class/CPC Symbol	**US Subclass/CPC Group**	**Date**	**Examiner**

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 529

Web Search History

date, time	web site	search string
7/5/2023 1:23:46 PM	Google Scholar	"(accent OR speech OR voice) conversion"
7/5/2023 1:23:55 PM	Google Scholar	"(accent OR speech OR voice) conversion" AND (phoneme OR phone)
7/5/2023 1:23:59 PM	Google Scholar	"(accent OR speech OR voice) conversion" AND (phoneme OR phone) [before:2021]
7/5/2023 1:56:13 PM	Google Scholar	[before:2021] "accent conversion" AND (phoneme OR phone)
7/5/2023 2:49:04 PM	Google Scholar	[before:2021] "accent conversion" AND (phoneme OR phone) AND (map OR mapping)
7/5/2023 2:49:19 PM	Google Scholar	"accent conversion" AND (phoneme OR phone) AND (map OR mapping) [before:2021]

IPR2026-00272
Krisp Technologies EX1002 Page 530

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L43	2886	(accent phoneme) near7 conver$4 AND (learn$3 train$3)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:07 PM
L44	589	(accent phoneme) near7 conver$4 SAME (pronounc$7 pronunc$7) AND (learn$3 train$3)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:07 PM
L45	765	(accent phoneme) near7 (conver$4 chang$3 alter$3 morph$3) SAME (pronounc$7 pronunc$7) AND (learn$3 train$3) AND (speech voice)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:08 PM
L46	660	phoneme near7 (conver$4 chang$3 alter$3 morph$3) SAME (pronounc$7 pronunc$7) AND (learn$3 train$3) AND (speech voice)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:08 PM
L47	350	phoneme near7 (conver$4 chang$3 alter$3 morph$3) near9 (pronounc$7 pronunc$7) AND (learn$3 train$3) AND (speech voice)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:08 PM
L48	253	phoneme near7 (conver$4 chang$3 alter$3 morph$3) near9 (pronounc$7 pronunc$7) AND (learn$3 train$3) AND (speech voice) near7 (conver$4 chang$3 alter$3 morph$3)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:11 PM
L49	166	phoneme near7 (conver$4 chang$3 alter$3 morph$3) near9 (pronounc$7 pronunc$7) AND (learn$3 train$3) near7 (algorithm model) AND (speech voice) near7 (conver$4 chang$3 alter$3 morph$3)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:11 PM
L50	50	phoneme near7 (conver$4 chang$3	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:22 PM

IPR2026-00272
Krisp Technologies EX1002 Page 531

		alter$3 morph$3) near9 (pronounc$7 pronunc$7) AND ((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab.					
L51	9085	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab.	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:23 PM
L52	1483	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND phoneme	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:23 PM
L53	424	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND phoneme SAME (pronounc$7 pronunc$7)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:24 PM
L54	45	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND phoneme SAME (pronounc$7 pronunc$7) SAME map$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 12:24 PM
L55	48	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND (phoneme phone monophone diphone triphone) SAME (pronounc$7 pronunc$7) SAME map$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 01:40 PM
L56	408	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).bsum. AND (phoneme phone monophone diphone triphone) SAME (pronounc$7 pronunc$7) SAME map$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 01:47 PM
L57	171	((speech voice speaker) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND (phoneme phone	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 01:47 PM

IPR2026-00272
Krisp Technologies EX1002 Page 532

		monophone diphone triphone) near9 map$4					
L58	16	((accent pronunciation pronounc$3) near7 (conver$4 chang$3 alter$3 morph$3)).ab. AND (phoneme phone monophone diphone triphone) near9 map$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 01:54 PM
L59	61	((accent pronunciation pronounc$3) near7 (conver$4 chang$3 alter$3 morph$3)) SAME (phoneme phone monophone diphone triphone) near9 map$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/05 02:11 PM
L60	833	accent SAME (pronounc$7 pronunc$7) SAME (phone phoneme triphone diphone monophone)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/10 01:50 PM
L61	184	accent SAME (pronounc$7 pronunc$7) SAME (phone phoneme triphone diphone monophone) AND (conver$4 morph$3 transform$7 chang$3 alter$5) SAME mapping	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/10 01:51 PM
L62	46	accent SAME (pronounc$7 pronunc$7) SAME (phone phoneme triphone diphone monophone) AND (conver$4 morph$3 transform$7 chang$3 alter$5) SAME mapping near9 (phone phoneme triphone diphone monophone)	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/10 01:51 PM
L63	20	62 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2023/07/11 12:24 PM
L64	3	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO,	OR	ON	ON	2023/07/11 12:24 PM

			NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)				

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 07/14/2023

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
07/14/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

BadumK@pepperlaw.com
USPTOinboxroc@pepperlaw.com

IPR2026-00272
Krisp Technologies EX1002 Page 535

Office Action Summary	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 04/17/2023.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-13 and 15-22 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-13 and 15-22 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Continued Examination Under 37 CFR 1.114

2. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on 04/17/2023 has been entered.

Response to Arguments

3. Applicant's arguments filed 01/04/2023 have been fully considered but they are not persuasive. Regarding arguments on page 11 of the Remarks, Examiner notes that Dirac is directed to accent translation. While Dirac does not explicitly teach the changing of phonemes, the newly applied Hwang reference teaches that different accents may have corresponding different phonemes. Therefore, when combined, the mapping and changing of one accent to another would also correspond to changing of the phonemes.

4. Regarding arguments on pages 12-13, Examiner notes that the accent translation system 130 itself is being updated and refined using machine learning. This system is used in col. 6 line 56 – col. 7 line 6 to determine the non-text linguistic representations for the various phonemes. The system is further used to synthesize the fourth audio data based on the trained and updated models. Therefore,

the same system, which is updated with machine learning, is used to perform both of the algorithms claimed. Examiner further notes that no details of the algorithms themselves are claimed other than the data they are trained with. Therefore, the machine learning used to update the language models would be used for both the derivation of the non-text linguistic representations as well as the synthesis steps.

5. Applicant's arguments with respect to claim(s) 1-13 and 15-22 have been considered but are moot because the new ground of rejection does not rely on any reference applied in the prior rejection of record for any teaching or matter specifically challenged in the argument.

Claim Rejections - 35 USC § 103

6. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

7. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

8. Claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac, in view of Hwang (US 2014/0258462 A1).

Regarding claim 1, Dirac teaches:

A computing device comprising:

at least one microphone (Fig. 6 element 621, col. 8 lines 16-32, where a microphone is used);

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used);

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used); and

program instructions stored on the non-transitory computer-readable medium that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) such that the computing device is configured to:

receive, via the at least one microphone, speech content having a first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content

to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized fourth audio data into a synthesized version of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the first and second phonemes are different phonemes;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Hwang teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an "ar" phoneme is replaced with an "ah" phoneme);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Regarding claim 2, Dirac in view of Hwang teaches:

The computing device of claim I, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to apply, to the derived non-text linguistic representation of the speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent) and a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52,

col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 8, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used) provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

receive, via at least one microphone, speech content having a first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

derive, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content, synthesize, using a second machine learning-algorithm

trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized fourth audio data into a synthesized version of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the first and second phonemes are different phonemes;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Hwang teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an "ar" phoneme is replaced with an "ah" phoneme);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Regarding claim 11, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent) and a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 17, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

receiving, via at least one microphone, speech content having a first accent (col. 8 lines 15-32, where accented speech is received at a microphone);

deriving, using a first machine-learning algorithm trained with first audio data comprising the first accent, a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio

characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content, synthesizing, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

converting the synthesized fourth audio data into a synthesized version of the speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;

wherein the first and second phonemes are different phonemes;

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

Hwang teaches:

wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an "ar" phoneme is replaced with an "ah" phoneme);

wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Regarding claim 21, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor such that the computing device is further configured to synthesize the fourth audio data representative of the speech content having the second accent and the set of prosodic features, and the synthesized version of the speech content having the second accent has the set of prosodic features (Dirac col. 12 lines 52-62, where at least one of the characteristics is adjusted, where the one adjusted characteristic is chosen to be pitch, while the stress or prosody remains unadjusted).

9. Claims 3, 12, and 22 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and further in view of Turk et al. (US 2006/0129399 A1), hereinafter referred to as Turk.

Regarding claim 3, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to

Dirac in view of Hwang does not teach:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content.

Turk teaches:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the alignment of Turk

(Turk para [0059]) when performing the mapping of Dirac in view of Hwang (Dirac col. 6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

Regarding claim 12, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to

Dirac in view of Hwang does not teach:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content.

Turk teaches:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the alignment of Turk (Turk para [0059]) when performing the mapping of Dirac in view of Hwang (Dirac col. 6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

Regarding claim 22, Dirac in view of Hwang teaches:

The method of claim 19, further comprising

Dirac in view of Hwang does not teach:

mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content.

Turk teaches:

mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the alignment of Turk (Turk para [0059]) when performing the mapping of Dirac in view of Hwang (Dirac col. 6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

10. Claim 5 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and further in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac in view of Hwang teaches:

The computing device of claim 1,

Dirac in view of Hwang does not teach:

wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Hwang (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

11. Claims 7 and 16 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and further in view of Sajjan et al. (Sajjan, S. C., & Vijaya, C. (2016, March). Continuous Speech Recognition of Kannada language using triphone modeling. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 451-455). IEEE.), hereinafter referred to as Sajjan.

Regarding claim 7, Dirac in view of Hwang teaches:

The computing device of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received) and the program instructions are executable by the at least one processor such that the computing device is further configured to determine, for each frame in the speech content, a respective (i) monophone (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent) sound detected in the frame based on the non-text learned linguistic representation for the first accent.

Dirac in view of Hwang does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

(ii) triphone (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used) sound detected in the frame

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the analysis of Sajjan

(Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Hwang (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

Regarding claim 16, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received) and the program instructions, when executed by the at least one processor, further cause the computing device to determine, for each frame in the speech content, a respective (i) monophone based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Dirac in view of Hwang does not teach:

(ii) triphone sound detected in the frame

Sajjan teaches:

ii) triphone sound detected in the frame (Page 453 section VB, Page 454 Section VC, where monophone and triphone HMMs are used)

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the analysis of Sajjan (Sajjan Page 453-454 Sections VB-C) on the inputs of Dirac in view of Hwang (Dirac col. 8 lines 15-32), in order to achieve an acceptable recognition rate (Sajjan page 451 section I second paragraph).

12. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and further in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac in view of Hwang teaches:

The computing device of claim 1,

Dirac in view of Hwang does not teach:

wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac in view of Hwang teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Hwang does not teach:

wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac in view of Hwang teaches:

The method of claim 19,

Dirac in view of Hwang does not teach:

further comprising:

receiving, in real time, continuous speech content having the first accent; and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

further comprising:

receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

13. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2009/0204395 A1 para [0092-93] teaches receiving pronunciation and prosody information and determining whether the phoneme is to be altered.

14. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

UNITED STATES PATENT AND TRADEMARK OFFICE

USPTO Automated Interview Request (AIR)

Aug 17 2023

This paper requesting to schedule and/or conduct an interview is appropriate because:

This submission is requested to be accepted as an authorization for this interview to communicate via the internet. Recognizing that Internet communications are not secure, I hereby authorize the USPTO to communicate with the undersigned concerning scheduling of the interview via video conference, instant messaging, or electronic mail, and to conduct the interview in accordance with office practice including video conferencing.

Name(s):
Nicholas Gallo

S-signature:
/Nicholas Gallo/

Registration Number:
60758

U.S. Application Number:
17460145

Confirmation Number:
3314

E-mail Address:
nicholas.gallo@troutman.com

Phone Number:
+1 5852702133

Proposed Time of Interview:
8-28-2023 1:00 PM ET

Alternative Proposed Time(s) of Interview:
8-29-2023 10:00 AM ET

Alternative Proposed Time(s) of Interview:
8-30-2023 2:00 PM ET

Prefered Interview Type:
Telephonic

I am the applicant or applicant's representative for this application.

Topic for Discussion:
I would like to discuss the pending prior art rejections along with some proposed claim amendments. I have prepared an interview agenda, which I would like to e-mail to the Examiner upon scheduling of the interview. Thank you.

uspto UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 09/01/2023

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
09/01/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

BadumK@pepperlaw.com
USPTOinboxroc@pepperlaw.com

IPR2026-00272
Krisp Technologies EX1002 Page 559

Applicant-Initiated Interview Summary	**Application No.** 17/460,145	**Applicant(s)** Serebryakov et al.		
	Examiner BRYAN S BLANKENAGEL	**Art Unit** 2658	**AIA (First Inventor to File) Status** Yes	**Page** 1 of 1

All Participants (applicant, applicants representative, PTO personnel)	**Title**	**Type**
BRYAN S BLANKENAGEL	Primary Examiner	Telephonic
NICHOLAS J. GALLO	Attorney of Record	

Date of Interview: 28 August 2023

Issues Discussed:

Proposed Amendment(s)

Applicant and Examiner discussed the proposed claims, and Examiner agreed that they would overcome the prior art rejection and would require further search and consideration.

☑ Attachment

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

Applicant is reminded that a complete written statement as to the substance of the interview must be made of record in the application file. It is the applicants responsibility to provide the written statement, unless the interview was initiated by the Examiner and the Examiner has indicated that a written summary will be provided. See MPEP 713.04
Please further see:
MPEP 713.04
Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews, paragraph (b)
37 CFR § 1.2 Business to be transacted in writing

Applicant recordation instructions: The formal written reply to the last Office action must include the substance of the interview. (See MPEP section 713.04). If a reply to the last Office action has already been filed, applicant is given a non-extendable period of the longer of one month or thirty days from this interview date, or the mailing date of this interview summary form, whichever is later, to file a statement of the substance of the interview.

Examiner recordation instructions: Examiners must summarize the substance of any interview of record. A complete and proper recordation of the substance of an interview should include the items listed in MPEP 713.04 for complete and proper recordation including the identification of the general thrust of each argument or issue discussed, a general indication of any other pertinent matters discussed regarding patentability and the general results or outcome of the interview, to include an indication as to whether or not agreement was reached on the issues raised.

IPR2026-00272
Krisp Technologies EX1002 Page 560

17/460,145 - Telephonic Interview Agenda - For Discussion Purposes Only

I. Technology Overview

A. This technology utilizes machine learning models to convert input speech in a first accent into a synthesized version of the input speech in a second accent in real-time. As a result of a frame-wise breakdown of captured speech across multiple speakers having an accent, this technology can develop a learned linguistic representation of speech having the accent that is not speaker-specific. With the learned linguistic representation, this technology is able to work at the level of encoded linguistic data, does not need to predict and generate output speech as a midpoint for the conversion, and therefore advantageously facilitates accent conversion that accounts for pronunciation with extremely low latency.

II. Proposed Claim Amendments

1. (Currently Amended) A ~~computing device~~ system, comprising:
~~at least one microphone;~~
at least one processor; and
non-transitory computer-readable medium[[; and]] comprising program instructions stored thereon ~~the non-transitory computer-readable medium~~ that are executable by the at least one processor to cause ~~such that~~ the ~~computing device is configured~~ system to:
train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;
~~receive, via the at least one microphone, speech content having a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;~~
apply the ~~derive, using a~~ first machine-learning algorithm ~~trained with first audio data comprising the first accent,~~ to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;
based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the speech content, wherein the first and second phonemes are different phonemes; and
convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content.

IPR2026-00272
Krisp Technologies EX1002 Page 561

2. (New) The system of claim 1, wherein the training further comprises aligning and classifying each of a plurality of frames of the captured speech content according to monophone and triphone sounds of the captured speech content.

III. Prior Art Rejections - 10,163,451 to Dirac and 2014/0258462 to Hwang

A. The "application" limitation of claim 1

i. The Office cited Dirac for using the first machine-learning algorithm to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the speech content. Specifically, Dirac discloses that "the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques."

a. On page 5 of the Office action, the Office acknowledges that Dirac does not disclose speech content having the first accent that comprises a set of phonemes associated with a first pronunciation of the speech content. Since Dirac does not disclose that the speech content has a set of phonemes associated with a first pronunciation, Dirac inherently cannot derive a non-text linguistic representation of such a set of phonemes.

b. The Office asserts that the accent audio characteristics of Dirac disclose the linguistic representations of the claims, but not that the accent audio characteristics are derived by applying the machine learning employed to update and refine the accent translation models, as required by the above-identified claim limitations with respect to the linguistic representation.

1. Dirac does not disclose that the accent translation models themselves are machine learning models or that the accent translation models are applied to generate the accent audio characteristics, and instead teaches performing a fast Fourier transform and determining frequency and amplitude in relation to time to derive the accent audio characteristics.

ii. Even assuming that Hwang discloses that an accent is associated with specific phonemes of different pronunciation, this teaching does not address the deficiencies of Dirac nor disclose the claimed derivation. Specifically, Hwang also does not disclose applying a machine-learning algorithm to such speech content having a set of phonemes, let alone to derive a non-text linguistic representation of the set of phonemes.

B. The "training" limitation of claim 1

i. Dirac does not disclose that the machine learning employed to update and refine the accent translation models is trained using speech content captured from different speakers having a first accent or that such training includes aligning and classifying frames of captured speech content corresponding to respective ones of the different speakers.

C. Claim 2

i. Dirac does not disclose that any aligning and classifying of frames of the captured speech content is according to monophone and triphone sounds of the captured speech

IPR2026-00272
Krisp Technologies EX1002 Page 562

content. In fact, the Office has acknowledged that none of the cited prior art references discloses monophone and triphone sounds except Sajjan, which merely discloses that a triphone sound can be detected and says nothing about using such detection to align and classify frames to facilitate training of a machine learning model.

IPR2026-00272
Krisp Technologies EX1002 Page 563

PTO/SB/06 (09-11)
Approved for use through 1/31/2014. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/460,145	Filing Date 08/27/2021	☐ To be Mailed

ENTITY: ☐ LARGE ☑ SMALL ☐ MICRO

APPLICATION AS FILED - PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A		N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A		N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*		x $50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*		x $240 =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).				
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))					
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL	

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT	09/06/2023	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	* 22	Minus	** 21	= 1		x $40 =	40
	Independent (37 CFR 1.16(h))	* 3	Minus	*** 3	= 0		x $192 =	0
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	40

		(Column 1)		(Column 2)	(Column 3)			
AMENDMENT		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA		RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=		x $0 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=		x $0 =	
	☐ Application Size Fee (37 CFR 1.16(s))							
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							
							TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.	LIE
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".	/VICTORIA JOHNSON/
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".	
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.	

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

PATENT
Docket No. **264248.000100**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventors	:	Serebryakov and Zhang	Examiner: Bryan S. Blankenagel
Applicant	:	Sanas.ai Inc.	
Serial No.	:	17/460,145	
Cnfrm. No.	:	3314	Art Unit: 2658
Filed	:	April 27, 2021	
For	:	REAL-TIME ACCENT CONVERSION MODEL	

AMENDMENT IN RESPONSE TO NON-FINAL OFFICE ACTION

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Commissioner:

In response to the July 14, 2023, non-final Office action, please amend the above-identified patent application as follows:

Amendments to the Claims are reflected in the listing of claims which begins on page 2 of this paper.

Remarks begin on page 10 of this paper.

IPR2026-00272
Krisp Technologies EX1002 Page 565

Amendments to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

1. (Currently Amended) A ~~computing device~~ system, comprising:

~~at least one microphone;~~

at least one processor; and

non-transitory computer-readable medium[[; and]] comprising program instructions stored thereon ~~the non-transitory computer-readable medium~~ that are executable by the at least one processor to cause ~~such that~~ the ~~computing device is configured~~ system to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

~~receive, via the at least one microphone, speech content having a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;~~

apply the ~~derive, using a~~ first machine-learning algorithm ~~trained with first audio data comprising the first accent,~~ to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different

IPR2026-00272
Krisp Technologies EX1002 Page 566

from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

2. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

3. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

4. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

7. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions are executable by the at least one processor such that the computing device is further configured to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames ~~frame~~ in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to~~:~~

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

~~receive, via at least one microphone, speech content having a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;~~

apply the ~~derive, using a~~ first machine-learning algorithm ~~trained with first audio data comprising the first accent,~~ to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

11. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the received

speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

12. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

13. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.

14. (Canceled)

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

16. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions, when executed by the at least one processor, further cause the computing device to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames ~~frame~~ in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

17. (Currently Amended) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to:

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

19. (Currently Amended) A method ~~carried out by a computing device~~, ~~the method~~ comprising:

training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

~~receiving, via at least one microphone, speech content having a first accent, wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content;~~

applying the ~~deriving, using a~~ first machine-learning algorithm ~~trained with first audio data comprising the first accent,~~ to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

IPR2026-00272
Krisp Technologies EX1002 Page 571

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

20. (Previously Presented) The method of claim 19, further comprising:

receiving, in real time, continuous speech content having the first accent; and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

21. (Currently Amended) The computing device of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor such that the computing device is further configured to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the synthesized version of the received speech content having the second accent has the set of prosodic features.

22. (Currently Amended) The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic

representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

23. (New) The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions and the method further comprises:

aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

REMARKS

Applicant has amended claims 1-12, 16-17, 19, and 21-22 and added new claim 23, as set forth above. No new matter has been added by way of these amendments. Applicant appreciates the courtesy provided by the Examiner of the interview on August 28, 2023 ("Interview"), a summary of which is provided herein. In view of the above amendments and the following remarks, reconsideration of the pending Office action is respectfully requested.

The Office has rejected, under 35 U.S.C. § 103, claims 1-2, 4, 6, 8, 10-11, 13, 15, 17, 19, and 21 as being obvious in view of U.S. Patent No. 10,163,451 to Dirac et al. ("Dirac") in view of U.S. Patent Application Publication No. 2014/0258462 to Hwang ("Hwang"); claims 3, 12, and 22 as being obvious in view of Dirac and Hwang in further view of U.S. Patent Application Publication No. 2006/0129399 to Turk et al. ("Turk"); claim 5 as being obvious in view of Dirac, and Hwang in further view of U.S. Patent No. 10,614,826 to Huffman et al. ("Huffman"); claims 7 and 16 as being obvious in view of Dirac and Hwang in further view of "Continuous Speech Recognition of Kannada Language Using Triphone Modeling" to Sajjan et al. ("Sajjan"); and claims 9, 18, and 20 as being obvious in view of Dirac and Hwang in further view of U.S. Patent Application Publication No. 2016/0203827 to Leff et al. ("Leff"). Applicant respectfully disagrees with these rejections, as explained in detail below.

Dirac, Hwang, Turk, Huffman, Sajjan, and Leff, alone or in combination, do not disclose or suggest applying "the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content," as recited by independent claims 1, 10, and 19.

The Office cited Dirac for using a first machine-learning algorithm to derive a non-text linguistic representation of a set of phonemes associated with a first pronunciation of speech content. Specifically, Dirac discloses in col. 7, lines 56-60 that "the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques."

IPR2026-00272
Krisp Technologies EX1002 Page 574

However, on page 5 of the Office action, the Office acknowledges that Dirac does not disclose speech content having a first accent that comprises a set of phonemes associated with a first pronunciation of speech content. Since Dirac does not disclose that the speech content has a set of phonemes associated with a first pronunciation, Dirac inherently cannot derive a non-text linguistic representation of such a set of phonemes.

Further, the Office also asserts on page 5 of the Office action that the accent audio characteristics of Dirac disclose the linguistic representations of the above-identified claim limitations, but not that the accent audio characteristics are derived by applying the machine learning that Dirac discloses is employed to update and refine the accent translation models, as required by the above-identified claim limitations with respect to the linguistic representation. In fact, Dirac does not disclose that the accent translation models themselves are machine learning models or that the accent translation models are applied to generate the accent audio characteristics, and instead teaches in col. 5, lines 28-62, for example, performing a fast Fourier transform and determining frequency and amplitude in relation to time to derive the accent audio characteristics.

Even assuming, *arguendo*, that Hwang discloses that an accent is associated with specific phonemes of different pronunciation, as asserted by the Office, this teaching does not address the deficiencies of Dirac nor disclose the derivation of the above-identified claim limitations. Specifically, Hwang also does not disclose applying a machine-learning algorithm to such speech content having a set of phonemes, let alone to derive a non-text linguistic representation of the set of phonemes. Turk, Huffman, Sajjan, and Leff also do not disclose or suggest the above-identified claim limitations.

Dirac, Hwang, Turk, Huffman, Sajjan, and Leff, alone or in combination, also do not disclose or suggest training "a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers," as recited by independent claims 1, 10, and 19.

The Office has asserted that Dirac discloses using a first machine-learning model to derive a non-text linguistic representation of a set of phonemes associated with a pronunciation of speech content in col. 5, lines 9-27, col. 6, line 56 – col. 7, line 6, and col. 7, lines 52-60. Applicant has carefully reviewed Dirac and is only able to identify two portions that reference machine learning, which include col. 7, lines 56-60 reproduced and discussed above and col. 2, lines 45-49, which discloses that "the accent translation system may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other algorithm refinement techniques."

However, Dirac does not disclose that the machine learning employed to update and refine the accent translation models is trained using speech content captured from different speakers having a first accent or that such training includes aligning and classifying frames of captured speech content corresponding to respective ones of the different speakers. In fact, Dirac is completely silent as to how the disclosed machine learning is trained. Hwang, Turk, Huffman, Sajjan, and Leff also do not disclose or suggest the above-identified claim limitations. Applicant appreciates the Office's indication in the Interview that the amendments reflected above overcome the pending prior art rejections of the independent claims.

This technology utilizes machine learning models to convert input speech in a first accent into a synthesized version of the input speech in a second accent in real-time. As a result of a frame-wise breakdown of captured speech across multiple speakers having an accent, this technology can develop a learned linguistic representation of speech having the accent that is not speaker-specific. With the learned linguistic representation, this technology is able to work at the level of encoded linguistic data, does not need to predict and generate output speech as a midpoint for the conversion, and therefore advantageously facilitates accent conversion that accounts for pronunciation with extremely low latency that is not possible based on the cited prior art.

In view of the foregoing amendments and remarks, the Office is respectfully requested to reconsider and withdraw the rejection of independent claims 1, 10, and 19. Since claims 2-9 and 21, 11-13 and 15-18, and 20 and 22-23 depend from and contain the limitations of claims 1, 10, and 19, respectively, they are distinguishable over the cited references and are patentable at least in the same manner as claims 1, 10, and 19.

In view of all of the foregoing, Applicant submits that this case is in condition for allowance and such allowance is earnestly solicited. The Office is encouraged to contact the undersigned to discuss any deficiencies the Office believes prevent allowability of the above-identified application.

Respectfully submitted,

Date: September 6, 2023

/Nicholas J. Gallo/
Nicholas J. Gallo
Registration No. 60,758

Troutman Pepper Hamilton Sanders LLP
70 Linden Oaks, Suite 210
Rochester, NY 14625
Telephone: (585) 270-2133
Facsimile: (800) 541-9525

Electronic Acknowledgement Receipt

EFS ID:	48542849
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	11951
Filer:	Nicholas Joseph Gallo
Filer Authorized By:	
Attorney Docket Number:	264248.000100
Receipt Date:	06-SEP-2023
Filing Date:	27-AUG-2021
Time Stamp:	10:51:02
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		20230906_Amendment_Response_Non_Final_OA_264248_000100.pdf	179435 080073eb83f0e3e1e6018a1341305ba0058f110b	yes	13

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment/Request for Reconsideration-After Non-Final Rejection	1	1
	Claims	2	9
	Applicant Arguments/Remarks Made in an Amendment	10	13
Warnings:			
Information:			
	Total Files Size (in bytes):	179435	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

United States Patent and Trademark Office

Office of the Chief Financial Officer

Document Code:WFEE

User :Victoria Johnson

Sale Accounting Date:09/11/2023

Sale Item Reference Number	Effective Date
17460145	09/06/2023

Document Number	Fee Code	Fee Code Description	Amount Paid	Payment Method
I202399B46545400	1202	CLAIMS IN EXCESS OF 20	$100.00	Deposit Account

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	CPC Classification	US Classification
*	A	US-20220301542-A1	09-2022	SUNG; Junesig	G10L13/02	1/1
*	B	US-20150170642-A1	06-2015	Peng; Fuchun	G10L15/187	704/235
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

IPR2026-00272
Krisp Technologies EX1002 Page 581

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022	08/04/2022	01/17/2023	07/11/2023	09/12/2023			
	1	✓	✓	✓	✓	✓	✓			
	2	✓	✓	✓	✓	✓	✓			
	3	✓	✓	✓	✓	✓	✓			
	4	✓	✓	✓	✓	✓	✓			
	5	✓	✓	✓	✓	✓	✓			
	6	✓	✓	✓	✓	✓	✓			
	7	✓	✓	✓	✓	✓	✓			
	8	✓	✓	✓	✓	✓	✓			
	9	✓	✓	✓	✓	✓	✓			
	10	✓	✓	✓	✓	✓	✓			
	11	✓	✓	✓	✓	✓	✓			
	12	✓	✓	✓	✓	✓	✓			
	13	✓	✓	✓	✓	✓	✓			
	14	✓	✓	✓	✓	-	-			
	15	✓	✓	✓	✓	✓	✓			
	16	✓	✓	✓	✓	✓	✓			
	17	✓	✓	✓	✓	✓	✓			
	18	✓	✓	✓	✓	✓	✓			
	19	✓	✓	✓	✓	✓	✓			
	20	✓	✓	✓	✓	✓	✓			
	21				✓	✓	✓			
	22					✓	✓			
	23						✓			

IPR2026-00272
Krisp Technologies EX1002 Page 582

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	**Date**	**Examiner**
G10L13/02	09/12/2023	BB
G10L15/02	09/12/2023	BB

CPC Combination Sets - Searched*		
Symbol	**Date**	**Examiner**

US Classification - Searched*			
Class	**Subclass**	**Date**	**Examiner**

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	**Date**	**Examiner**
Performed inventor search in PE2E, EAST	09/12/2023	BB
Performed text search	09/12/2023	BB
Performed text limited CPC search	09/12/2023	BB
Updated search	09/12/2023	BB

Interference Search			
US Class/CPC Symbol	**US Subclass/CPC Group**	**Date**	**Examiner**

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 583

Web Search History

date, time	web site	search string
9/12/2023 2:23:17 PM	Google Scholar	accent AND (training OR learning) AND (model OR algorithm)
9/12/2023 2:23:23 PM	Google Scholar	accent AND (training OR learning) AND (model OR algorithm) [before:2021]
9/12/2023 2:23:33 PM	Google Scholar	[before:2021] accent AND (training OR learning) AND (model OR algorithm) AND align
9/12/2023 2:23:40 PM	Google Scholar	[before:2021] accent AND (training OR learning) AND (model OR algorithm) AND align AND (classify OR label)
9/12/2023 2:23:50 PM	Google Scholar	[before:2021] accent AND (training OR learning) AND (model OR algorithm) AND align AND (classify OR label) AND (pronounce OR pronunciation)
9/12/2023 2:23:58 PM	Google Scholar	[before:2021] accent AND (training OR learning) AND (model OR algorithm) AND align AND (classify OR label) AND (pronounce OR pronunciation) AND (phoneme OR triphone)
9/12/2023 2:24:06 PM	Google Scholar	[before:2021] accent AND (training OR learning) AND (model OR algorithm) AND align AND (classify OR label) AND (pronounce OR pronunciation) AND (phoneme OR triphone) AND synthesize

IPR2026-00272
Krisp Technologies EX1002 Page 584

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L65	1224	(train$3 learn$3) near7 (algorithm model) SAME (speech utterance voice speak$3) SAME accent	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:23 PM
L66	4	(train$3 learn$3) near7 (algorithm model) SAME (speech utterance voice speak$3) SAME accent SAME align$4 SAME classif$7	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:24 PM
L67	24	(train$3 learn$3) near7 (algorithm model) SAME (speech utterance voice speak$3) SAME accent AND (train$3 learn$3) SAME align$4 near9 frame	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:25 PM
L68	17	(train$3 learn$3) near7 (algorithm model) SAME (speech utterance voice speak$3) SAME accent SAME align$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:29 PM
L69	18	(train$3 learn$3) near7 (algorithm model) SAME accent SAME align$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:41 PM
L70	66	(train$3 learn$3) SAME accent SAME align$4	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 12:41 PM
L71	15	70 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2023/09/12 02:20 PM
L72	3	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO; JPO; DERWENT; IBM_TDB)	OR	ON	ON	2023/09/12 02:20 PM

PE2E SEARCH - Search History (Interference)

There are no Interference searches to show.

IPR2026-00272
Krisp Technologies EX1002 Page 586

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 09/18/2023

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
09/18/2023	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

BadumK@pepperlaw.com
USPTOinboxroc@pepperlaw.com

IPR2026-00272
Krisp Technologies EX1002 Page 587

Office Action Summary	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 09/06/2023.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☑ This action is **FINAL.** 2b) ☐ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-13 and 15-23 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☑ Claim(s) 1-13 and 15-23 is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☐ Claim(s) _____ are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☑ The drawing(s) filed on 08/27/2021 is/are: a)☑ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☑ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Response to Arguments

2. Applicant's arguments with respect to claim(s) 1-13 and 15-23 have been considered but are moot because the new ground of rejection does not rely on any reference applied in the prior rejection of record for any teaching or matter specifically challenged in the argument.

3. Applicant's arguments filed 09/06/2023 have been fully considered but they are not persuasive. Regarding arguments on pages 10-11 of the Remarks, Examiner notes that Dirac is not being viewed alone in the rejection, but in combination with other references. The combination of references is relied upon to teach the claimed limitations.

Claim Rejections - 35 USC § 103

4. In the event the determination of the status of the application as subject to AIA 35 U.S.C. 102 and 103 (or as subject to pre-AIA 35 U.S.C. 102 and 103) is incorrect, any correction of the statutory basis for the rejection will not be considered a new ground of rejection if the prior art relied upon, and the rationale supporting the rejection, would be the same under either status.

5. The following is a quotation of 35 U.S.C. 103 which forms the basis for all obviousness rejections set forth in this Office action:

> A patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the

claimed invention pertains. Patentability shall not be negated by the manner in which the invention was made.

6. Claims 1-2, 4, 6-8, 10-11, 13, 15-17, 19, 21, and 23 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac et al. (US 10,163,451 B2), hereinafter referred to as Dirac, in view of Hwang (US 2014/0258462 A1), and further in view of Peng et al. (US 2015/0170642 A1), hereinafter referred to as Peng.

Regarding claim 1, Dirac teaches:

A system, comprising:

at least one processor (Fig. 11 elements 10a, 10b, col. 16 lines 17-29, where a processor is used); and

non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used) comprising program instructions stored thereon that are executable by the at least one processor (col. 16 lines 17-40, where instructions are stored and executed) to cause the system to:

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content (col. 8 lines 15-32, where accented speech is received at a microphone), to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data

comprising the second accent, fourth audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

comprising a set of phonemes associated with a first pronunciation of the received speech content;

wherein the first and second phonemes are different phonemes;

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

Hwang teaches:

comprising a set of phonemes associated with a first pronunciation of the received speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an "ar" phoneme is replaced with an "ah" phoneme);

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Peng teaches:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent (para [0018], where the system is trained to associate pronunciations, and para [0031], where the accent is from a group of users), wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers (para [0025-26], where alignment of frames is performed, and para [0030], where the accent group identifier is determined);

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the training of Peng (Peng para [0018]) on the system of Dirac in view of Hwang (Dirac col. 5 lines 9-27) to associate actual pronunciations to an expected pronunciation and perform replacements or substitutions before processing an utterance (Peng para [0003]).

Regarding claim 2, Dirac in view of Hwang and Peng teaches:

The system of claim I, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 4, Dirac in view of Hwang and Peng teaches:

The system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 6, Dirac in view of Hwang and Peng teaches:

The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent) and a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines

15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 7, Dirac in view of Hwang and Peng teaches:

The system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received) and the program instructions are executable by the at least one processor such that the computing device is further configured to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm (Peng para [0027], where a sequence of three phonemes as well as single phonemes are used in the training, which includes alignment and classifying as in para [0025-26] and [0030]); and

determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound (Peng para [0027], where a sequence of three phonemes is used) detected in the frame based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Regarding claim 8, Dirac in view of Hwang and Peng teaches:

The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 10, Dirac teaches:

A non-transitory computer-readable medium (col. 16 lines 17-29, where a computer-accessible medium is used) provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content (col. 8 lines 15-32, where accented speech is received at a microphone), to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content (col. 6 lines

27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

comprising a set of phonemes associated with a first pronunciation of the received speech content;

wherein the first and second phonemes are different phonemes;

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

Hwang teaches:

comprising a set of phonemes associated with a first pronunciation of the received speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an “ar” phoneme is replaced with an “ah” phoneme);

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received

speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Peng teaches:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent (para [0018], where the system is trained to associate pronunciations, and para [0031], where the accent is from a group of users), wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers (para [0025-26], where alignment of frames is performed, and para [0030], where the accent group identifier is determined);

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the training of Peng (Peng para [0018]) on the system of Dirac in view of Hwang (Dirac col. 5 lines 9-27) to associate actual pronunciations to an expected pronunciation and perform replacements or substitutions before processing an utterance (Peng para [0003]).

Regarding claim 11, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the

derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent (Dirac col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Regarding claim 13, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent (Dirac col. 3 lines 36-60, where the audio samples are obtained from various individuals).

Regarding claim 15, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent (Dirac col. 8 lines 33-52, where a first user inputs indication of their accent) and a second user input indicating a selection of the second accent (Dirac col. 8 lines 33-52, col. 9 lines 15-38, where a first user inputs indication of their accent, and the first or second speaker inputs indication of the second speaker's accent).

Regarding claim 16, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein the first machine- learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received) and

the program instructions, when executed by the at least one processor, further cause the computing device to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm (Peng para [0027], where a sequence of three phonemes as well as single phonemes are used in the training, which includes alignment and classifying as in para [0025-26] and [0030]); and

determine, for each of a plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound (Peng para [0027], where a sequence of three phonemes is used) detected in the frame based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

Regarding claim 17, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device (Dirac Fig. 6 element 632, 612, where the output audio with the second accent is transmitted to the second device).

Regarding claim 19, Dirac teaches:

A method carried out by a computing device, the method comprising:

applying the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content (col. 8 lines 15-32, where accented speech is received at a microphone), to derive a non-

text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, col. 6 line 56 – col. 7 line 6, where the audio characteristics are determined for various phonemes of the accents, and col. 7 lines 52-60, where machine learning is used to update when new samples are received);

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent, and col. 7 lines 52-60, where the models use machine learning), wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content (col. 6 lines 27-55, where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings); and

converting the synthesized fourth audio data into a synthesized version of the received speech content having the second accent (col. 12 line 52 - col. 13 line 18, where the accent translation model adjusts audio characteristics from a first accent to a second accent).

Dirac does not teach:

training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises

aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

comprising a set of phonemes associated with a first pronunciation of the received speech content;

wherein the first and second phonemes are different phonemes;

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

Hwang teaches:

comprising a set of phonemes associated with a first pronunciation of the received speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Boston accent);

wherein the first and second phonemes are different phonemes (para [0087], [0089], where an "ar" phoneme is replaced with an "ah" phoneme);

wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content (para [0087], [0089], where an accent is associated with specific phonemes with different pronunciations, such as a Southern accent).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac by using the phonemes of Hwang (Hwang para [0086-90]) when considering the accents of Dirac (Dirac col. 5 lines 9-27), so that changing the accent would include changing phonemes where needed, so that the system may respond to the user request for customization of the content (Hwang para [0089]).

Peng teaches:

training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent (para [0018], where the system is trained to associate pronunciations, and para [0031], where the accent is from a group of users), wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers (para [0025-26], where alignment of frames is performed, and para [0030], where the accent group identifier is determined);

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang by using the training of Peng (Peng para [0018]) on the system of Dirac in view of Hwang (Dirac col. 5 lines 9-27) to associate actual pronunciations to an expected pronunciation and perform replacements or substitutions before processing an utterance (Peng para [0003]).

Regarding claim 21, Dirac in view of Hwang and Peng teaches:

The computing device of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor such that the computing device is further configured to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the synthesized version of the received speech content having the second accent has the set of prosodic features (Dirac col. 12 lines 52-62, where at least one of the characteristics is adjusted, where the one adjusted characteristic is chosen to be pitch, while the stress or prosody remains unadjusted).

Regarding claim 23, Dirac in view of Hwang and Peng teaches:

The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent (Dirac col. 5 lines 9-27, where the accent audio characteristics of the accent sample sets are the linguistic representations, and col. 7 lines 52-60, where machine learning is used to update when new samples are received) and the program instructions and the method further comprises:

aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm (Peng para [0027], where a sequence of three phonemes as well as single phonemes are used in the training, which includes alignment and classifying as in para [0025-26] and [0030]); and

determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone sound and (ii) triphone sound (Peng para [0027], where a sequence of three phonemes is used) detected in the frame based on the non-text learned linguistic representation for the first accent (Dirac col. 6 line 56 - col. 7 line 6, col. 9 line 66 - col. 10 line 12, where phonemes are spoken differently depending on the accent).

7. Claims 3, 12, and 22 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and Peng, and further in view of Turk et al. (US 2006/0129399 A1), hereinafter referred to as Turk.

Regarding claim 3, Dirac in view of Hwang and Peng teaches:

The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to

Dirac in view of Hwang and Peng does not teach:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

Turk teaches:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by using the alignment of Turk (Turk para [0059]) when performing the mapping of Dirac in view of Hwang and Peng (Dirac col. 6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

Regarding claim 12, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to

Dirac in view of Hwang and Peng does not teach:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

Turk teaches:

map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by using the alignment of Turk (Turk para [0059]) when performing the mapping of Dirac in view of Hwang and Peng (Dirac col. 6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

Regarding claim 22, Dirac in view of Hwang and Peng teaches:

The method of claim 19, further comprising

Dirac in view of Hwang and Peng does not teach:

mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

Turk teaches:

mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content (para [0059], where the two utterances are aligned so that phonemes may be mapped to each other, where the alignment is considered the mapping of one frame to another).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by using the alignment of Turk (Turk para [0059]) when performing the mapping of Dirac in view of Hwang and Peng (Dirac col.

6 lines 27-55), in order to ensure accuracy when determining if phonemes of different accents should be mapped to each other (Turk para [0059]).

8. Claim 5 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and Peng, and further in view of Huffman et al. (US 10,614,826 B2), hereinafter referred to as Huffman.

Regarding claim 5, Dirac in view of Hwang and Peng teaches:

The system of claim 1,

Dirac in view of Hwang and Peng does not teach:

wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

Huffman teaches:

wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent (col. 4 lines 14-30 and 45-56, where voice, including accent, is transformed to that of a target speaker).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by using the voice conversion method of Huffman (Huffman col. 4 lines 14-30) in the accent conversion of Dirac in view of Hwang and Peng (Dirac col. 12 line 52 - col. 13 line 18) to allow real-time transformation of a source voice to optimally mimic a target voice (Huffman col. 4 lines 14-30).

9. Claims 9, 18, and 20 is/are rejected under 35 U.S.C. 103 as being unpatentable over Dirac, in view of Hwang, and Peng, and further in view of Leff et al. (US 2016/0203827 A1), hereinafter referred to as Leff.

Regarding claim 9, Dirac in view of Hwang and Peng teaches:

The system of claim 1,

Dirac in view of Hwang and Peng does not teach:

wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang and Peng (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 18, Dirac in view of Hwang and Peng teaches:

The non-transitory computer-readable medium of claim 10,

Dirac in view of Hwang and Peng does not teach:

wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang and Peng (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Regarding claim 20, Dirac in view of Hwang and Peng teaches:

The method of claim 19,

Dirac in view of Hwang and Peng does not teach:

further comprising:

receiving, in real time, continuous speech content having the first accent; and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

Leff teaches:

further comprising:

receiving, in real time, continuous speech content having the first accent (para [0093], where real-time conversion is performed); and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent (para [0093], where real-time conversion is performed, with a delay of 50ms, or in a range from 20-250 ms).

It would have been obvious to one of ordinary skill in the art before the effective filing date of the claimed invention to modify the system of Dirac in view of Hwang and Peng by including the delay of Leff (Leff para [0093]) in the conversion of Dirac in view of Hwang and Peng (Dirac col. 12 line 52 - col. 13 line 18), so that a live conversion can be performed with the delay having no appreciable effect on the flow of conversation between two individuals (Leff para [0093]).

Conclusion

10. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2022/0301542 A1 para [0101] teaches alignment of frames as well as training a speech model to approach an accent of a target speaker.

11. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

12. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional

questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

PTO/AIA/31 (12-22)
Approved for use through 05/31/2024. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

NOTICE OF APPEAL FROM THE EXAMINER TO THE PATENT TRIAL AND APPEAL BOARD

Docket Number (Optional): 264248.000100

I hereby certify that this correspondence is being facsimile transmitted to the USPTO, transmitted via the USPTO's patent electronic filing system, or deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to "Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450" [37 CFR 1.8(a)] on _____

Signature _____

Typed or printed name _____

First Named Inventor	
Maxim Serebryakov	
Application Number: 17/460,145	Filed: 08-27-2021
For: Real-Time Accent Conversion Model	
Art Unit: 2658	Examiner: BLANKENAGEL, BRYAN S

Applicant hereby **appeals** to the Patent Trial and Appeal Board from the last decision of the examiner.

The fee for this Notice of Appeal is (37 CFR 41.20(b)(1)) $ 840.00

☑ Applicant asserts small entity status. See 37 CFR 1.27. Therefore, the fee shown above is reduced by 60%, and the resulting fee is: $ 336.00

☐ Applicant certifies micro entity status. See 37 CFR 1.29. Therefore, the fee shown above is reduced by 80%, and the resulting fee is: $ _____
Form PTO/SB/15A or B or equivalent must either be enclosed or have been submitted previously.

☐ A check in the amount of the fee is enclosed.

☐ Payment by credit card. Form PTO-2038 is attached.

☑ The Director is hereby authorized to charge any fees which may be required, or credit any overpayment to Deposit Account No. 500436.

☐ Payment made via USPTO's patent electronic filing system (Patent Center or EFS-Web).

☐ A petition for an extension of time under 37 CFR 1.136(a) (PTO/AIA/22 or equivalent) is enclosed.
For extensions of time in reexamination proceedings, see 37 CFR 1.550.

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

I am the

☐ applicant

☑ attorney or agent of record Registration number 60,758

☐ attorney or agent acting under 37 CFR 1.34 Registration number _____

Signature /Nicholas J. Gallo/

Typed or printed name Nicholas J. Gallo

Telephone Number 585-270-2133

Date September 28, 2023

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications. Submit multiple forms if more than one signature is required, see below*.

☑ * Total of 1 forms are submitted.

A Federal agency may not conduct or sponsor, and a person is not required to respond to, nor shall a person be subject to a penalty for failure to comply with an information collection subject to the requirements of the Paperwork Reduction Act of 1995, unless the information collection has a currently valid OMB Control Number. The OMB Control Number for this information collection is 0651-0031. Public burden for this form is estimated to average 12 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the information collection. Send comments regarding this burden estimate or any other aspect of this information collection, including suggestions for reducing this burden to the Chief Administrative Officer, United States Patent and Trademark Office, P.O. Box 1450, Alexandria, VA 22313-1450 or email InformationCollection@uspto.gov. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. The United States Patent and Trademark Office (USPTO) collects the information in this record under authority of 35 U.S.C. 2. The USPTO's system of records is used to manage all applicant and owner information including name, citizenship, residence, post office address, and other information with respect to inventors and their legal representatives pertaining to the applicant's/owner's activities in connection with the invention for which a patent is sought or has been granted. The applicable Privacy Act System of Records Notice for the information collected in this form is COMMERCE/PAT-TM-7 Patent Application Files, available in the Federal Register at 78 FR 19243 (March 29, 2013). https://www.govinfo.gov/content/pkg/FR-2013-03-29/pdf/2013-07341.pdf

Routine uses of the information in this record may include disclosure to: 1) law enforcement, in the event that the system of records indicates a violation or potential violation of law; 2) a Federal, state, local, or international agency, in response to its request; 3) a contractor of the USPTO having need for the information in order to perform a contract; 4) the Department of Justice for determination of whether the Freedom of Information Act (FOIA) requires disclosure of the record; 5) a Member of Congress submitting a request involving an individual to whom the record pertains, when the individual has requested the Member's assistance with respect to the subject matter of the record; 6) a court, magistrate, or administrative tribunal, in the course of presenting evidence, including disclosures to opposing counsel in the course of settlement negotiations; 7) the Administrator, General Services Administration (GSA), or their designee, during an inspection of records conducted by GSA under authority of 44 U.S.C. 2904 and 2906, in accordance with the GSA regulations and any other relevant (i.e., GSA or Commerce) directive, where such disclosure shall not be used to make determinations about individuals; 8) another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)); 9) the Office of Personnel Management (OPM) for personnel research purposes; and 9) the Office of Management and Budget (OMB) for legislative coordination and clearance.

If you do not furnish the information requested on this form, the USPTO may not be able to process and/or examine your submission, which may result in termination of proceedings, abandonment of the application, and/or expiration of the patent.

Additional Uses

Additional USPTO uses of the information in this record may include disclosure to: 1) the International Bureau of the World Intellectual Property Organization, if the record is related to an international application filed under the Patent Cooperation Treaty; 2) the public i) after publication of the application pursuant to 35 U.S.C. 122(b), ii) after issuance of a patent pursuant to 35 U.S.C. 151, iii) if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections, or an issued patent, or iv) without publication of the application or patent under the specific circumstances provided for by 37 CFR 1.14(a)(1)(v)-(vii); and/or 3) the National Archives and Records Administration, for inspection of records.

Electronic Patent Application Fee Transmittal

Application Number:	17460145
Filing Date:	27-Aug-2021
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Filer:	Nicholas Joseph Gallo/Mary Frances Heeks
Attorney Docket Number:	264248.000100

Filed as Small Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
NOTICE OF APPEAL	2401	1	336	336
Post-Allowance-and-Post-Issuance:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
		Total in USD ($)		**336**

Electronic Acknowledgement Receipt	
EFS ID:	48655635
Application Number:	17460145
International Application Number:	
Confirmation Number:	3314
Title of Invention:	Real-Time Accent Conversion Model
First Named Inventor/Applicant Name:	Maxim Serebryakov
Customer Number:	11951
Filer:	Nicholas Joseph Gallo/Mary Frances Heeks
Filer Authorized By:	Nicholas Joseph Gallo
Attorney Docket Number:	264248.000100
Receipt Date:	28-SEP-2023
Filing Date:	27-AUG-2021
Time Stamp:	14:56:40
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$336
RAM confirmation Number	E20239RE57197423
Deposit Account	
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Notice of Appeal Filed	20230928_Notice_of_Appeal_264248_000100.pdf	194594 d42498d42b8db1e27a0dd07a9b9bdab837b2daea	no	2
Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	37967 9a9a93ffbc1d65a32bee9101efc0efc80d09b1bf	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	232561		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/460,145	**11/22/2023 04:18:00 PM Z ET**	**264248.000100**

Title of Invention

Real-Time Accent Conversion Model

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	3314	FILED BY	Marlene Chappe
PATENT CENTER #	63362146	FILING DATE	08/27/2021
CUSTOMER #	11951	FIRST NAMED INVENTOR	Maxim Serebryakov
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	Nicholas Gallo

Documents

TOTAL DOCUMENTS: 2

DOCUMENT		PAGES	DESCRIPTION	SIZE (KB)
20231122_Appeal_Brief_264248_000100.pdf		23	-	185 KB
20231122_Appeal_Brief_264248_000100-AP.B.pdf	(1-14)	14	Appeal Brief Filed	166 KB
20231122_Appeal_Brief_264248_000100-CLM.pdf	(15-23)	9	Claims	98 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)

20231122_Appeal_Brief_264248_000100.pdf	ACF5E4D26D1A52BA0E13571E3B2A3D46C86F9AFAEC7DEDF3060E80ACBB21D38CDC24B8C85FD3989234EFE15D595539678B5B93142AF220B4B07A850318524B47
20231122_Appeal_Brief_264248_000100-AP.B.pdf	A21ACD08CD2EBF8B4F886B722CA115A78CD896A3874FF66F537E37EDF000BF2EDC26E362DCE509DE368368F852B031FC5180C61EDE60B9AE098B1AD5802DD5DF
20231122_Appeal_Brief_264248_000100-CLM.pdf	D75D407301D87830D227B23E03DBE667E697C6E69CD8FEC11EF65417B8735A76B8FAA69046ED9DED255B663F83F4DC8174C5900F720C827D0349DB3741C8A89B

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PATENT
Docket No.: **264248.000100**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventors	:	Serebryakov and Zhang	Examiner: Bryan S. Blankenagel
Applicant	:	Sanas.ai Inc.	
Serial No.	:	17/460,145	
Cnfrm. No.	:	3314	Art Unit: 2658
Filed	:	April 27, 2021	
For	:	REAL-TIME ACCENT CONVERSION MODEL	

APPEAL BRIEF

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

Pursuant to 37 CFR § 41.37, Appellant hereby files an appeal brief. Pursuant to 37 CFR § 41.20(b)(2)(i), no fee is due. Should it be determined that a fee is required, you are hereby authorized to charge any fee to Deposit Account No. 50-0436.

I. REAL PARTY IN INTEREST

The real party in interest is Sanas.ai Inc., the assignee of record, having a principal place of business at 437 Lytton Avenue, Palo Alto, California 94301.

II. RELATED APPEALS AND INTERFERENCES

None.

III. STATUS OF CLAIMS

Claims 1-13 and 15-23 are currently pending and rejected in the above-identified application and are the subject of the present appeal. No claims have been allowed.

IPR2026-00272
Krisp Technologies EX1002 Page 621

IV. STATUS OF AMENDMENTS

On September 6, 2023, Appellant filed a response to the non-final Office action mailed on July 14, 2023 (the "Non-Final Office Action"). The Office entered the response and issued a final Office action mailed on September 18, 2023 (the "Final Office Action"). Appellant filed a Notice of Appeal on September 28, 2023.

V. SUMMARY OF CLAIMED SUBJECT MATTER

With respect to independent claim 1, a system is disclosed, by way of example in element 100 of FIG. 1, element 200 of FIG. 2, element 300 of FIG. 3, and paragraphs [0017]-[0048] of the above-identified application. The system includes at least one processor as disclosed in element 102 of FIG. 1 and paragraphs [0017]-[0019], for example. The system also includes a non-transitory computer-readable medium including program instructions stored thereon that are executable by the at least one processor, as disclosed in element 104 of FIG. 1 and paragraphs [0009]-[0010], [0019], and [0051].

A first machine-learning algorithm is trained with first audio data comprising speech content captured from a plurality of different speakers having a first accent, as explained in paragraph [0033] of the above-identified application. The training includes aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers, as also described n paragraph [0033] as well as paragraphs [0039] and [0042]. The first machine-learning algorithm is then applied to speech content received via at least one microphone to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, as disclosed in elements 402-406 of FIG. 4 and paragraphs [0032]-[0042. The speech content includes a set of phonemes associated with a first pronunciation of the received speech content as disclosed in original claims 1, 10, and 19, for example.

Based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, fourth audio data representative of the received speech content having the second accent is synthesized, as described in element 408 of FIG. 4 and paragraphs [0043]-[0045] of the above-identified application. The fourth audio data is synthesized using a second machine-learning-algorithm trained with (i) second

audio data comprising the first accent and (ii) third audio data comprising the second accent, as also explained in element 408 of FIG. 4 and paragraphs [0043]-[0045]. Additionally, synthesizing the fourth audio data includes mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second, different phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, as described in paragraphs [0041]-[0043] and [0048].

The synthesized fourth audio data is then converted into a synthesized version of the received speech content having the second accent, as described in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045] of the above-identified application, for example. The synthesized version of the received speech content having the second accent includes the updated set of phonemes associated with the second pronunciation of the received speech content, as also explained in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045].

With respect to independent claim 10, a non-transitory computer-readable medium provisioned with program instructions that can be executed by at least one processor of a computing device is disclosed in element 104 of FIG. 1 and paragraphs [0009]-[0010], [0019], and [0051] of the above-identified application. A first machine-learning algorithm is trained with first audio data comprising speech content captured from a plurality of different speakers having a first accent, as explained in paragraph [0033] of the above-identified application. The training includes aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers, as also described n paragraph [0033] as well as paragraphs [0039] and [0042].

The first machine-learning algorithm is then applied to speech content received via at least one microphone to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, as disclosed in elements 402-406 of FIG. 4 and paragraphs [0032]-[0042. The speech content includes a set of phonemes

associated with a first pronunciation of the received speech content as disclosed in original claims 1, 10, and 19, for example.

Based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, fourth audio data representative of the received speech content having the second accent is synthesized, as described in element 408 of FIG. 4 and paragraphs [0043]-[0045] of the above-identified application. The fourth audio data is synthesized using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, as also explained in element 408 of FIG. 4 and paragraphs [0043]-[0045]. Additionally, synthesizing the fourth audio data includes mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second, different phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, as described in paragraphs [0041]-[0043] and [0048].

The synthesized fourth audio data is then converted into a synthesized version of the received speech content having the second accent, as described in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045] of the above-identified application, for example. The synthesized version of the received speech content having the second accent includes the updated set of phonemes associated with the second pronunciation of the received speech content, as also explained in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045].

With respect to independent claim 19, a method is disclosed in FIG. 4 and paragraphs [0031]-[0049] of the above-identified application. A first machine-learning algorithm is trained with first audio data comprising speech content captured from a plurality of different speakers having a first accent, as explained in paragraph [0033] of the above-identified application. The training includes aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers, as also described n paragraph [0033] as well as paragraphs [0039] and [0042]. The first machine-learning algorithm is then applied to speech content received via at least one microphone to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the

IPR2026-00272
Krisp Technologies EX1002 Page 624

received speech content, as disclosed in elements 402-406 of FIG. 4 and paragraphs [0032]-[0042. The speech content includes a set of phonemes associated with a first pronunciation of the received speech content as disclosed in original claims 1, 10, and 19, for example.

Based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, fourth audio data representative of the received speech content having the second accent is synthesized, as described in element 408 of FIG. 4 and paragraphs [0043]-[0045] of the above-identified application. The fourth audio data is synthesized using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, as also explained in element 408 of FIG. 4 and paragraphs [0043]-[0045]. Additionally, synthesizing the fourth audio data includes mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second, different phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, as described in paragraphs [0041]-[0043] and [0048].

The synthesized fourth audio data is then converted into a synthesized version of the received speech content having the second accent, as described in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045] of the above-identified application, for example. The synthesized version of the received speech content having the second accent includes the updated set of phonemes associated with the second pronunciation of the received speech content, as also explained in element 410 of FIG. 4 and paragraphs [0025], [0027], and [0045].

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. Whether claims 1-13 and 15-23 were improperly rejected under 35 U.S.C. § 103 as being unpatentable over U.S. Patent No. 10,163,451 to Dirac et al. ("Dirac") in view of U.S. Patent Application Publication No. 2014/0258462 to Hwang ("Hwang") in further view of U.S. Patent Application Publication No. 2015/0170642 to Peng et al. ("Peng").

VII. ARGUMENT

IPR2026-00272
Krisp Technologies EX1002 Page 625

A. The Office has Failed to Set Forth a *Prima Facie* Case of Obviousness to Reject Claims 1-13 and 15-23 as Being Unpatentable Over Dirac, Hwang, and Peng

A person of ordinary skill in the art in possession of Dirac would not be motivated to look to Peng and, in any event, Dirac has multiple deficiencies with respect to the limitations of the independent claims even if combined with Hwang and Peng as proposed by the Office. The rejection of claims 1-13 and 15-23 under § 103 is therefore respectfully traversed, as explained in detail below.

1. Peng's Speech-to-Text technology Cannot be Combined with Dirac and Even if Combined Fails to Teach Training the First Machine-Learning Algorithm of the Independent Claims

Dirac, Hwang, and Peng, alone or in combination, do not disclose or suggest, "train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent," as recited by claims 1 and 10 or "training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent," as recited by claim 19.

The Office has acknowledged on page 5 of the Final Office Action that Dirac and Hwang do not disclose or suggest the above-identified claim limitations but has asserted that Peng addresses the deficiencies of those references. Peng discloses "a system 100 for identifying a substitute pronunciation for replacement in an actual phonetic transcription of a term by a user" (emphasis added), as Peng explains in paragraph [0018]. Peng teaches in paragraph [0033] that the system advantageously utilizes a "confusion matrix database 106 that is generated when the system 100 is in the training mode to obtain a text-based transcription of the user's utterance of a term based on the substitute pronunciation." (Emphasis added). Thus, as Peng teaches in paragraph [0038], "the language models 174 generate the text-based transcription 168 of the audio data 160 corresponding to the user speaking the additional term in the natural language based on the updated phonetic transcription." (Emphasis added).

By converting audio speech to text, Peng's disclosure solves a different problem in a manner expressly taught away from by the instant application. For example, the above-identified application discloses in paragraph [0006]:

> Some other approaches have involved a speech-to-text (STT) conversion of input speech as a midpoint, followed by a text-to-

> speech (TTS) conversion to generate the output audio content. However, this type of STT-TTS approach cannot capture many of the nuances of input speech that can provide information beyond the meaning of the words themselves, such as the prosody or emotion of the speaker. Further, a STT-TTS approach generally involves a degree of latency (e.g., up to several seconds) that makes it impractical for use in real-time communication scenarios such as an ongoing conversation (e.g., a phone call). (Emphasis added)

As another example, the above-identified application discloses in paragraph [0040] that "it should be noted that the ASR engine 302 may differ from some other speech recognition models that are configured predict and generate output speech, such as a speech-to-text model." Numerous other portions of the instant application make clear that conversion to text is advantageously avoided by the claimed technology.

Significantly, a person of ordinary skill in the art in possession of Dirac would not be motivated to look to Peng to achieve the claimed invention because Dirac similarly teaches away from speech-to-text transcription. Dirac discloses in the abstract that "[a]n input audio portion in a first spoken language may be received. It may be determined whether the input audio portion is substantially associated with the first accent, and if so, an output audio portion substantially associated with the second accent in the first spoken language may be outputted based, at least in part, on the translation model." (Emphasis added). Thus, the output of the Dirac's technology is audio, and not text as taught by Peng. The remainder of Dirac similarly discloses techniques for accent translation that do not contemplate any conversion of speech to text.

On page 6 of the Final Office Action, the Office has asserted that "[i]t would have been obvious to one of ordinary skill in the art…to modify the system of Dirac in view of Hwang by using the training of Peng (Peng para [0018]) on the system of Dirac in view of Hwang…." However, Appellant submits that a person of ordinary skill in the art in possession of Dirac would not be motivated to look to Peng because Peng purports to solve a problem not contemplated by Dirac and relating to representing differing pronunciations in text-based transcriptions. The combination proposed by the Office does not provide any teaching as to how the system disclosed by Dirac, which is configured to generate and output non-text linguistic representations (e.g., audio speech content), could possibly be combined with, and use the training of, Peng's text-based transcription system.

Even assuming, *arguendo*, that a person of ordinary skill in the art in possession of Dirac and Hwang would have been motivated look to Peng, that reference does not disclose or suggest the above-identified claim limitations in paragraphs [0018] and [0031], or elsewhere, as asserted by the Office. In paragraph [0018], Peng discloses that "when the system 100 operates in the training mode, the system 100 is trained to associate a received pronunciation of a term, or a portion of the term, by a user of a given language belonging to a particular like-pronunciation group with an expected pronunciation of those terms by an idealized native speaker in the given language." (Emphasis added). Peng also discloses in paragraph [0031]:

> The accent group identifier 113 provides the accent ID 124 to the confusion matrix manager 114. The confusion matrix manager 114 stores the accent ID 124 in the confusion matrix database 106. In some examples, the accent ID 124 is assigned to, or associated with, the user. In some examples, the accent ID 124 is assigned to a group of users, the group of users including the user. In some examples, the substitute pronunciation 122 is designated as the substitute pronunciation for the corresponding portion of the actual phonetic transcription 116 for a group of users, the group of users including the user. For example, the group of users can be associated with a similar characteristic, e.g., including a like-pronunciation group. Furthermore, the accent group identifier 113 associates the accent ID 124 with the substitute pronunciation 122. Specifically, the substitute pronunciation 122 for the corresponding portion of the actual phonetic transcription 116 can be associated with the accent ID 124. In some examples, a mapping is generated between the substitute pronunciation 122 and the accent ID 124. (Emphasis added)

As illustrated above, Peng discloses training a system to associate a pronunciation of a term by a user with an expected pronunciation of that term by an idealized native speaker. In Peng, users are assigned to a like-pronunciation group based on a group identifier and accent ID, which are correlated in a confusion matrix database to substitute pronunciations for phonetic transcriptions. Based on a mapping between substitute pronunciations and the accent ID, phonetic transcriptions can be generated that align with an expected pronunciation by an idealized native speaker.

However, Peng is silent with respect to any machine-learning algorithms. Specifically, the training mode disclosed by Peng is irrelevant to training a machine-learning algorithm, let alone with first audio data comprising speech content captured from different speakers having a first accent, as required by the pending claims. Peng discloses in paragraph

[0033] that "[t]he system 100 can utilize the confusion matrix database 106 that is generated when the system 100 is in the training mode to obtain a text-based transcription of the user's utterance of a term based on the substitute pronunciation." Appellant submits that the training mode disclosed by Peng generates a confusion matrix database and not a trained machine-learning algorithm.

Even if that deficiency is ignored, the training in Peng is to associate a received pronunciation of term(s) by a user, of a given language and belonging to a like-pronunciation group, with an expected pronunciation of those term(s) by an idealized native speaker in the given language, and not with audio data comprising speech content captured from different speakers having the same accent, as required by the above-identified claim limitations. In other words, Peng's training mode is based on an association of a user having one pronunciation the Office equates with an accent and an idealized native speaker having another, different pronunciation.

On page 5 of the Final Office Action, the Office has asserted that paragraph [0018] of Peng discloses a system trained to associate pronunciations and paragraph [0031] of Peng discloses an accent from a group of users. However, this analysis ignores that the pronunciations associated in the training mode are not from different speakers having the same accent, as recited in the above-identified claim limitations. Instead, the pronunciations associated in the training mode are from a user and idealized native speaker, each of which has a different accent. Moreover, Peng's generic disclosure that a group of users can have the same accent is irrelevant to training anything, and specifically does not teach training a machine-learning algorithm with audio data comprising speech content captured from different speakers of such a group of users.

2. Peng's Classifies Users, and Not Frames, and the Disclosed Alignment is Between Speakers Having Different Accents in Contrast to the Requirements of the Independent Claims

Dirac, Hwang, and Peng, alone or in combination, also do not disclose or suggest, "wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers [having a first accent]," as recited by claims 1, 10, and 19. The Office also has acknowledged that Dirac and Hwang do not disclose or suggest these claim limitations but has asserted that Peng addresses the deficiencies of those references in paragraphs [0025]-[0026] and [0030].

Peng discloses in paragraph [0026] that "the expected phonetic transcription 116 of the idealized native speaker of the natural language is aligned with the actual phonetic transcription of the particular user," but the above-identified claim limitations require aligning and classifying frames of captured speech content corresponding to respective different speakers that have the same accent. The idealized native speaker in Peng clearly does not have the same accent as the 'particular user.'

Moreover, Peng discloses in paragraph [0030] that "[t]he accent group identifier 113 can generate, or update, an accent identifier (ID) 124 based on the user characteristics 120. For example, for a particular geographic location associated with the user, e.g., a postal code associated with the user or a current city of the user, the accent group identifier 113 can associate a particular accent identifier with the user." Accordingly, the accent group identifier in Peng is used to classify users based on user characteristics, and not to classify frames of captured speech content.

Further, Peng's disclosure that users can belong to an accent group says nothing about aligning and classifying frames of speech content corresponding to respective ones of those users. Thus, even if somehow combined with the system of Dirac in view of Hwang, Peng fails to address the deficiencies of those references with respect to the pending claims.

3. Dirac's Generic Machine-learning Disclosure is Insufficient to Suggest the Derivation of the Independent Claims

Dirac, Hwang, and Peng, alone or in combination, also do not disclose or suggest "apply the first machine-learning algorithm to speech content received via at least one microphone…to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content," as recited by independent claims 1 and 10 or "applying the first machine-learning algorithm to speech content received via at least one microphone…to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content," as recited by independent claims 19.

The Office has cited col. 8, lines 15-32, col. 5, lines 9-27, col. 6, line 56-col. 7, line 6, and col. 7, lines 52-60 of Dirac as disclosing the above-identified claim limitations. Initially, the Office has acknowledged that Dirac fails to disclose training the first machine-learning algorithm, as claimed. Since Dirac does not train the claimed first machine-learning algorithm,

Dirac inherently cannot disclose applying that trained algorithm, let alone to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content.

Additionally, on page 5 of the Non-Final Office Action, the Office acknowledged that Dirac does not disclose speech content having a first accent that comprises a set of phonemes associated with a first pronunciation of speech content: "Dirac does not teach: wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content…." Since Dirac does not disclose that the speech content has a set of phonemes associated with a first pronunciation, Dirac inherently cannot derive a non-text linguistic representation of such a set of phonemes.

Appellant has carefully reviewed Dirac and is only able to identify two portions that reference machine-learning, which include col. 7, lines 56-60: "the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine-learning or other associated techniques" and col. 2, lines 45-49: "the accent translation system may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine-learning or other algorithm refinement techniques."

Further, the Office also asserts on page 3 of the Final Office Action that the accent audio characteristics of Dirac disclose the linguistic representations of the above-identified claim limitations: "where the accent audio characteristics of the accent sample sets are the linguistic representations." However, Dirac does not disclose or suggest that the accent audio characteristics are derived by applying the machine-learning that Dirac discloses is employed to update and refine the accent translation models, as required by the above-identified claim limitations.

In fact, Dirac does not disclose that the accent translation models themselves are machine-learning models or that the accent translation models are applied to generate the accent audio characteristics, and instead teaches in col. 5, lines 28-62, for example, performing a fast Fourier transform and determining frequency and amplitude in relation to time to derive the accent audio characteristics. Hwang and Peng also do not disclose or suggest the above-identified claim limitations.

4. Dirac Does Not Map Non-Text Linguistic Representations of Different Phonemes Associated with Different Pronunciations as Required by the Independent Claims

Dirac, Hwang, and Peng, alone or in combination, also do not disclose or suggest, "mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the speech content that is different from the first pronunciation of the speech content," as recited by independent claims 1, 10, and 19.

The Office has acknowledged on page 4 of the Non-Final Office Action that Dirac does not disclose a set of phonemes associated with a first pronunciation of speech content having a first accent or an updated set of phonemes associated with a second pronunciation of the speech content having a second accent: "Dirac does not teach: wherein the speech content having the first accent comprises a set of phonemes associated with a first pronunciation of the speech content…wherein the speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the speech content." Thus, Dirac inherently cannot teach mapping first and second non-text linguistic representations of a first phoneme of such a set of phonemes and a second phoneme of such an updated set of phonemes, respectively.

However, The Office has asserted on page 4 of the Final Office Action that Dirac discloses the above-identified claim limitations in col. 6, lines 27-55: "where the adjustment of audio characteristics of a first accent to more closely resemble audio characteristics of a second accent are considered mappings…." The Board's attention is respectfully directed to col. 10, lines 7-12 in which Dirac discloses that "it may be determined that speakers having the first accent tend to speak a particular phoneme with a higher pitch, while speakers having the second accent tend to [speak] that same phoneme with a lower pitch. In some examples, the accent translation model 641 may include instructions to decrease pitch whenever this particular phoneme is detected." (Emphasis added).

Thus, the audio characteristics adjusted in Dirac are for the same phoneme originating with speakers having different accents, and not multiple sets of phonemes, let alone associated with different pronunciations of speech content. Therefore, Dirac cannot disclose

mapping non-text linguistic representations of different phonemes from such multiple phoneme sets, as required by the pending claims.

Dirac does not disclose mapping between multiple sets of non-text linguistic representations of different phonemes because it is not concerned with, nor capable of compensating for, different pronunciations. In fact, neither "pronunciation," nor any form thereof, is found anywhere in Dirac's disclosure. Hwang and Peng also do not disclose or suggest the above-identified claim limitations.

The claimed technology advantageously utilizes machine-learning to convert input speech in a first accent into a synthesized version of the input speech in a second accent in real-time. As a result of a frame-wise breakdown of captured speech across multiple speakers having an accent, this technology can develop a learned linguistic representation of speech having the accent that is not speaker-specific. With the learned linguistic representation, this technology can work at the level of encoded linguistic data, does not need to predict and generate output speech or text as a midpoint for the conversion, and therefore advantageously facilitates accent conversion that accounts for pronunciation with extremely low latency, which is not possible based on the disclosures of the cited prior art references.

Accordingly, the Office has failed to set forth a *prima facie* case of obviousness to reject independent claims 1, 10, and 19 under Dirac, Hwang, and Peng. Since claims 2-9 and 21, 11-13 and 15-18, and 20 and 22-23 depend from and contain the limitations of claims 1, 10, and 19, respectively, they are patentable over the cited references at least in the same manner as claims 1, 10, and 19. As such, the rejection of claims 2-9 and 21, 11-13 and 15-18, and 20 and 22-23 as being obvious also should be reversed.

VIII. CLAIMS APPENDIX

A clean copy of the claims 1-13 and 15-23 involved in the appeal is included in the Claims Appendix.

IX. EVIDENCE APPENDIX

None.

X. RELATED PROCEEDINGS APPENDIX

None.

XI. CONCLUSION

For at least the foregoing reasons, the rejection of the appealed claims 1-13 and 15-23 in the Final Office Action should be reversed.

Respectfully submitted,

Date: November 22, 2023

/Nicholas J. Gallo/
Nicholas J. Gallo
Registration No. 60,758

Troutman Pepper Hamilton Sanders LLP
70 Linden Oaks, Suite 210
Rochester, New York 14625
Telephone: 585.270.2133
Facsimile: 800.541.9525

CLAIMS APPENDIX

CLEAN COPY OF CLAIMS ON APPEAL

1. A system, comprising:

at least one processor; and

non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor to cause the system to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different

IPR2026-00272
Krisp Technologies EX1002 Page 635

from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

2. The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

3. The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

4. The system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. The system of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

6. The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

7. The system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions are executable by the at least one processor such that the computing device is further configured to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. The system of claim 1, wherein the program instructions are executable by the at least one processor such that the computing device is further configured to:

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

10. A non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first

pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

11. The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

12. The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

13. The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having

the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.

14. (Canceled)

15. The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

16. The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions, when executed by the at least one processor, further cause the computing device to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

17. The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device.

IPR2026-00272
Krisp Technologies EX1002 Page 640

18. The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to:

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

19. A method comprising:

training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

applying the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising the second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated

IPR2026-00272
Krisp Technologies EX1002 Page 641

with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

20. The method of claim 19, further comprising:

receiving, in real time, continuous speech content having the first accent; and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

21. The computing device of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor such that the computing device is further configured to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the synthesized version of the received speech content having the second accent has the set of prosodic features.

22. The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content

in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

23. The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions and the method further comprises:

aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 643

Notice of References Cited	Application/Control No. 17/460,145	Applicant(s)/Patent Under Reexamination Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date YYYY-MM-DD	Name	CPC Classification	US Classification
*	A	US-20200169591-A1	2020-05-28	Ingel; Ben Avi	G10L13/08	1/1
	B					
	C					
	D					
	E					
	F					
	G					
	H					
	I					
	J					
	K					
	L					
	M					

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date YYYY-MM-DD	Country	Name	CPC Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in YYYY-MM-DD format are publication dates. Classifications may be US or foreign.

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

CPC							
Symbol						**Type**	**Version**
G10L	/	13	/	02		F	2013-01-01
G10L	/	15	/	02		I	2013-01-01
G06N	/	20	/	20		I	2019-01-01
G10L	/	25	/	27		I	2013-01-01
G10L	/	2015	/	022		A	2013-01-01

CPC Combination Sets								
Symbol						**Type**	**Set**	**Ranking**	**Version**
	/		/						

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	22	
/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	30 January 2024	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	2

U.S. Patent and Trademark Office — Part of Paper No.: 20240130

IPR2026-00272
Krisp Technologies EX1002 Page 645

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

INTERNATIONAL CLASSIFICATION					
CLAIMED					
G10L	/	13	/	02	
G10L	/	15	/	02	

NON-CLAIMED					
	/		/		

US ORIGINAL CLASSIFICATION	
CLASS	**SUBCLASS**
704	200

CROSS REFERENCES(S)						
CLASS	**SUBCLASS (ONE SUBCLASS PER BLOCK)**					

NONE (Assistant Examiner) (Date)	**Total Claims Allowed:** 22	
/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658 (Primary Examiner) 30 January 2024 (Date)	O.G. Print Claim(s) 1	O.G. Print Figure 2

IPR2026-00272
Krisp Technologies EX1002 Page 646

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIMS

Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original
1	1	11	10	19	19										
2	2	12	11	20	20										
3	3	13	12	10	21										
4	4	14	13	21	22										
5	5	-	14	22	23										
6	6	15	15												
7	7	16	16												
8	8	17	17												
9	9	18	18												

NONE		**Total Claims Allowed:**	
(Assistant Examiner)	(Date)	22	
/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	30 January 2024	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	2

U.S. Patent and Trademark Office — Part of Paper No.: 20240130

IPR2026-00272
Krisp Technologies EX1002 Page 647

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

CPC - Searched*		
Symbol	**Date**	**Examiner**
G10L13/02	01/30/2024	BB
G10L15/02	01/30/2024	BB

CPC Combination Sets - Searched*		
Symbol	**Date**	**Examiner**

US Classification - Searched*			
Class	**Subclass**	**Date**	**Examiner**

* See search history printout included with this form or the SEARCH NOTES box below to determine the scope of the search.

Search Notes		
Search Notes	**Date**	**Examiner**
Performed inventor search in PE2E, EAST	01/30/2024	BB
Performed text search	01/30/2024	BB
Performed text limited CPC search	01/30/2024	BB
Updated search	01/30/2024	BB
Appeal Brief Conference ended with agreement to allow	01/30/2024	BB

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 648

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	Art Unit
	BRYAN S BLANKENAGEL	2658

Interference Search			
US Class/CPC Symbol	**US Subclass/CPC Group**	**Date**	**Examiner**
G10L	13/02	01/30/2024	BB
G10L	15/02	01/30/2024	BB

/BRYAN S BLANKENAGEL/ Primary Examiner, Art Unit 2658	

IPR2026-00272
Krisp Technologies EX1002 Page 649

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	17/460,145	Serebryakov et al.
	Examiner	**Art Unit**
	BRYAN S BLANKENAGEL	2658

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

CLAIMS

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/20/2021	01/31/2022	08/04/2022	01/17/2023	07/11/2023	09/12/2023	01/30/2024		
1	1	✓	✓	✓	✓	✓	✓	=		
2	2	✓	✓	✓	✓	✓	✓	=		
3	3	✓	✓	✓	✓	✓	✓	=		
4	4	✓	✓	✓	✓	✓	✓	=		
5	5	✓	✓	✓	✓	✓	✓	=		
6	6	✓	✓	✓	✓	✓	✓	=		
7	7	✓	✓	✓	✓	✓	✓	=		
8	8	✓	✓	✓	✓	✓	✓	=		
9	9	✓	✓	✓	✓	✓	✓	=		
11	10	✓	✓	✓	✓	✓	✓	=		
12	11	✓	✓	✓	✓	✓	✓	=		
13	12	✓	✓	✓	✓	✓	✓	=		
14	13	✓	✓	✓	✓	✓	✓	=		
-	14	✓	✓	✓	✓	-	-	-		
15	15	✓	✓	✓	✓	✓	✓	=		
16	16	✓	✓	✓	✓	✓	✓	=		
17	17	✓	✓	✓	✓	✓	✓	=		
18	18	✓	✓	✓	✓	✓	✓	=		
19	19	✓	✓	✓	✓	✓	✓	=		
20	20	✓	✓	✓	✓	✓	✓	=		
10	21				✓	✓	✓	=		
21	22					✓	✓	=		
22	23						✓	=		

IPR2026-00272
Krisp Technologies EX1002 Page 650

PE2E SEARCH - Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
L73	3098	accent AND (pronunciation pronounc$3) AND phoneme	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 03:12 PM
L74	759	accent AND (pronunciation pronounc$3) AND phoneme AND train$3 SAME machine near3 learn$3	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 03:12 PM
L75	134	accent AND (pronunciation pronounc$3) AND phoneme AND train$3 SAME machine near3 learn$3 AND conver$4 near7 (second target)	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 03:23 PM
L76	21	accent AND (pronunciation pronounc$3) AND phoneme AND train$3 SAME machine near3 learn$3 AND conver$4 near7 (second target) near7 (voice accent)	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 03:23 PM
L77	51	accent AND (pronunciation pronounc$3) AND phoneme AND train$3 SAME machine near3 learn$3 AND (conver$4 chang$3 morph$3) near7 (second target) near7 (voice accent)	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 03:29 PM
L78	16	77 AND (G10L13/02 G10L15/02).cpc.	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 04:29 PM
L79	3	((serebryakov zhang).in. ("sanas.ai").asn.) AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT; USOCR; FIT (AU, AP, AT, BE, BG, BR, BY, CA, CH, CN, CS, CU, CZ, DD, DE, DK, EA, EE, EP, ES, FI, FR, GB, HR, HU, ID, IE, IL, IS, IT, JP, KR, LT, LU, LV, MA, OA, RU, SU, WO, MC, MD, MY, NL, NO, NZ, PH, PL, PT, RO, RS, SE, SG, SI, SK, TH, TN, TR, TW, UA, VN); FPRS; EPO;	OR	ON	ON	2024/01/30 04:29 PM

IPR2026-00272
Krisp Technologies EX1002 Page 651

			JPO; DERWENT; IBM_TDB)				

PE2E SEARCH - Search History (Interference)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	British Equivalents	Time Stamp
N1	4	(G10L13/02 G10L15/02).cpc. AND (accent AND synthesi$4 AND conver$4 AND machine near3 learn$3).clm.	(US-PGPUB; USPAT)	OR	ON	ON	2024/01/30 04:29 PM

Web Search History

date, time	web site	search string
1/30/2024 4:34:35 PM	Google Scholar	accent and (pronunciation OR pronounce) AND phoneme
1/30/2024 4:34:44 PM	Google Scholar	accent and (pronunciation OR pronounce) AND phoneme [before:2021]
1/30/2024 4:36:30 PM	Google Scholar	[before:2021] accent and (pronunciation OR pronounce) AND phoneme AND (conversion OR morphing OR changing)
1/30/2024 4:38:39 PM	Google Scholar	[before:2021] accent and (pronunciation OR pronounce) AND phoneme AND (conversion OR morphing OR changing) AND align
1/30/2024 4:38:43 PM	Google Scholar	[before:2021] accent and (pronunciation OR pronounce) AND phoneme AND (conversion OR morphing OR changing) AND align AND classify
1/30/2024 4:38:50 PM	Google Scholar	[before:2021] accent and (pronunciation OR pronounce) AND phoneme AND (conversion OR morphing OR changing) AND align AND classify AND speech
1/30/2024 4:38:57 PM	Google Scholar	[before:2021] accent and (pronunciation OR pronounce) AND phoneme AND (conversion OR morphing OR changing) AND align AND classify AND speech AND synthesize

IPR2026-00272
Krisp Technologies EX1002 Page 653

Bibliographic Data

Application No: 17/460,145

Foreign Priority claimed: ○ Yes ◉ No

35 USC 119 (a-d) conditions met: ☐ Yes ☑ No ☐ Met After Allowance

Verified and Acknowledged: _____ (Examiner's Signature) _____ (Initials)

Title: Real-Time Accent Conversion Model

FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
08/27/2021	704	2658	264248.000100
RULE			

APPLICANTS

Sanas.ai Inc., Pleasanton, CA,

INVENTORS

Maxim Serebryakov, Palo Alto, CA, UNITED STATES

Shawn Zhang, Pleasanton, CA, UNITED STATES

CONTINUING DATA

This application has PRO of 63185345 05/06/2021

FOREIGN APPLICATIONS

IF REQUIRED, FOREIGN LICENSE GRANTED**

09/08/2021

**** SMALL ENTITY ****

STATE OR COUNTRY

UNITED STATES

ADDRESS

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625
UNITED STATES

FILING FEE RECEIVED

$3,000

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

11951 7590 02/09/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	
ART UNIT	PAPER NUMBER
2658	

DATE MAILED: 02/09/2024

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

TITLE OF INVENTION: Real-Time Accent Conversion Model

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	$480	$0.00	$0.00	$480	05/09/2024

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the ENTITY STATUS shown above. If the ENTITY STATUS is shown as SMALL or MICRO, verify whether entitlement to that entity status still applies.

If the ENTITY STATUS is the same as shown above, pay the TOTAL FEE(S) DUE shown above.

If the ENTITY STATUS is changed from that shown above, on PART B - FEE(S) TRANSMITTAL, complete section number 5 titled "Change in Entity Status (from status indicated above)".

For purposes of this notice, small entity fees are 40% the amount of undiscounted fees, and micro entity fees are 20% the amount of undiscounted fees.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Maintenance fees are due in utility patents issuing on applications filed on or after Dec. 12, 1980. It is patentee's responsibility to ensure timely payment of maintenance fees when due. More information is available at www.uspto.gov/PatentMaintenanceFees.

PTOL-85 (Rev. 11/23)

IPR2026-00272
Krisp Technologies EX1002 Page 655

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via the USPTO patent electronic filing system.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications. **Because electronic patent issuance may occur shortly after issue fee payment, any desired continuing application should preferably be filed prior to payment of this issue fee in order not to jeopardize copendency.**

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

11951 7590 02/09/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via the USPTO patent electronic filing system or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

TITLE OF INVENTION: Real-Time Accent Conversion Model

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	$480	$0.00	$0.00	$480	05/09/2024

EXAMINER	ART UNIT	CLASS-SUBCLASS
BLANKENAGEL, BRYAN S	2658	704-200000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

❑ Change of correspondence address (or Change of Correspondence Address form PTO/AIA/122 or PTO/SB/122) attached.

❑ "Fee Address" indication (or "Fee Address" Indication form PTO/AIA/47 or PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1_____
2_____
3_____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ❑ Individual ❑ Corporation or other private group entity ❑ Government

4a. Fees submitted: ❑Issue Fee ❑Publication Fee (if required)

4b. Method of Payment: *(Please first reapply any previously paid fee shown above)*

❑ Electronic Payment via the USPTO patent electronic filing system ❑ Enclosed check ❑ Non-electronic payment by credit card (Attach form PTO-2038)

❑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. _____

5. **Change in Entity Status** (from status indicated above)

❑ Applicant certifying micro entity status. See 37 CFR 1.29

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.

❑ Applicant asserting small entity status. See 37 CFR 1.27

NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.

❑ Applicant changing to regular undiscounted fee status.

NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

IPR2026-00272
Krisp Technologies EX1002 Page 656

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 02/09/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

DATE MAILED: 02/09/2024

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(Applications filed on or after May 29, 2000)

The Office has discontinued providing a Patent Term Adjustment (PTA) calculation with the Notice of Allowance.

Section 1(h)(2) of the AIA Technical Corrections Act amended 35 U.S.C. 154(b)(3)(B)(i) to eliminate the requirement that the Office provide a patent term adjustment determination with the notice of allowance. See Revisions to Patent Term Adjustment, 78 Fed. Reg. 19416, 19417 (Apr. 1, 2013). Therefore, the Office is no longer providing an initial patent term adjustment determination with the notice of allowance. The Office will continue to provide a patent term adjustment determination with the Issue Notification Letter that is mailed to applicant approximately three weeks prior to the issue date of the patent, and will include the patent term adjustment on the patent. Any request for reconsideration of the patent term adjustment determination (or reinstatement of patent term adjustment) should follow the process outlined in 37 CFR 1.705.

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

IPR2026-00272
Krisp Technologies EX1002 Page 657

OMB Clearance and PRA Burden Statement for PTOL-85 Part B

The Paperwork Reduction Act (PRA) of 1995 requires Federal agencies to obtain Office of Management and Budget approval before requesting most types of information from the public. When OMB approves an agency request to collect information from the public, OMB (i) provides a valid OMB Control Number and expiration date for the agency to display on the instrument that will be used to collect the information and (ii) requires the agency to inform the public about the OMB Control Number's legal significance in accordance with 5 CFR 1320.5(b).

The information collected by PTOL-85 Part B is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450. Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. The United States Patent and Trademark Office (USPTO) collects the information in this record under authority of 35 U.S.C. 2. The USPTO's system of records is used to manage all applicant and owner information including name, citizenship, residence, post office address, and other information with respect to inventors and their legal representatives pertaining to the applicant's/ owner's activities in connection with the invention for which a patent is sought or has been granted. The applicable Privacy Act System of Records Notice for the information collected in this form is COMMERCE/PAT-TM-7 Patent Application Files, available in the Federal Register at 78 FR 19243 (March 29, 2013).
https://www.govinfo.gov/content/pkg/FR-2013-03-29/pdf/2013-07341.pdf

Routine uses of the information in this record may include disclosure to:

1) law enforcement, in the event that the system of records indicates a violation or potential violation of law;
2) a federal, state, local, or international agency, in response to its request;
3) a contractor of the USPTO having need for the information in order to perform a contract;
4) the Department of Justice for determination of whether the Freedom of Information Act (FOIA) requires disclosure of the record;
5) a Member of Congress submitting a request involving an individual to whom the record pertains, when the individual has requested the Member's assistance with respect to the subject matter of the record;
6) a court, magistrate, or administrative tribunal, in the course of presenting evidence, including disclosures to opposing counsel in the course of settlement negotiations;
7) the Administrator, General Services Administration (GSA), or their designee, during an inspection of records conducted by GSA under authority of 44 U.S.C. 2904 and 2906, in accordance with the GSA regulations and any other relevant (i.e., GSA or Commerce) directive, where such disclosure shall not be used to make determinations about individuals;
8) another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c));
9) the Office of Personnel Management (OPM) for personnel research purposes; and
10) the Office of Management and Budget (OMB) for legislative coordination and clearance.

If you do not furnish the information requested on this form, the USPTO may not be able to process and/or examine your submission, which may result in termination of proceedings, abandonment of the application, and/or expiration of the patent.

Notice of Allowability	Application No. 17/460,145	Applicant(s) Serebryakov et al.	
	Examiner BRYAN S BLANKENAGEL	Art Unit 2658	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☑ This communication is responsive to 11/22/2023.
   ☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2. ☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

3. ☑ The allowed claim(s) is/are 1-13 and 15-23 . As a result of the allowed claim(s), you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information , please see **http://www.uspto.gov/patents/init_events/pph/index.jsp** or send an inquiry to **PPHfeedback@uspto.gov.**

4. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

   Certified copies:

   a) ☐ All b) ☐ Some* c) ☐ None of the:

   1. ☐ Certified copies of the priority documents have been received.
   2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
   3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____ .

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____ .

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☑ Notice of References Cited (PTO-892)
2. ☐ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____.
3. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material _____.
4. ☐ Interview Summary (PTO-413), Paper No./Mail Date. _____.
5. ☐ Examiner's Amendment/Comment
6. ☑ Examiner's Statement of Reasons for Allowance
7. ☐ Other _____.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

IPR2026-00272
Krisp Technologies EX1002 Page 659

DETAILED ACTION

Notice of Pre-AIA or AIA Status

1. The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Response to Arguments

2. Applicant's arguments with respect to claim(s) 1-13 and 15-23 have been considered but are moot because the rejections have been withdrawn.

Allowable Subject Matter

3. Claims 1-13 and 15-23 allowed.

4. The following is an examiner's statement of reasons for allowance: The closest prior art of Dirac, Hwang, and Peng do not teach the limitations of the claims. After reviewing the arguments filed in the Appeal Brief of 11/22/2023, Examiner finds the arguments persuasive, and that the cited references do not teach the limitations of the claims, such as the training and application of the machine learning model and the synthesizing of fourth audio data using the method described in the claims. Hence, none of the cited prior art either alone or in combination thereof teaches the combination of limitations found in the claims.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably accompany the issue fee. Such submissions should be clearly labeled "Comments on Statement of Reasons for Allowance."

Conclusion

5. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 2020/0169591 A1 para [0142] teaches training a machine learning model to analyze audio data including accent, while para [0009] teaches introducing an accent into a dubbed version of a media stream.

6. Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRYAN S BLANKENAGEL whose telephone number is (571)270-0685. The examiner can normally be reached 8:00am-5:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

PATENT
Docket No.: 264248.000100

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventors	:	SEREBRYAKOV and ZHANG	
Applicant	:	SANAS.AI INC.	
Serial No.	:	17/460,145	Examiner: BRYAN S BLANKENAGEL
Cnfrm. No.	:	3314	Art Unit: 2658
Filed	:	27 August 2021	
For	:	REAL-TIME ACCENT CONVERSION MODEL	

AMENDMENT AFTER ALLOWANCE UNDER 37 CFR 1.312

Mail Stop Issue Fee
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In response to the February 9, 2024, Notice of Allowance, please amend the above-identified patent application as follows:

Amendments to the Claims begin on page 2 of this paper.

Remarks begin on page 9 of this paper.

IPR2026-00272
Krisp Technologies EX1002 Page 662

Amendments to the Claims

1. (Currently Amended) A system, comprising:

at least one processor; and

non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor to cause the system to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising [[the]] a second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

IPR2026-00272
Krisp Technologies EX1002 Page 663

2. (Currently Amended) The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

3. (Currently Amended) The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

4. (Previously Presented) The system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

5. (Previously Presented) The system of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

6. (Currently Amended) The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

7. (Currently Amended) The system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to:

IPR2026-00272
Krisp Technologies EX1002 Page 664

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

8. (Currently Amended) The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to transmit the synthesized version of the received speech content having the second accent to a second computing device.

9. (Currently Amended) The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to:

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

10. (Currently Amended) A non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising [[the]] a second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

11. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

12. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

13. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.

14. (Canceled)

15. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

16. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions, when executed by the at least one processor, further cause the computing device to:

align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

17. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device.

18. (Previously Presented) The non-transitory computer-readable medium of claim 10, wherein the program instructions, when executed by the at least one processor, further cause the computing device to:

receive, in real time, continuous speech content having the first accent; and

continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

19. (Currently Amended) A method comprising:

training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

applying the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising [[the]] a second accent, fourth audio data representative of the received speech content having the second accent, wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

~~convert~~ converting the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

20. (Previously Presented) The method of claim 19, further comprising:

receiving, in real time, continuous speech content having the first accent; and

continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

21. (Currently Amended) The ~~computing device~~ system of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor to further cause the system ~~such that the computing device is further configured~~ to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the synthesized version of the received speech content having the second accent has the set of prosodic features.

22. (Previously Presented) The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

23. (Currently Amended) The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and ~~the program instructions and~~ the method further comprises:

aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

IPR2026-00272
Krisp Technologies EX1002 Page 669

REMARKS

As set forth above, Applicant has amended claim 1-3, 6-7, 8-10, 19, 21, and 23 to correct minor informalities relating to antecedent basis issues. Applicant submits that these amendments result in no new matter nor any change in claim scope. Accordingly, the Office is respectfully requested to consider and enter these claim amendments.

Applicant appreciates the Office's Notice of Allowance of February 9, 2024 and notes that it has contemporaneously submitted an issue fee transmittal and issue fee payment. Applicant looks forward to the prompt issuance of the instant application.

Respectfully submitted,

Date: February 14, 2024

/Nicholas J. Gallo/
Nicholas J. Gallo
Registration No. 60,758

TROUTMAN PEPPER HAMILTON SANDERS LLP
70 Linden Oaks, Suite 210
Rochester, New York 14625
Telephone: (585) 270-2133
Facsimile: (800) 541-9525

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), by mail or fax, or via the USPTO patent electronic filing system.

By mail, send to: Mail Stop ISSUE FEE
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

By fax, send to: (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications. **Because electronic patent issuance may occur shortly after issue fee payment, any desired continuing application should preferably be filed prior to payment of this issue fee in order not to jeopardize copendency.**

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

11931 7590 02/09/2024
Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being transmitted to the USPTO via the USPTO patent electronic filing system or by facsimile to (571) 273-2885, on the date below.

(Typed or printed name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

TITLE OF INVENTION: Real-Time Accent Conversion Model

APPLN. TYPE	ENTITY STATUS	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	SMALL	$480	$0.00	$0.00	$480	05/09/2024

EXAMINER	ART UNIT	CLASS-SUBCLASS
BLANKENAGEL, BRYAN S	2658	704-200000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

☐ Change of correspondence address (or Change of Correspondence Address form PTO/AIA/122 or PTO/SB/122) attached.

☐ "Fee Address" indication (or "Fee Address" Indication form PTO/AIA/47 or PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) The names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) The name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 **Troutman Pepper Hamilton Sanders LLP**
2
3

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document must have been previously recorded, or filed for recordation, as set forth in 37 CFR 3.11 and 37 CFR 3.81(a). Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE: **SANAS.AI INC.**

(B) RESIDENCE: (CITY and STATE OR COUNTRY): **Pleasanton, CA**

Please check the appropriate assignee category or categories (will not be printed on the patent) : ☐ Individual ☒ Corporation or other private group entity ☐ Government

4a. Fees submitted: ☒ Issue Fee ☐ Publication Fee (if required)

4b. Method of Payment: *(Please first reapply any previously paid fee shown above)*

☒ Electronic Payment via the USPTO patent electronic filing system ☐ Enclosed check ☐ Non-electronic payment by credit card (Attach form PTO-2038)

☒ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment to Deposit Account No. **500436**

5. **Change in Entity Status** (from status indicated above)

☐ Applicant certifying micro entity status. See 37 CFR 1.29

NOTE: Absent a valid certification of Micro Entity Status (see forms PTO/SB/15A and 15B), issue fee payment in the micro entity amount will not be accepted at the risk of application abandonment.

☐ Applicant asserting small entity status. See 37 CFR 1.27

NOTE: If the application was previously under micro entity status, checking this box will be taken to be a notification of loss of entitlement to micro entity status.

☐ Applicant changing to regular undiscounted fee status.

NOTE: Checking this box will be taken to be a notification of loss of entitlement to small or micro entity status, as applicable.

NOTE: This form must be signed in accordance with 37 CFR 1.31 and 1.33. See 37 CFR 1.4 for signature requirements and certifications.

Authorized Signature: **/Nicholas J. Gallo/** Date: **February 14, 2024**

Typed or printed name: **Nicholas J. Gallo** Registration No. **60,758**

IPR2026-00272
Krisp Technologies EX1002 Page 671

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC PAYMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/460,145	**02/14/2024 02:13:30 PM Z ET**	**264248.000100**

Title of Invention

Real-Time Accent Conversion Model

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	3314	FILED BY	Mary Frances Heeks
PATENT CENTER #	64316664	AUTHORIZED BY	Nicholas Gallo
CUSTOMER #	11951	FILING DATE	08/27/2021
CORRESPONDENCE ADDRESS	-	FIRST NAMED INVENTOR	Maxim Serebryakov

Payment Information

PAYMENT METHOD	PAYMENT TRANSACTION ID	PAYMENT AUTHORIZED BY
DA / 500436	**E20242DE14336688**	**Nicholas Gallo**

PRE-AUTHORIZED ACCOUNT	PRE-AUTHORIZED CATEGORY
500436	**37 CFR 1.16 (National application filing, search, and examination fees); 37 CFR 1.17 (Patent application and reexamination processing fees); 37 CFR 1.19 (Document supply fees); 37 CFR 1.21 (Miscellaneous fees and charges)**

FEE CODE	DESCRIPTION	ITEM PRICE($)	QUANTITY	ITEM TOTAL($)
2501	UTILITY ISSUE FEE	480.00	1	480.00
			TOTAL AMOUNT:	**$480.00**

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
17/460,145	**02/14/2024 02:13:30 PM Z ET**	**264248.000100**

Title of Invention

Real-Time Accent Conversion Model

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	-
CONFIRMATION #	3314	FILED BY	Mary Frances Heeks
PATENT CENTER #	64316664	FILING DATE	08/27/2021
CUSTOMER #	11951	FIRST NAMED INVENTOR	Maxim Serebryakov
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	Nicholas Gallo

Documents

TOTAL DOCUMENTS: 4

DOCUMENT		PAGES	DESCRIPTION	SIZE (KB)
20240214_312_Amendment_264248_000100.pdf		9	-	131 KB
Warning: The attached file contains comments or annotations. Comments are not recommended and may cause processing problems with the document.				
20240214_312_Amendment_264248_000100-A.NA.pdf	(1-1)	1	Amendment after Notice of Allowance (Rule 312)	104 KB
20240214_312_Amendment_264248_000100-CLM.pdf	(2-8)	7	Claims	124 KB
20240214_312_Amendment_264248_000100-REM.pdf	(9-9)	1	Applicant Arguments/Remarks Made in an Amendment	103 KB
20240214_Issue_Fee_Transmit_264248_000100.pdf		1	Issue Fee Payment (PTO-85B)	152 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)
20240214_312_Amendment_264248_000100.pdf	18CEA2BB1836A5B7B9DFFDD1FFAF5B0C94A4ECC9F24C65329CD16775F52CF088E1654B166BEE712FFE72A528B1A4E2E277416ED8A0DB422DEBA6643D7B361D85
20240214_312_Amendment_264248_000100-A.NA.pdf	180496A3439EC65E22858372060A86F7617AADDBF03B11B0CD296FC4225699C76768467295C525C77A037AFF63F394256849088DD2C287A48A338C3FCCC2642F
20240214_312_Amendment_264248_000100-CLM.pdf	F10C2510FE3962B64AB22CEB3BC8D67792EB11A823E73E17646DAD639D3892BAA69680890649B582C2DF5D0D56B60C43F0D04ED8CB5DE58E59E98B5537C2AB6F
20240214_312_Amendment_264248_000100-REM.pdf	A36B25469F031D37EBCDC4E18CEF58453F2A5DB7BA72572452EB3A48EB1011F11E770580DA55198F121206A18785E2A063C387F7D25E2EF59367A2A7988CB598
20240214_Issue_Fee_Transmit_264248_000100.pdf	714FC36CADE7A46998623D67649E17BF596C7A1C9CF82E8CE2BB187349969E53F4608A6DD2222C0374124E8238BFD3BA9F630E5BA4343FB3DC3009ECEF94C88B

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 02/26/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
02/26/2024	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

BadumK@pepperlaw.com
USPTOinboxroc@pepperlaw.com

IPR2026-00272
Krisp Technologies EX1002 Page 676

Response to Rule 312 Communication	Application No.	Applicant(s)	
	17/460,145	Serebryakov et al.	
	Examiner	Art Unit	AIA (FITF) Status
	BRYAN S BLANKENAGEL	2658	Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

1. ☑ The amendment filed on 14 February 2024 under 37 CFR 1.312 has been considered, and has been:
   a) ☑ entered.
   b) ☐ entered as directed to matters of form not affecting the scope of the invention.
   c) ☐ disapproved because the amendment was filed after the payment of the issue fee.
      Any amendment filed after the date the issue fee is paid must be accompanied by a petition under 37 CFR 1.313(c)(1) and the required fee to withdraw the application from issue.
   d) ☐ disapproved. See explanation below.
   e) ☐ entered in part. See explanation below.
   f) ☐ not entered because the supplemental or corrected Application Data sheet (ADS)
      ☐ was not accompanied by a petition to accept an unintentionally delayed claim under 37 CFR 1.55 or 27 CFR 1.78;
      ☐ did not identify the information being changed in accordance with 37 CFR 1.76(c)(2);
      ☐ was not properly signed in accordance with 37 CFR 1.76(e) (or 37 CFR 1.33(b) for applications filed prior to September 16, 2012).

/BRYAN S BLANKENAGEL/
Primary Examiner, Art Unit 2658

IPR2026-00272
Krisp Technologies EX1002 Page 677

OK to enter /B.B./ 02/20/2024

PATENT
Docket No.: 264248.000100

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Inventors	:	SEREBRYAKOV and ZHANG	
Applicant	:	SANAS.AI INC.	
Serial No.	:	17/460,145	Examiner: BRYAN S BLANKENAGEL
Cnfrm. No.	:	3314	Art Unit: 2658
Filed	:	27 August 2021	
For	:	REAL-TIME ACCENT CONVERSION MODEL	

AMENDMENT AFTER ALLOWANCE UNDER 37 CFR 1.312

Mail Stop Issue Fee
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In response to the February 9, 2024, Notice of Allowance, please amend the above-identified patent application as follows:

Amendments to the Claims begin on page 2 of this paper.

Remarks begin on page 9 of this paper.

IPR2026-00272
Krisp Technologies EX1002 Page 678

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	04/02/2024	11948550	264248.000100	3314

11951 7590 03/13/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

ISSUE NOTIFICATION

The projected patent number and issue date are specified above. The patent will issue electronically. The electronically issued patent is the official patent grant pursuant to 35 U.S.C. § 153. The patent may be accessed on or after the issue date through Patent Center at https://patentcenter.uspto.gov/. The patent will be available in both the public and the private sides of Patent Center. Further assistance in electronically accessing the patent, or about Patent Center, is available by calling the Patent Electronic Business Center at 1-888-217-9197.

The USPTO is implementing electronic patent issuance with a transition period, during which period the USPTO will mail a ceremonial paper copy of the electronic patent grant to the correspondence address of record. Additional copies of the patent (i.e., certified and presentation copies) may be ordered for a fee from the USPTO's Certified Copy Center at https://certifiedcopycenter.uspto.gov/index.html. The Certified Copy Center may be reached at (800)972-6382.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment is 0 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Center (https://patentcenter.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Patents Stakeholder Experience (OPSE), Stakeholder Support Division (SSD) at (571)-272-4200.

IPR2026-00272
Krisp Technologies EX1002 Page 679

INVENTOR(s) (Please see PATENT CENTER site https://patentcenter.uspto.gov for additional inventors):

Maxim Serebryakov, Palo Alto, CA;
Shawn Zhang, Pleasanton, CA;

APPLICANT(s) (Please see PATENT CENTER site https://patentcenter.uspto.gov for additional applicants):

Sanas.ai Inc., Pleasanton, CA;

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The USA offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to encourage and facilitate business investment. To learn more about why the USA is the best country in the world to develop technology, manufacture products, and grow your business, visit SelectUSA.gov.

IPR2026-00272
Krisp Technologies EX1002 Page 680

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/460,145	08/27/2021	Maxim Serebryakov	264248.000100	3314

11951 7590 04/02/2024

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Suite 210
Rochester, NY 14625

EXAMINER	
BLANKENAGEL, BRYAN S	

ART UNIT	PAPER NUMBER
2658	

NOTIFICATION DATE	DELIVERY MODE
04/02/2024	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

BadumK@pepperlaw.com
USPTOinboxroc@pepperlaw.com

IPR2026-00272
Krisp Technologies EX1002 Page 681

APPLICATION NO.	ISSUE DATE	PATENT NO.
17/460,145	02-Apr-2024	11948550

Troutman Pepper Hamilton Sanders LLP (Rochester)
70 Linden Oaks
Rochester, NY 14625

EGRANT NOTIFICATION

Your electronic patent grant (eGrant) is now available, which can be accessed via Patent Center at https://patentcenter.uspto.gov

The electronic patent grant is the official patent grant under 35 U.S.C. 153. For more information, please visit https://www.uspto.gov/electronicgrants

IPR2026-00272
Krisp Technologies EX1002 Page 682