

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.,
Petitioner

v.

Sanas.ai Inc.,
Patent Owner

Petition for Post Grant Review Case No. PGR2026-00032
U.S. Patent No. 12,412,561

**PETITION FOR POST-GRANT REVIEW
OF U.S. PATENT NO. 12,412,561**

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	SUMMARY OF THE '561 PATENT.....	1
	DESCRIPTION OF THE '561 PATENT	1
	SUMMARY OF THE PROSECUTION HISTORY	2
	LEVEL OF SKILL OF A PERSON HAVING ORDINARY SKILL IN THE ART.....	2
III.	REQUIREMENTS UNDER 37 C.F.R. § 42.104	2
	GROUND FOR STANDING UNDER 37 C.F.R. § 42.104(A)	2
	IDENTIFICATION OF CHALLENGE UNDER 37 C.F.R. § 42.104(B)	3
	CLAIM CONSTRUCTION UNDER 37 C.F.R. § 42.104(B)(3)	3
IV.	SHOWING OF ANALOGOUS, PRIOR ART	3
	OVERVIEW OF <i>FEINAUER</i>	5
	OVERVIEW OF <i>FAN</i>	5
	OVERVIEW OF <i>PRABHAVALKAR</i>	9
	OVERVIEW OF <i>SARGSYAN</i>	10
	OVERVIEW OF <i>KALINLI</i>	10
	OVERVIEW OF <i>WATTS</i>	10
V.	GROUND 1: CLAIMS 1-2, 6-7, 11-12, AND 16-18 ARE OBVIOUS OVER <i>FEINAUER</i>, <i>PRABHAVALKAR</i>, AND <i>FAN</i>.....	11
	CLAIM 1	11
	1. Claim 1(Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”	11
	2. Claim 1(a)(1): “extracting, from a chunk of a stream of chunks of an input audio signal”	13
	a) Feinauer’s Teachings.....	13
	b) Prabhavalkar’s Teachings	13
	c) Motivation to Combine Feinauer and Prabhavalkar	15
	3. Claim 1(a)(2): “extracting, ... acoustic features”	16
	a) Feinauer’s Teachings.....	16
	b) Fan’s Teachings.....	16
	c) Motivation to Combine Feinauer-Prabhavalkar and Fan.....	17
	4. Claim 1(a)(3): “extracting...linguistic features, wherein the linguistic features have a reduced accent;”	18
	a) Fan’s Teachings.....	18
	b) Motivation to Combine Feinauer-Prabhavalkar and Fan.....	19
	5. Claim 1(b): “synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and”	20
	a) Fan’s Teachings.....	21
	b) Motivation to Combine Fan’s speaker embedding.....	23

6.	<i>Claim 1(c): “generating, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal.”</i>	27
a)	<i>Feinauer’s Teachings</i>	28
b)	<i>Fan’s Teachings</i>	30
c)	<i>Motivation to Combine Fan’s vocoder</i>	31
CLAIM 2		32
7.	<i>Claim 2: “The method of claim 1, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”</i>	32
CLAIM 6		33
1.	<i>Claim 6(Pre): “A computing system, comprising an input device, a vocoder, memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”</i>	33
2.	<i>Claim 6(a): “extract, from a chunk of a stream of chunks of an input audio signal obtained via the input device, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”</i>	33
3.	<i>Claim 6(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”</i>	34
4.	<i>Claim 6(c): “and generate, via the vocoder and based on the spectrum representation, an output chunk of an output audio signal.”</i>	34
CLAIM 7		34
8.	<i>Claim 7: “The computing system for real-time correction of claim 6, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”</i>	34
CLAIM 11		34

9.	<i>Claim 11(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”</i>	34
10.	<i>Claim 11(a): “extract, from a chunk of a stream of chunks of an input audio signal, linguistic features and acoustic features, wherein the linguistic features have a reduced accent;”</i>	35
11.	<i>Claim 11(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”</i>	35
12.	<i>Claim 11(c): “generate, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal;”</i>	35
13.	<i>Claim 11(d): “and provide the output audio signal for output via an auditory channel of an output device.”</i>	35
CLAIM 12		37
14.	<i>Claim 12: “The non-transitory computer-readable storage medium of claim 11, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”</i>	37
CLAIM 16		37
15.	<i>Claim 16(Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”</i>	37
16.	<i>Claim 16(a): “dividing a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”</i>	37
17.	<i>Claim 16(b): “extracting, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”</i>	38
	a) <i>Prabhavalkar’s Teachings</i>	39
	b) <i>Motivation to Combine Prabhavalkar’s Context</i>	39
18.	<i>Claim 16(c): “obtaining a speaker embedding for a human speaker; and”</i>	40
19.	<i>Claim 16(d): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”</i>	41
CLAIM 17		41
20.	<i>Claim 17(Pre): “A computing system, comprising memory comprising instructions stored thereon, and one or more</i>	

	<i>processors coupled to the memory and configured to execute the stored instructions to:</i>	41
21.	<i>Claim 17(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”</i>	41
22.	<i>Claim 17(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”</i>	41
23.	<i>Claim 17(c): “obtain a speaker embedding for a human speaker; and generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”</i>	42
CLAIM 18		42
24.	<i>Claim 18(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”</i>	42
25.	<i>Claim 18(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”</i>	42
26.	<i>Claim 18(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”</i>	42
27.	<i>Claim 18(c): “obtain a speaker embedding for a human speaker; and”</i>	43
28.	<i>Claim 18(d): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”</i>	43
VI.	GROUND 2: CLAIMS 3, 8, 13 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND SARGSYAN	43
CLAIM 3		43
1.	<i>Claim 3: “The method of claim 1, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”</i>	43
	a) Sargsyan’s Teachings	44
	a) Motivation to Combine Sargsyan’s sampling rate.....	45
CLAIM 8		46
2.	<i>Claim 8: “The computing system for real-time correction of claim 6, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”</i>	46
CLAIM 13		47
1.	<i>Claim 13: “The non-transitory computer-readable storage medium of claim 11, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”</i>	47

VII.	GROUND 3: CLAIMS 4, 9, AND 14 ARE OBVIOUS OVER <i>FEINAUER</i>, <i>PRABHAVALKAR</i>, <i>FAN</i>, AND <i>KALINLI</i>	47
	CLAIM 4	47
	1. Claim 4: “The method of claim 1, further comprising processing, in parallel via at least two processing units of the computing system, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.”	47
	a) Kalinli’s Teachings	47
	b) Motivation to Combine	48
	CLAIM 9	49
	2. Claim 9: “The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.”	49
	CLAIM 14	50
	3. Claim 14	50
VIII.	GROUND 4: CLAIMS 5, 10, 15 ARE OBVIOUS OVER <i>FEINAUER</i>, <i>PRABHAVALKAR</i>, <i>FAN</i>, AND <i>WATTS</i>	50
	CLAIM 5	50
	1. Claim 5: “The method of claim 1, further comprising: extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, and generating the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and providing, as an input to at least one of the modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.”	50
	a) Motivation to Combine Watts	52
	CLAIM 10	53
	2. Claim 10	53
	CLAIM 15	53
	3. Claim 15	53
IX.	CONCLUSION	54
X.	MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)	55
	REAL PARTY-IN-INTEREST	55
	RELATED MATTERS	55
	LEAD AND BACK-UP COUNSEL	56
	37 C.F.R. § 42.8(B)(4) – SERVICE INFORMATION	56

TABLE OF AUTHORITIES

Cases

<i>Penumbra, Inc. v. RapidPulse, Inc.</i> , IPR2021-01466, Paper 34 (PTAB March 10, 2023)	6
<i>Phillips v. AWH Corp.</i> , 415 F.3d 1303 (Fed. Cir. 2005) (en banc)	3

Statutes

35 U.S.C § 102(a)(1)	5
35 U.S.C § 102(a)(2)	6, 9
35 U.S.C. § 102(d)(2)	6

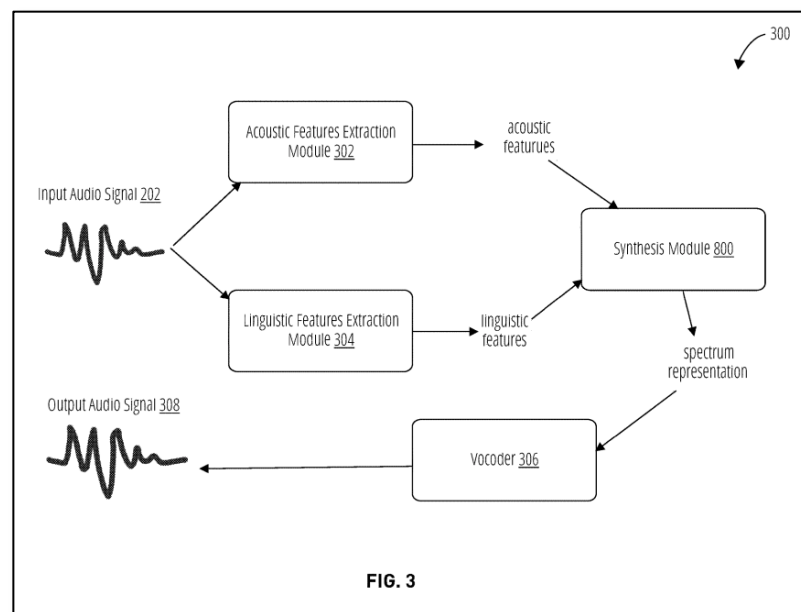
I. INTRODUCTION

Krisp Technologies, Inc. (“Petitioner”), respectfully requests post grant review (“PGR”) of claims 1-18 (“Challenged Claims”) of U.S. Patent No. 12,412,561 (“’561 Patent”).

II. SUMMARY OF THE ’561 PATENT

Description of the ’561 Patent

The ’561 Patent describes “real-time correction of an accent in an input audio signal.” EX1001, Abstract. “FIG. 3 is a block diagram showing a pipeline 300 for real-time correction of an accent in input audio signals, according to an example embodiment. The pipeline 300 (also referred to as a computational graph) may include an acoustic features extraction module 302, a linguistic features extraction module 304, a synthesis module 800, and a vocoder 306.” EX1001, 6:61-67.



EX1001, Fig. 3.

Summary of the Prosecution History

The '561 Patent claims priority to Provisional Application No. 63/297,901 filed January 10, 2022. The application that resulted as the '561 Patent faced one office action with a double patenting rejection for claims 1, 3-4, 7-8, 13-14, 17-18, and 20-27 to U.S. Patent No. 11,715,457. EX1002, 149-157. In response, patent owner filed a terminal disclaimer. EX1002, 158-159. The application was then allowed on August 5, 2025. EX1002, 170-171.

Level of Skill of a Person Having Ordinary Skill in the Art

A PHOSITA at the time of the '561 Patent would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of speech processing and machine learning, or a bachelor's degree in the same fields with at least four years of experience in the field of speech processing. Additional education or experience might substitute for the above requirements. EX1003, ¶27.

III. REQUIREMENTS UNDER 37 C.F.R. § 42.104

Grounds for Standing Under 37 C.F.R. § 42.104(a)

Petitioner certifies the '561 Patent is eligible for post-grant review and that the Petitioner is not barred or estopped from requesting a PGR challenging the claims of the '561 Patent.

Identification of Challenge Under 37 C.F.R. § 42.104(b)

In view of the prior art and evidence presented, the Challenged Claims of the '561 Patent are unpatentable and should be canceled. 37 C.F.R. § 42.204(b)(1). Petitioner requests the Challenged Claims be found unpatentable on the following grounds. 37 C.F.R. § 42.204(b)(2).

Proposed Grounds of Unpatentability	Exhibits
<u>Ground 1:</u> Claims 1-2, 6-7, 11-12, and 16-18 are obvious over <i>Feinauer, Prabhavalkar</i> and <i>Fan</i> .	1004, 1005, 1006
<u>Ground 2:</u> Claims 3, 8 and 13 are obvious over <i>Feinauer, Prabhavalkar, Fan</i> and <i>Sargsyan</i> .	1004, 1005, 1006, 1007
<u>Ground 3:</u> Claims 4, 9, and 14 are obvious over <i>Feinauer, Prabhavalkar, Fan</i> , and <i>Kalinli</i>	1004, 1005, 1006, 1008
<u>Ground 4:</u> Claims 5, 10 and 15 are obvious over <i>Feinauer, Prabhavalkar, Fan</i> and <i>Watts</i> .	1004, 1005, 1006, 1009

Claim Construction Under 37 C.F.R. § 42.104(b)(3)

In this proceeding, claims are interpreted under the same standard applied by Article III courts (i.e., the Phillips standard). 37 C.F.R. § 42.100(b); *Phillips v. AWH Corp.*, 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc). For purposes of this proceeding, Petitioner applies the plain and ordinary meaning of all claim terms as understood by a PHOSITA, unless specifically noted below.

IV. SHOWING OF ANALOGOUS, PRIOR ART

The prior art is analogous to the claimed invention of the '561 Patent. The prior art is from the same field of endeavor of the '561 patent, namely speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3; EX1004-Feinauer, Title, Abstract,

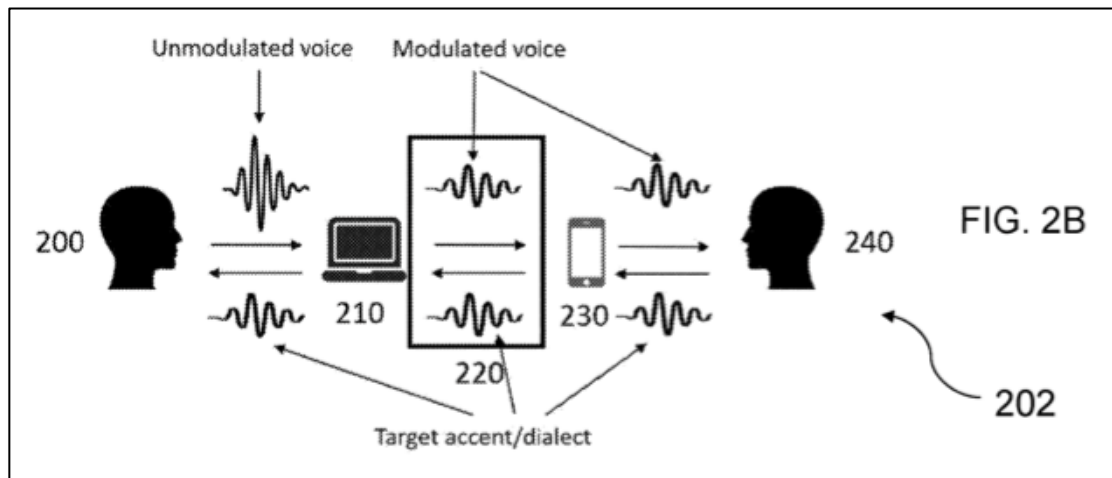
[0001], [0091]; EX1005-Prabhavalkar, 1:15-25, 2:41-3:56, 6:56-7:4, 15:16-22, EX1006-Fan, Abstract, [0134], [0201]; EX1007-Sargsyan, Abstract, [0003]-[0004], [0027], Fig. 14, EX1008-Watts, Abstract, 1:16-17; EX1009-Kalinli, Abstract, [0048]. EX1003, ¶¶61, 63, 69, 74, 80, 83, 87.

The prior art is also reasonably pertinent to at least one problem facing the '561 Patent inventors, namely at least one of:

- (A) real-time audio modification: EX1001, Title, Abstract, 1:19-22, 4:28-30, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086]; EX1005-Prabhavalkar, 17:33-51, 18:33-37, 20:16-21; 22:16-27;
- (B) accent conversion; EX1001, Title, Abstract, 1:19-22, 4:28-30, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086]; EX1006-Fan, Abstract, [0120], [0135], [0201];
- (C) feature extraction from audio data; EX1001, 4:21-23, 15:11-21, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086]; EX1005-Prabhavalkar, 2:44-50, 11:36-41, Fig. 2; EX1006-Fan, Abstract, [0120], [0135], [0201], EX1007-Sargsyan, [0045], [0070], [0072], EX1008-Watts, 4:37-42, 5:48-51, 6:21-23, Fig. 2; EX1009-Kalinli, Abstract, [0048], [0050]. EX1003, ¶¶64-66, 70-71, 75-77, 81, 84, 88.

Overview of *Feinauer*

U.S. Patent Appl. No. 2020/0193971 to Feinauer et al. ("*Feinauer*") was filed on December 13, 2018, and published on June 18, 2020, and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA. *Feinauer* teaches "methods for accent and dialect modification." EX1004-Feinauer, Abstract.



EX1004-Feinauer, Fig. 2B.

Overview of *Fan*

U.S. Patent Application Publication No. 2023/0223006 to Fan et. al. ("*Fan*")¹ was filed March 20, 2023, published July 13, 2023, and is a continuation of

¹ EX1006 is the U.S. Patent Appl. to Fan. Also included is the Chinese priority application, EX1013, and the certified English translation of the Chinese priority application, EX1013 at 1, with a certification that the translation is true and correct, EX1013 at 63. Also included is the PCT priority application, EX1015, and the certified English translation of the PCT priority application, EX1015 at 1, with a certification that the translation is true and correct, EX1015 at 64. EX1014 shows

PCT/CN2021/117945 (EX1015), which was filed on September 13, 2021, and claims priority to Chinese application 202010996501.5 (EX1013) filed September 21, 2020. *See* EX1014-FanHistory; EX1003, ¶¶67, 72.

Fan is entitled to an effective filing date of September 21, 2020 (and therefore qualifies as prior art to the '561 Patent) because *Fan* was filed as a continuation of a then co-pending PCT application which designated the U.S., includes a specific reference to the priority benefits of both its prior PCT and Chinese applications (including in an ADS), names at least one common inventor (Fan) with its prior PCT and Chinese applications, and the corresponding PCT and Chinese applications describe the subject matter relied on in *Fan* as prior art. MPEP §§ 1895, 2154.01; 35 U.S.C. §§ 102(a)(2) and 102(d)(2). *Penumbra, Inc. v. RapidPulse, Inc.*, IPR2021-01466, Paper 34, at 29-35 (PTAB March 10, 2023) (precedential as to section II.E.3).

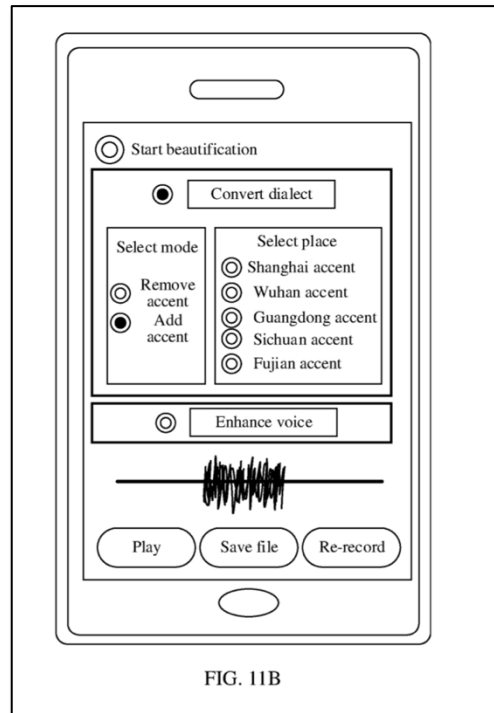
Petitioner provides the following chart showing support in the *Fan* priority applications for the citations to *Fan* herein. Mr. Schmandt compared the translations of both priority PTC and Chinese applications to *Fan* and determined that the *Fan* disclosures relied on in this Petition are fully supported by both the PCT and Chinese applications. EX1003, ¶¶67, 72.

Fan made a foreign priority claim within 12 months of the foreign priority application (at 3-4) and the Office acknowledged same (at 10).

<i>Fan</i> Publication (EX1006) (figure or paragraph)	<i>Fan</i> PCT (EX1015) (page, figure or paragraph)	<i>Fan</i> Priority (EX1013) (page, paragraph)
Abstract	47, Abstract and cover page at (57)	39, Abstract
Fig. 4	50, Fig. 4	42, Fig. 4
Fig. 5	51, Fig. 5	43, Fig. 5
Fig. 7	54, Fig. 7	46, Fig. 7
Fig. 9	55, Fig. 9	47, Fig. 9
Fig. 11B	58, Fig. 11B	50, Fig. 11B
Fig. 12	59, Fig. 12	51, Fig. 12
Fig. 13	59, Fig. 13	51, Fig. 13
[0007]	2, ¶2	2, ¶2
[0010]	3, ¶2	3, ¶1
[0023]	6, ¶2	6, ¶2
[0024]	6, ¶3	6, ¶3
[0025]	6, ¶4	6, ¶4
[0026]	6, ¶5	6, ¶5
[0045]	8, ¶6	8, ¶4
[0046]	9, ¶1	9, ¶1
[0047]	9, ¶2	9, ¶2
[0048]	9, ¶3	9, ¶3
[0049]	9, ¶4	9, ¶4
[0050]	10, ¶1	9, ¶5
[0053]	10, ¶4	10, ¶3
[0075]	13, ¶5	13, ¶4
[0081]	14, ¶1	14, ¶1
[0082]	14, ¶2	14, ¶2
[0085]	14, ¶5	14, ¶5
[0089]	15, ¶4	15, ¶4
[0090]	15, ¶5	15, ¶5
[0091]	16, ¶1	15, ¶6
[0092]	16, ¶2	16, ¶1
[0093]	16, ¶3	16, ¶2
[0094]	16, ¶4	16, ¶3
[0095]	17, ¶3	16, ¶4
[0097]	17, ¶3	17, ¶2
[0099]	18, ¶2	18, ¶2

[0106]	19, ¶6	19, ¶5
[0113]	21, ¶1	20, ¶6
[0116]	21, ¶4	21, ¶3
[0120]	22, ¶4	22, ¶3
[0124]	23, ¶2	22, ¶7
[0125]	23, ¶3	23, ¶1
[0126]	23, ¶4	23, ¶2
[0127]	23, ¶5	23, ¶3
[0128]	24, ¶1	23, ¶4
[0129]	24, ¶2	24, ¶1
[0134]	25, ¶2	25, ¶1
[0135]	25, ¶3	25, ¶2
[0152]	28, ¶6	28, ¶4
[0153]	28, ¶7	28, ¶5
[0154]	28, ¶8	28, ¶6
[0155]	28, ¶9	28, ¶7
[0156]	28, ¶10	28, ¶8
[0158]	29, ¶2	28, ¶10
[0159]	29, ¶3	29, ¶1
[0160]	29, ¶4	29, ¶2
[0161]	29, ¶5	29, ¶3
[0162]	29, ¶6	29, ¶4
[0166]	30, ¶2	29, ¶8
[0167]	30, ¶3	30, ¶1
[0170]	30, ¶6	30, ¶4
[0171]	30, ¶7	30, ¶5
[0201]	35, ¶3	35, ¶3

Fan teaches “a dialect conversion mode, for adding an accent to or removing an accent from the first voice.” EX1006-Fan, Abstract.



EX1006-Fan, Fig. 11B.

Overview of *Prabhavalkar*

U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“*Prabhavalkar*”) was filed on September 20, 2021, and published on October 1, 2024, qualifying as prior art under 35 U.S.C § 102(a)(2) AIA.

Prabhavalkar determines, “from the audio data for the utterance, speech frames,” and “successively processing chunks that each include different sets of speech frames with the speech recognition model.” EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. *Prabhavalkar* envisions “each chunk comprising a same predetermined number of speech frames each chunk having a first predetermined number of speech frames representing speech.” EX1005-*Prabhavalkar*, 6:65-7:4.

Overview of *Sargsyan*

U.S. Patent Appl. No. 2020/0066296 to Sargsyan et al. (“*Sargsyan*”) was filed August 21, 2019. *Sargsyan* claims priority to provisional application No. 62/720,702 filed August 21, 2018. *Sargsyan* is prior art to the ’561 Patent under 35 U.S.C. § 102(a)(1) AIA.

Sargsyan teaches “speech enhancement and noise suppression” EX1007-Sargsyan, Abstract. *Sargsyan* teaches to “expand low bitrate audio (e.g., 8 kHz) in conference calls on the server side and make it so that people calling in from low bitrate cellular lines sound much richer for other participants in the call (from 16 kHz to 44.1 kHz).” EX1007-Sargsyan, [0027].

Overview of *Kalinli*

U.S. Patent Application No. 2012/0116756 to Kalinli (“*Kalinli*”) was filed November 10, 2010, and published May 10, 2012. *Kalinli* is prior art to the ’561 Patent under 35 U.S.C. § 102(a)(1) AIA.

Kalinli teaches “spoken language processing method for tone/intonation recognition,” using parallel processing. EX1009-Kalinli, Abstract, [0048].

Overview of *Watts*

U.S. Patent No. 7,319,959 to Watts (“*Watts*”) was filed May 14, 2003, and published on January 15, 2008. *Watts* is prior art to the ’561 Patent under 35 U.S.C. § 102(a)(1) AIA.

Watts teaches “processing an audio signal” to determine “phoneme-level classification.” EX1008-Watts, Abstract.

V. GROUND 1: CLAIMS 1-2, 6-7, 11-12, AND 16-18 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, AND FAN*.

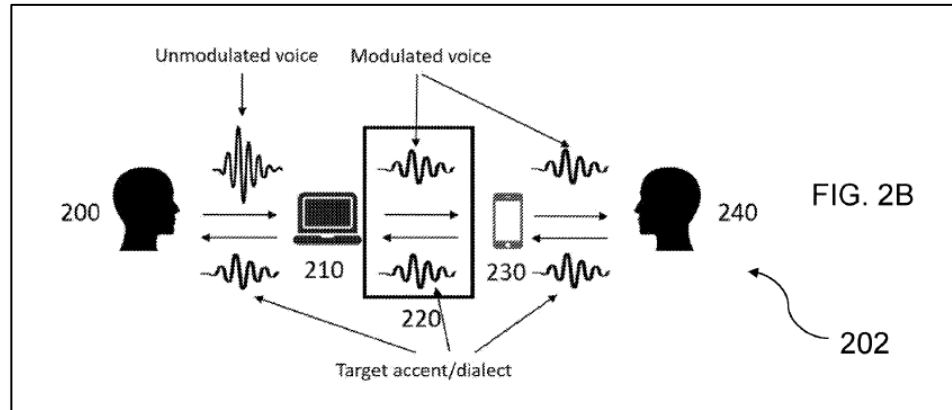
Claim 1

1. Claim 1(Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”

To the extent the preamble is limiting, *Feinauer* teaches a *method for real-time accent correction* in an input audio signal² that receives an unmodulated voice signal, processes it, and selects “a target dialect and accent to use to modify [the] voice communications based on a context.” EX1004-Feinauer, [0001]-[0003], [0046], [0049]. For example, *Feinauer’s* accent conversion method and system may modify spoken words to within a spoken target accent strength threshold, such as from a Japanese accent to a German accent during an audio/video call. EX1004-Feinauer, [0036], [0041], [0086]. As illustrated below, *Feinauer’s* method

² Claim language is *italicized*. All drawing annotations added. Citations to a claim limitation and/or section number are to the limitation’s mapping in this Petition and are incorporated by reference, including the Motivations to Combine (MTC) and showing of reasonable expectation of success (REOS) provided in the cross-referenced section. All REOS include the benefits described for the respective MTC. All emphases added unless otherwise noted.

specifically “*modifies*] *an accent* and/or dialect *in real-time* to sound more similar to an accent and/or dialect which would achieve a desired outcome.”



EX1004-Feinauer, [0035]-[0036], [0041], Fig. 2B. *Feinauer* explains its process may be performed within a “lag duration...that sounds like *real-time* to a user (e.g. 100ms).” EX1004-Feinauer, [0086]. A PHOSITA thus would have recognized *Feinauer* teaches a method for real-time correction of accent in an input audio signal. EX1003, ¶90.

Feinauer’s speech conversion process may be implemented as “software or programmable code embedded in a computer readable medium that is executed by a processor,” and is thus *implemented by a computing system*. EX1003, ¶91 citing EX1004-Feinauer, [0031]-[0032], [0043], [0121], Fig. 1.

2. Claim 1(a)(1): “extracting, from a chunk of a stream of chunks of an input audio signal”

a) Feinauer’s Teachings

Feinauer teaches receiving an unmodulated voice signal via a microphone. EX1004-Feinauer, [0031]-[0032], [0042], [0044]-[0046], [0056], Fig. 1; EX1003, ¶¶92-93. *Feinauer* is clear that thereafter the unmodulated voice signal is processed “to provide a modulated voice with a target accent.” EX1004-Feinauer, [0045], Figs. 2B. Though *Feinauer* expressly states its method “*may modify the accent* and/or dialect of a conversation, or portions thereof, *on any level of granularity*, including but not limited on a per sound basis, per word basis, set or string of words basis, on a whole or portion of a sentence structure basis,” *Feinauer* provides no further detail regarding how its input unmodulated voice signal is divided into these “portions” for processing. EX1004-Feinauer, [0054], [0057]; EX1003, ¶94. A PHOSITA would have looked to *Prabhavalkar* for detailed description of implementing known methods for dividing a voice signal into input chunks.

b) Prabhavalkar’s Teachings

Prabhavalkar describes methods of speech processing using neural networks. EX1005-Prabhavalkar, 1:15-25, 2:41-3:56. *Prabhavalkar*’s techniques are applicable to a variety of speech processing applications, including using neural networks to form a sequence-to-sequence model that “performs acoustic modelling, pronunciation modelling, and language modelling.” EX1005-Prabhavalkar, 3:22-24.

Notably, *Prabhavalkar* specifically includes embodiments comprising a “sequence-to-sequence model without finite state transducers, a lexicon, or a text normalization module.” EX1005-Prabhavalkar, 3:25-29; EX1003, ¶95.

Prabhavalkar’s system includes a computer system and microphone that receives an utterance from a user and generates audio data representing the utterance. EX1005-Prabhavalkar, 11:19-45, Fig. 2. *Prabhavalkar* then determines, “from the audio data for the utterance, **speech frames**,” and “successively processing **chunks** that each include different sets of speech frames with the speech recognition model.” EX1005-Prabhavalkar, 6:56-7:4, 15:16-22; EX1003, ¶96. *Prabhavalkar* explicitly envisions “**each chunk comprising a same predetermined number of speech frames**, each chunk having a first predetermined number of speech frames representing speech.” *Id.* In a specific example, *Prabhavalkar* teaches a *speech audio signal is divided into a stream of input chunks, each including 5 predetermined (pre-defined) frames of the speech audio signal.* EX1005-Prabhavalkar, 20:14-20, 22:1-38; EX1003, ¶96.

Prabhavalkar characterizes its approach as “streaming,” wherein chunks of speech are processed “while audio data of an utterance is still being received and without waiting for the utterance to be completed.” EX1005-Prabhavalkar, 6:4-23, 18:28-41, 2:14-40. A PHOSITA would have thus recognized *Prabhavalkar* teaches

performing a *real-time* speech processing method, like *Feinauer*. EX1003, ¶97 *citing* EX1005-*Prabhavalkar*, 6:24-28, 18:28-41.

c) Motivation to Combine *Feinauer* and *Prabhavalkar*

A PHOSITA would have been motivated to employ *Prabhavalkar*'s method of dividing audio into input chunks as an operable method of performing sampling (dividing speech into "portions") in *Feinauer*'s real-time accent conversion method. EX1003, ¶¶98-99 *citing* EX1004-*Feinauer*, [0057]; EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22, 17:33-51, 18:18-41, 22:1-38. A PHOSITA would have been motivated by *Prabhavalkar*'s express statement that a chunk-based approach may "limit the amount of latency incurred," increasing speed and performance compared to the unmodified *Feinauer* method. EX1003, ¶100 *citing* EX1005-*Prabhavalkar*, 17:33-51. A PHOSITA would have recognized that applying *Prabhavalkar*'s chunking techniques to *Feinauer*'s system would have only required a reprogramming according to known techniques to produce the reasonable result of reducing the latency of *Feinauer*'s accent conversion method by processing frames in chunks having a predetermined number of frames, such as 5 or 10. EX1003, ¶101 *citing* EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22, 17:33-51, 25:55-62. A PHOSITA would have had a REOS at least because *Prabhavalkar* expressly describes the processor architecture necessary for implementing its method and extols its method's use in

speech applications, requiring only ordinary skill of a PHOSITA to combine. EX1003, ¶¶101-102 *citing* EX1005-Prabhavalkar, 25:55-62.

3. ***Claim 1(a)(2): “extracting, ... acoustic features”***

The '561 Patent provides examples of “acoustic features” including “pitch 206 (or main frequency (F0)), energy 208 (signal amplitude), and voice activity detection.” EX1001, 6:35-37, Fig. 2.

a) *Feinauer’s Teachings*

Feinauer teaches extracting “prosodic vocal features and characteristics” from input speech audio. EX1004-Feinauer, [0099], Fig. 6 (step 620); EX1003, ¶¶104-105. As discussed above regarding Claim 1(a)(1), a PHOSITA would have modified *Feinauer’s* speech “portions” to be *input chunks* per *Prabhavalkar*. See Section V.A.2.

Feinauer further envisions prosodic features including “any voice characteristics such as ***pitch, tone, volume, etc.***” EX1004-Feinauer, [0057]. EX1003, ¶105. A PHOSITA would have understood pitch, tone, and volume to each satisfy “acoustic features” of the input chunk of user’s speech. *Id.* *Feinauer* does not provide further detail regarding how acoustic features are extracted, or by which module.

b) *Fan’s Teachings*

Fan teaches a voice conversion method for “adding an accent to or removing an accent from” a user’s voice. EX1006-Fan, Abstract; EX1003, ¶106. Like *Feinauer*, *Fan* teaches modifying acoustic features such as the user’s tone, pitch,

rhythm, and speaking speed. EX1006-Fan, [0050], [0106]; EX1003, ¶106. *Fan* expressly achieves its accent modification using “hardware or the software includes one or more **modules**” including a plurality of sub-feature extraction models, such as a style separation model. EX1006-Fan, [0023], [0089]-[0090], Fig. 5; EX1003, ¶107.

c) Motivation to Combine *Feinauer-Prabhavalkar* and *Fan*

A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module (such as hardware employing *Fan*’s style separation model) as an operable method of extracting acoustic features including pitch, tone, volume, rhythm, and/or speaking speed in *Feinauer*’s real-time accent conversion method. EX1003, ¶108. A PHOSITA would have appreciated the lengthy detail provided by *Fan* regarding hardware, software, and neural network configurations capable of extracting acoustic features, leading to a REOS. EX1006-Fan, [0046]-[0050], [0090]-[0094], Figs. 5, 7; EX1003, ¶108. A PHOSITA would have recognized the modification constitutes applying a known technique (*Feinauer* and *Fan*’s steps of acoustic feature extraction) to a known device (*Fan*’s hardware and software modules) with predictable results, further motivating the combination. *Id.* *Feinauer* in view of *Fan* teaches *extracting, by an acoustic features extraction module (Fan*’s “style separation model”), *acoustic features* including pitch, volume, and tone (per *Feinauer*) as well as rhythm and speed (per *Fan*). *Fan* further teaches

its style separation module uses “attention” to “focus more on different key parts of an input instead of the entire input.” EX1006-Fan, [0048], [0090], [0094]; EX1003, ¶108.

4. Claim 1(a)(3): “extracting...linguistic features, wherein the linguistic features have a reduced accent;”

As discussed above, *Feinauer-Prabhavalkar-Fan* renders obvious extracting acoustic features from an input chunk. Further, *Feinauer* teaches identifying “individual words” in the user’s speech, and may determine “which words” to be accent-modified. EX1004-Feinauer, [0083], [0094], [0099]; EX1003, ¶109. A PHOSITA would have thus recognized *Feinauer* extracts linguistic features, but does not provide further detail regarding how linguistic features are extracted, or by which module.

The ’561 Patent states, “linguistic features 210 can be represented as Phonetic PosteriorGrams (PPGs).” EX1001, 7:6-26, Fig. 3.

a) Fan’s Teachings

Fan teaches a voice conversion method for “adding an accent to or removing an accent from” a user’s voice. EX1006-Fan, Abstract; EX1003, ¶111.

Fan teaches a voice feature extraction model that extracts phoneme posteriorgrams (PPGs) from the user’s speech representing the words spoken. EX1006-Fan, [0007], [0085]-[0086], Fig. 5; EX1003, ¶¶112-113. A PPG “describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and

is equivalent to recognizing the phoneme. The PPG feature is for retaining the content information of the first voice.” EX1006-Fan, [0007], [0085]; EX1003, ¶113. As discussed, *Fan* expressly performs its methods using “hardware or the software includes one or more **modules**” including a plurality of sub-feature extraction models, such as a voice feature extraction model. EX1006-Fan, [0023], [0085], Fig. 5; EX1003, ¶113.

As supported by the prior art and the ’561 specification, a PHOSITA would have recognized that PPGs extracted from a user’s speech carry a reduced accent. See EX1001, 7:19-21, 15:22-32; EX1003, ¶113 *citing* EX1012-Ding, Abstract. Thus, *Feinauer-Prabhavalkar-Fan* renders obvious *extracting linguistic features with a reduced accent from an input chunk*.

b) Motivation to Combine *Feinauer-Prabhavalkar* and *Fan*

A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module (such as hardware employing *Fan*’s voice feature extraction model) as an operable method of extracting linguistic features including PPGs in *Feinauer*’s real-time accent conversion method. EX1003, ¶114. A PHOSITA would have appreciated the lengthy detail provided by *Fan* regarding hardware, software, and neural network configurations capable of extracting linguistic features, leading to a REOS. EX1006-Fan, [0046]-[0050], [0090]-[0094], Figs. 5, 7; EX1003, ¶114. A PHOSITA would have recognized the modification

constitutes applying a known technique (*Feinauer* and *Fan*'s steps of linguistic feature extraction) to a known device (*Fan*'s hardware and software modules) with predictable results, further motivating the combination. *Id.*

5. Claim 1(b): “synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and”

As discussed in Claim 1(a)(2)-1(a)(3), *Feinauer-Prabhavalkar-Fan* teaches extracting acoustic features and linguistic features. *Feinauer-Prabhavalkar-Fan* further teaches a speaker embedding for a human user and renders obvious synthesizing a spectrum representation based on the above.

“a speaker embedding”

The '561 Patent describes a speaker embedding as a “multidimensional vector” that is “pretrained based on audio data including a recorded speech of a target speaker having a further accent.” EX1001, 15:33-38. The '561 Patent further describes the speaker embedding may include other “information on target speakers’ voice style not related to the accent of the target speaker.” EX1001, 11:39-41. A speaker embedding may be trained fixed, pre-trained fixed, or extracted via a pre-trained model from a speech audio signal in real-time. EX1001, 11:42-53. The speaker embedding may be utilized in such a way that voice acoustic features corresponding to the speaker embedding match voice acoustic features of a target speaker. *Id.*

Feinauer teaches adjusting the accent of the user to match a target accent.

EX1004-Feinauer, [0045]-[0046], [0096], [0103]-[0110] Figs. 2B, 7. *Feinauer's* system loads a relevant “translation model” corresponding to the particular target accent. EX1004-Feinauer, [0081], [0105], [0110], Fig. 7; EX1003, ¶117. *Feinauer's* translation model determines “attributes of a user before the target word was spoken so that their voice profile can be replicated in the real time modification and the replaced word flows with the vocal biometric profile of the user.” EX1004-Feinauer, [0038], [0065]; EX1003, ¶117.

a) *Fan's* Teachings

Similarly, *Fan* teaches that a user may modify the particular accent of input source speech, such as to a “Sichuan accent” or a “Guangdong accent.” EX1006-Fan, [0120], [0124], [0128]-[0130]; EX1003, ¶118. *Fan* envisions a “style embedding vector” that a PHOSITA would have recognized includes a representation of a target accent. EX1003, ¶118 *citing* EX1010-Accentron, 21; EX1006-Fan, [0050], [0094], [0127], Figs. 7, 9, 12.

Specifically, *Fan* explains a target voice (annotated green below) may be input to a style separation model (annotated red). EX1006-Fan, [0093]-[0094], [0127].

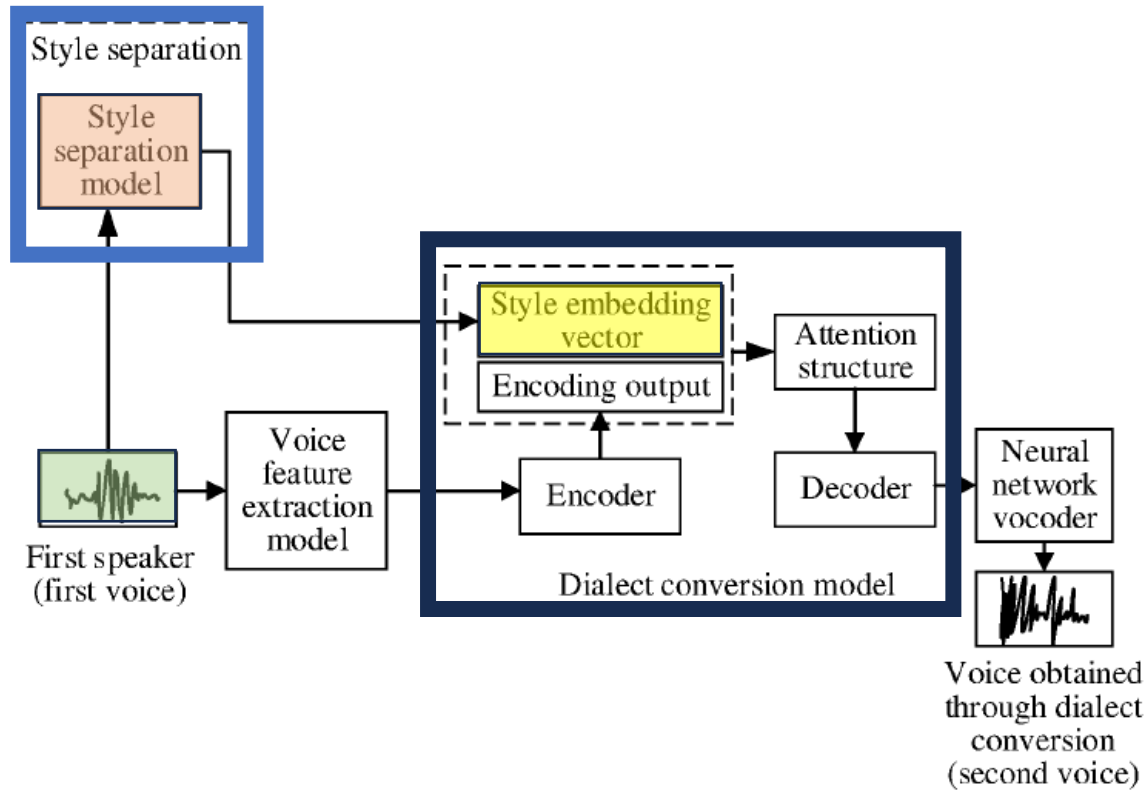


FIG. 9

EX1006-Fan, Fig. 9 (annotated) *see also* Figs. 7, 12. The style separation model extracts features from the target voice to output a style embedding vector (annotated yellow) that represents the timbre of the input (target) voice. EX1006-Fan, [0094], Figs. 7, 9, 12; EX1003, ¶¶113-114 (explaining intervening processing regarding attention in embodiments). *Fan* explicitly states that “timbre” comprises the accent of the voice. EX1006-Fan, [0050] (“The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.”). Thus, *Fan*’s style embedding vector (used in its dialect conversion model) represents the accent of a target speaker.

EX1003, ¶115 *citing* EX1010-Accentron, 21; EX1006-Fan, [0050], [0094], [0127], Figs. 7, 9, 12. For these reasons, *Fan's* style embedding vector teaches a *speaker embedding*.

In the exemplary embodiment corresponding to Figure 9, *Fan* describes linguistic and acoustic features combined with its speaker embedding (style embedding vector, yellow above) to produce a modified output. EX1006-Fan, [0126]-[0129]. Specifically, *Fan* explains the PPG feature (*e.g.*, linguistic content) of the first (input) voice and the style features (*e.g.*, acoustic content) of the first (input) voice are input into the dialect conversion model (annotated with a black box) corresponding to the selection operation, and the second (modified) voice is output by using the dialect (accent) conversion model corresponding to a style embedding vector (speaker embedding). EX1006-Fan, [0127], Fig. 9; EX1003, ¶116.

b) Motivation to Combine Fan's speaker embedding

In the proposed modified method, a PHOSITA would have been motivated and found it obvious to employ the speaker embedding taught by *Fan* as part of the “translation model” corresponding to a particular target accent in *Feinauer* as an operable method of modifying the source voice's timbre to match the timbre of the target, modifying the resulting accent. EX1004-Feinauer, [0081], [0105], Fig. 7; EX1006-Fan, [0010], [0081]-[0082], [0094], [0126]-[0129], [0160]-[0162], Figs. 9,

12-13; EX1003, ¶¶117-118. As discussed, *Feinauer* itself teaches extracting linguistic and acoustic features from a user's input speech. *See*, Sections V.A.3-V.A.4. *Feinauer* further teaches modifying an input, unmodulated audio signal to an output, modulated audio signal by adjusting the accent/dialect of the voice of a user (such as modifying the voice from an American accent to an Australian accent). EX1004-Feinauer, [0046]; EX1003, ¶¶117-118. *Fan* further provides detail as to how a PHOSITA would have implemented the modification, teaching "the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation." EX1006-Fan, [0127]. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan's* modules described functionality in *Fan's* style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. EX1003, ¶120.

A PHOSITA would have combined the teachings such that the method of *Feinauer*, in which accent modification is performed as detailed by *Fan* regarding Fig. 9, in which linguistic content (*e.g.*, comprising PPGs) is combined with acoustic features (*e.g.*, pitch, tone, and/or volume per Claim 1(a)(2)) and a speaker embedding (style embedding vector) to modify the accent (dialect) of the output

audio signal. EX1003, ¶¶120-121. A PHOSITA would have incorporated any necessary processing steps, software, and/or hardware recited by *Fan* into the modified *Feinauer* system to result in an operable system that produces a style embedding vector including a target accent for modifying source speech as taught by *Fan*. *Id.*

A PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer*'s translation model and *Fan*'s style embedding vector target template) according to known data storage and retrieval methods with predictable results, further motivating the combination. *Id.* Because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan*'s software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification. *Id.*

“synthesizing a spectrum representation”

Feinauer teaches “prosodic vocal features and characteristics of the dialects” extracted from the audio samples are converted into a “spectrogram.” EX1004-*Feinauer*, [0095], [0099], [0064]. *Feinauer* already teaches generating a spectrogram representing a signal's frequency spectrum. EX1003, ¶122. *Fan* provides additional detail, teaching producing such a spectrogram specifically of the acoustic features, linguistic features, and style (speaker) embedding vector. EX1006-*Fan*, [0053], [0097]-[0099], [0126]-[0127]. A PHOSITA would have appreciated that in the

Feinauer-Prabhavalkar-Fan modified system, the mapped acoustic features, linguistic features with a reduced accent, and speaker embedding for a human speaker are used to synthesize a mel spectrum representation, as in *Fan*. EX1003, ¶122.

Specifically, as discussed above regarding Fig. 9 and its corresponding embodiment, a linguistic content (*e.g.*, PPG features) of the first (input) voice and the acoustic content (style features) of the first (input) voice are input into a dialect conversion model, and the second (modified) voice is output by using the dialect (accent) conversion model corresponding to a style embedding vector (speaker embedding). EX1006-Fan, [0127], Fig. 9; EX1003, ¶116.

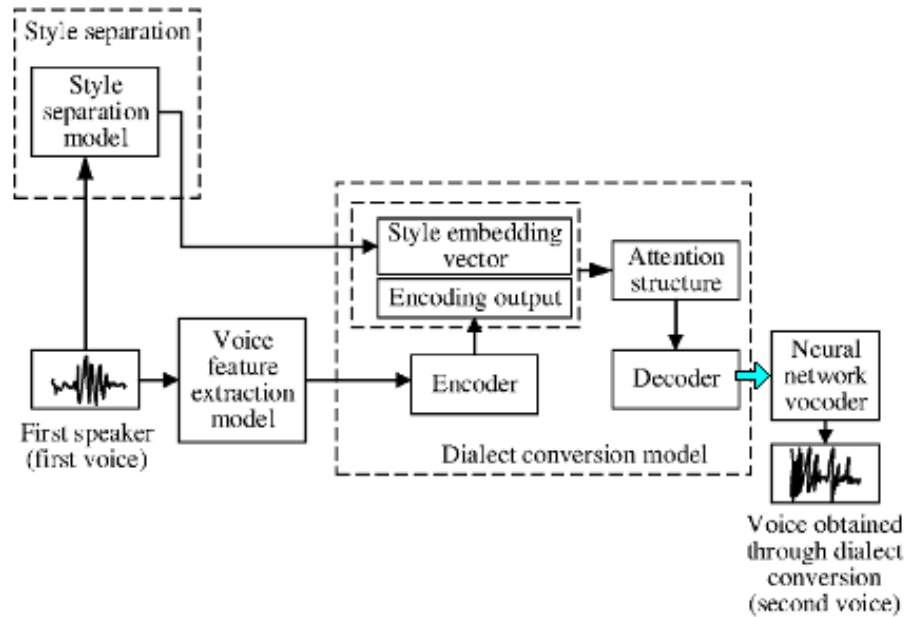


FIG. 9

EX1006-Fan, Fig. 9 (annotated) *see also* Figs. 7, 12. The output of the dialect conversion model, annotated with a blue arrow above, is contemplated by *Fan* as comprising a mel **spectrum** feature. EX1006-Fan, [0099], Fig. 9; EX1003, ¶¶123-124. Specifically, *Fan* explains that a mel spectrum feature is an intermediate representation that “contains information about an output voice” to be input to a vocoder, precisely as shown in Fig. 9. *Id.*

Thus, because a mel spectrum teaches a *spectrum representation*, and because the *Feinauer-Prabhavalkar-Fan* system as combined would generate such mel spectrum representations as output, the modified system teaches *synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker. Id.*

6. Claim 1(c): “generating, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal.”

The '561 Patent describes output chunks, like input chunks, comprising a plurality of frames. EX1001, 8:24-40, 9:10-14. Specifically, the '561 Patent describes output chunks as corresponding in size to that of the input chunk from which they were generated. EX1001, 14:13-18.

Feinauer teaches combining a target accent *speaker embedding*, *acoustic features* such as pitch, and *linguistic features* including spoken words into an output audio signal having a modified accent. EX1004-Feinauer, [0057]. In the *Feinauer-*

Prabhavalkar-Fan method, the output audio signal is generated in *output chunks* based on the spectrum representation by a *vocoder*. EX1005-*Prabhavalkar*, 18:1-18. In the *Feinauer-Prabhavalkar-Fan* method, the output audio signal is generated in *output chunks* comprising a plurality of frames corresponding to the input chunk from which it was generated. EX1005-*Prabhavalkar*, 18:1-18.

a) *Feinauer's* Teachings

Feinauer employs an accent modifier 318 to modify the accent in “portions” of a user’s speech, modified per Claim 1(a)(1) to be *Prabhavalkar's* input chunks. EX1004-*Feinauer*, [0057], Fig. 3; See Sections V.A.2-V.A.3 discussing “chunk-based” processing. Specifically, *Feinauer's* method generates a conversation output audio signal with a modified accent. EX1004-*Feinauer*, [0086], Fig. 5; EX1003, ¶127 (explaining this “broadcast” conversation teaches an output audio signal). Each of steps 512 (Fig. 5), 714 (Fig. 7), and 812 (Fig. 8) provide steps in which *Feinauer's* method generates an output audio signal based on a *speaker embedding*, (*Feinauer's* template, modified to include *Fan's* style embedding vector per Claim 1(b)), *acoustic features* (*Feinauer's* pitch, tone, and/or volume per Claim 1(a)(2)), and *linguistic features* (*Feinauer's* individual words, modified to be represented by *Fan's* PPGs per Claim 1(a)(3)). EX1004-*Feinauer*, [0086], [0110], [0117], Figs. 5-7; EX1003, ¶128. In the combined system, the spectrum representation as discussed in limitation 1(b) is input to a *vocoder*, as seen in *Fan*.

“generating an output chunk by a vocoder”

Prabhavalkar teaches a method of real-time speech processing in which input is processed as chunks, and an output corresponding to each chunk is produced sequentially before processing the next input chunk. EX1005-*Prabhavalkar*, 6:24-28, 18:1-41. Specifically, *Prabhavalkar* teaches its “chunk-based” processing technique consumes all frames in a first input chunk, processing them and outputting results “before it consumes the next chunk.” EX1005-*Prabhavalkar*, 18:20-23, 17:33-51; EX1003, ¶¶129-130. Thus, *Prabhavalkar*’s generates output from a given chunk (an “output chunk”) before moving on to processing succeeding chunks. *Id.* Based on this disclosure, a PHOSITA would have recognized or at least found obvious that *Prabhavalkar*’s “chunk-based approach,” which consumes input as chunks, correspondingly generates output as chunks. *Id.*

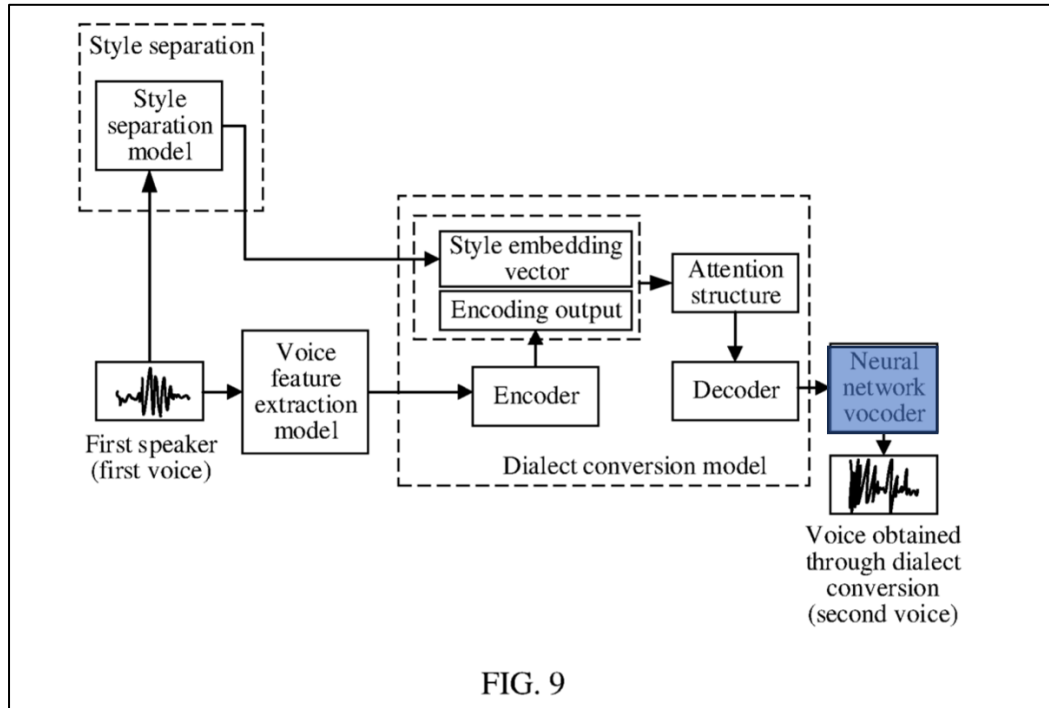
A PHOSITA would have been motivated to configure the *Feinauer-Prabhavalkar-Fan* system to generate its output voice modified with the target accent in “output chunks” as taught by *Prabhavalkar* to enable the increase in speed and performance intended when dividing the speech signal into input chunks as discussed regarding Claim 1(a)(1). EX1003, ¶¶131-132 *citing* EX1005-*Prabhavalkar*, 17:33-51. A MTC for *Prabhavalkar*’s chunk-based approach is provided above regarding Claim 1(a)(1) and is incorporated by reference for this limitation. *See* Section V.A.2. Specifically, regarding output chunks, Mr. Schmandt

explains that audio frames are short (~30ms) and are commonly used as input to signal processing algorithms (e.g., spectra extraction), while speech events such as phonemes or words are longer in duration than a single frame. EX1003, ¶¶131-132. Thus, to analyze speech looking for linguistic events, a PHOSITA would have found it advantageous to analyze a longer series of frames, together in sequence as an output chunk. *Id.*

b) *Fan's* Teachings

Fan teaches using a “neural network vocoder” to “convert an output feature of a voice processing neural network into a highly natural voice.” EX1006-Fan, [0049]. *Fan* teaches a mel³ “feature is converted into a playable voice form by using a neural network vocoder.” EX1006-Fan, [0126].

³ *Fan* discusses the third and fourth “melt” features. A PHOSITA would have recognized that “melt features” is a typographic error throughout *Fan's* specification, intending to say “mel features.” EX1003, ¶135.



EX1006-Fan, Fig. 9 (annotated); EX1003, ¶¶133-134.

c) Motivation to Combine *Fan's* vocoder

In the proposed modified method, A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module such as a vocoder to generate an output chunk based on *Feinauer's* spectrogram-spectrum representation as taught by *Fan*. EX1003, ¶135. A PHOSITA would have appreciated the lengthy detail provided by *Fan* regarding hardware, software, and neural network configurations capable of generating an output signal based on a spectrum representation using a vocoder, leading to a REOS. EX1006-Fan, [0049], [0099], [0126], Figs. 9 and 12; EX1003, ¶135. A PHOSITA would have recognized the modification constitutes applying a known technique (*Feinauer* and *Fan's* steps of

producing an output signal using a vocoder) to a known device (*Fan's* vocoder software and any necessary supporting hardware) with predictable results, further motivating the combination. *Id.* Because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan's* software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification. *Id.*

Claim 2

7. Claim 2: “The method of claim 1, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”

As discussed in Claim 1(b), *Feinauer's* spectrogram in *Feinauer-Prabhavalkar-Fan*, includes the linguistic features with a reduced accent. *Feinauer* teaches the acoustic features include a pitch of the input audio signal and Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal. EX1004-*Feinauer*, [0057], [0099]. *Feinauer* teaches to “convert the audio sample waveforms into a spectrogram of MFCC[s] [mel-frequency cepstral coefficients].” EX1004-*Feinauer*, [0099], [0095]. A PHOSITA would have understood *Feinauer's*

spectrogram of MFCC (mel frequency cepstral coefficients) teaches a melspectrogram. EX1003, ¶136.

Claim 6

1. ***Claim 6(Pre): “A computing system, comprising an input device, a vocoder, memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”***

As discussed in Claim 1(Pre), *Feinauer* teaches a computing device. *See* Section V.A.1. As discussed in Claim 1(c), *Feinauer-Prabhavalkar-Fan* teach a vocoder. *See* Section V.A.6. Further, *Feinauer*’s computing device includes an input device such as a microphone where speech is received and detected. EX1004-*Feinauer*, [0031], [0041]. *Feinauer* teaches a computing system device that includes a memory storing instructions that has “one or more processors, coupled to memory” that are configured to execute the stored instructions. EX1004-*Feinauer*, [0010], [0012], [0031]-[0032]. Thus, *Feinauer* also teaches the claimed *input device*.

2. ***Claim 6(a): “extract, from a chunk of a stream of chunks of an input audio signal obtained via the input device, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”***

As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent. *See* Sections V.A.2-V.A.4. As discussed in Claim 6[Pre], this input audio is obtained via *Feinauer*’s input device.

3. ***Claim 6(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”***

See Section V.A.5.

4. ***Claim 6(c): “and generate, via the vocoder and based on the spectrum representation, an output chunk of an output audio signal.”***

See Section V.A.6.

Claim 7

8. ***Claim 7: “The computing system for real-time correction of claim 6, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”***

See Section V.B.1.

Claim 11

9. ***Claim 11(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”***

As discussed in Claim 1(Pre), *Feinauer* teaches a computing device that includes a non-transitory computer-readable storage medium including instructions.

EX1004-Feinauer, [0031]-[0032], [0043], [0081], [0118]-[0121]; *See also* Section V.A.1.

10. Claim 11(a): “extract, from a chunk of a stream of chunks of an input audio signal, linguistic features and acoustic features, wherein the linguistic features have a reduced accent;”

As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent. *See* Sections V.A.2-V.A.4.

11. Claim 11(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”

See Section V.A.5.

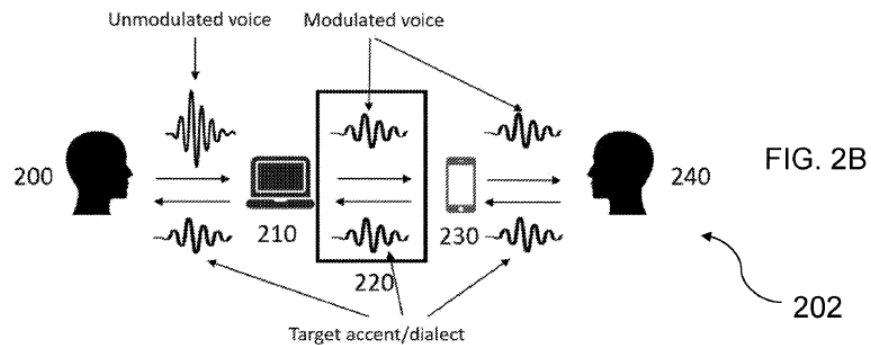
12. Claim 11(c): “generate, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal;”

See Section V.A.6.

13. Claim 11(d): “and provide the output audio signal for output via an auditory channel of an output device.”

Feinauer teaches to provide the output audio signal for output via speakers (auditory channel) of an output device. EX1004-Feinauer, [0014], [0016], [0018], [0041]-[0042]. *Feinauer* teaches “user 200 will have his or her voice modified from a first accent and/or dialect to a second accent and/or dialect” and that participants will communicate “with each other over the communication channel using

modulated audio signals of one or more of their voices in the accent and/or dialect selected and modified.” EX1004-Feinauer, [0043]. *See also* EX1004-Feinauer, [0031]-[0032]. Once the modified modulated signal is sent via *Feinauer’s* communication channel, a user may listen to the signal via “one or more input/output (I/O) devices 126 (e.g., a mouse, a keyboard, a microphone, one or more speakers...” EX1004-Feinauer, [0031]. A PHOSITA would have recognized “speakers” teach an auditory channel. EX1003, ¶146.



EX1004-Feinauer, Fig. 2B.

Claim 12

14. Claim 12: *“The non-transitory computer-readable storage medium of claim 11, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”*

See Section V.B.1.

Claim 16

15. Claim 16(Pre): *“A method for real-time accent correction, the method implemented by a computing system and comprising:”*

See Section V.A.1.

16. Claim 16(a): *“dividing a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”*

Feinauer in view of *Prabhavalkar* teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a 5 predetermined (pre-defined) number of frames of the speech audio signal. See Section V.A.2.

17. Claim 16(b): “extracting, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”

As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent. *Feinauer-Prabhavalkar-Fan* further teach a context associated with the input chunk. See Sections V.A.2-V.A.4.

The '561 Patent explains “context may include a pre-determined number of the frames belonging to chunks preceding the input chunk in the stream of input chunks.” EX1001, 15:11-16.

As discussed regarding Claim 1(a)(1), in the *Feinauer-Prabhavalkar* method, speech audio is divided into *input chunks* comprising a *pre-defined number of frames*. See Section V.A.2 citing EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Further, in the *Feinauer-Prabhavalkar-Fan* method, the style separation model (part of the mapped *acoustic feature module*) uses “attention” to “focus more on different key parts of an input instead of the entire input.” EX1006-Fan, [0048], [0090], [0094]; EX1003, ¶152. *Prabhavalkar* explains that “attention” may be computed using chunks immediately preceding each input chunk. EX1003, ¶152 citing EX1005-*Prabhavalkar*, 6:56-7:11. These preceding chunks (including a pre-determined number of frames) teach *a context associated with the input chunk* in the *Feinauer-Prabhavalkar-Fan* method.

a) *Prabhavalkar's Teachings*

In speech recognition, “attention” refers to a mechanism that allows a neural network model to dynamically focus on the most relevant segments of an audio input. EX1003, ¶153 *citing* EX1006-Fan, [0048], [0097].

Prabhavalkar teaches computing attention based speech frames in the current chunk and one or more previous chunks of the utterance. EX1005-*Prabhavalkar*, 6:56-7:11. In a specific example, *Prabhavalkar* “compute attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:5-11, 18:28-41. Because each chunk, in an embodiment, includes 5 speech frames, the “10 previous chunks” to a particular input chunk would include a pre-determined number (50) of speech frames. EX1005-*Prabhavalkar*, 6:65-7:4, 22:1-38; EX1003, ¶154. Thus, these predetermined number of frames belonging to chunks preceding the input chunk teach *a context associated with the input chunk*. EX1005-*Prabhavalkar*, 6:65-7:4, 17:33-51, 18:1-41, 22:1-38; EX1001, 15:11-16; EX1003, ¶154.

b) *Motivation to Combine Prabhavalkar's Context*

A PHOSITA would have been motivated to extract acoustic features using an attention-based approach as taught by *Feinauer-Prabhavalkar-Fan* by considering the context of the frames of the 10 previous chunks as taught by *Prabhavalkar* EX1003, ¶155 *citing* EX1005-*Prabhavalkar*, 22:1-38, 7:9-11, 17:33-51, 18:1-41. A

PHOSITA would have appreciated *Fan's* discussion of “attention” providing “focus” on “key parts” of an input as an express motivation to apply *Prabhavalkar's* “previous chunks” *context* approach to the system's acoustic feature extraction module, allowing the combined method to focus more on the key *current input chunk*. EX1003, ¶156 *citing* EX1006-*Fan*, [0048], [0097]. A PHOSITA would have recognized the modification constitutes applying a known technique (*Prabhavalkar's* context provided by frames of previous chunks) to a known device (hardware employing a style separation model, mapped as the *acoustic feature extraction module* of *Feinauer-Prabhavalkar-Fan*) with the predictable result of accounting for the context of previous chunks when extracting acoustic features from a particular input chunk. EX1003, ¶157. Because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan's* software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification. *Id.*

18. Claim 16(c): “obtaining a speaker embedding for a human speaker; and”

Feinauer-Prabhavalkar-Fan teaches to obtain a speaker embedding for a human speaker. *See* Section V.A.5.

19. Claim 16(d): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

Feinauer-Prabhavalkar-Fan teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. See Section V.A.6.

Claim 17

20. Claim 17(Pre): “A computing system, comprising memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”

See Section V.C.1.

21. Claim 17(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”

Feinauer in view of *Prabhavalkar* teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a 5 predetermined (pre-defined) number of frames of the speech audio signal. See Section V.A.2.

22. Claim 17(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”

See Sections V.G.2, V.A.2.

23. Claim 17(c): “obtain a speaker embedding for a human speaker; and generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

Feinauer-Prabhavalkar-Fan teaches to obtain a speaker embedding for a human speaker. See Section V.A.5. *Feinauer-Prabhavalkar-Fan* teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. See Section V.A.6.

Claim 18

24. Claim 18(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”

See Section V.E.1.

25. Claim 18(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”

Feinauer in view of *Prabhavalkar* teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a 5 predetermined (pre-defined) number of frames of the speech audio signal. See Section V.A.2.

26. Claim 18(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”

See Sections V.G.2, V.A.2.

27. Claim 18(c): “obtain a speaker embedding for a human speaker; and”

Feinauer-Prabhavalkar-Fan teaches to obtain a speaker embedding for a human speaker. See Section V.A.5.

28. Claim 18(d): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

Feinauer-Prabhavalkar-Fan teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. See Section V.A.6.

VI. GROUND 2: CLAIMS 3, 8, 13 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND SARGSYAN*

Claim 3

1. Claim 3: “The method of claim 1, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”

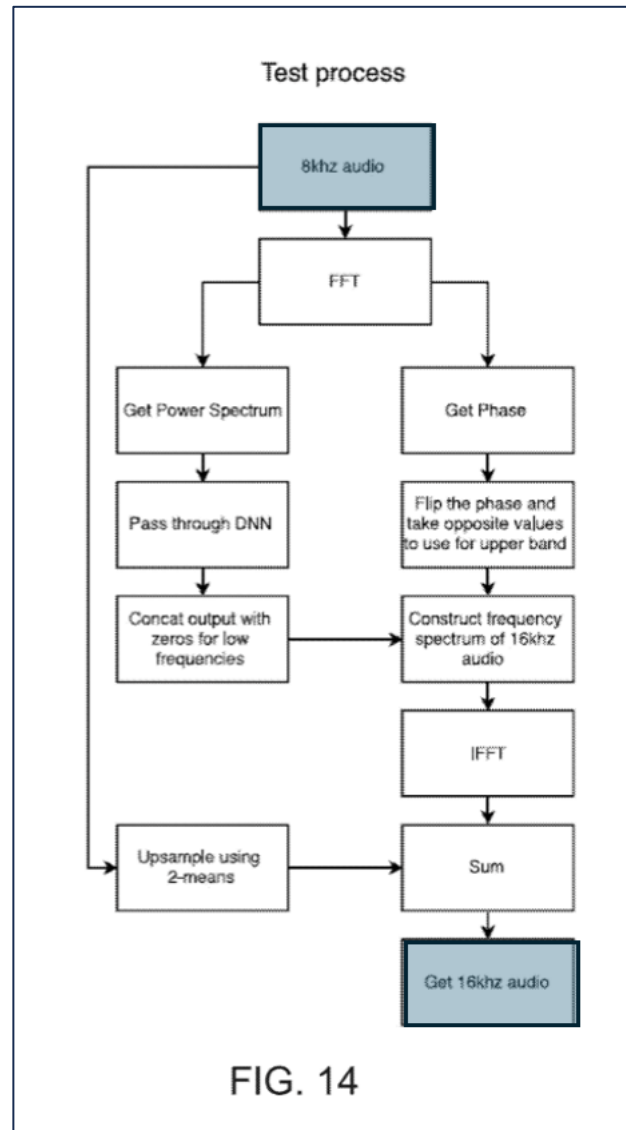
Feinauer’s processor is able to process digital and analog signals. EX1004-*Feinauer*, [0032]. A PHOSITA would have expected that *Feinauer’s* system would digitize the input audio signal a sampling rate to make the signal usable in speech processing. EX1003, ¶¶169-171 citing EX1005-*Prabhavalkar*, 26:57-59.

Feinauer does not specifically discuss its sample rate. However, *Feinauer-Prabhavalkar-Fan* in view of *Sargsyan* teach an input audio signal digitized with a

first sample rate and an output audio signal digitized with a second sample rate different from the first sample rate.

a) *Sargsyan's Teachings*

Sargsyan teaches a speech processing system in which input audio sampled at a first sample rate is converted to a second, different sample rate. EX1007-Sargsyan, [0036]. In a specific example, “8 kHz sampled audio” (*a first sample rate*) “is converted to 16 kHz” (*second sample rate different from the first sample rate*). *Id.* *Sargsyan* teaches sample rate conversion may improve the received output sound quality and/or save network traffic. EX1007-Sargsyan, [0027]; EX1003, ¶172. *Sargsyan's* sample rate conversion is illustrated in Fig. 14 below.



EX1007-Sargsyan, Fig. 14 (annotated).

a) Motivation to Combine *Sargsyan*'s sampling rate

A PHOSITA would have been motivated to employ *Sargsyan*'s method of converting a "8 kHz sampled audio" (*a first sample rate*) "to 16 kHz" (*second sample rate different from the first sample rate*) in *Feinauer*'s real-time accent conversion method. EX1003, ¶173. A PHOSITA would have recognized that applying *Sargsyan*'s sampling techniques to *Feinauer*'s system would have only

required a reprogramming according to known techniques to produce the reasonable result of saving network traffic in *Feinauer's* accent conversion method by expanding low bitrate audio. EX1003, ¶173 *citing* EX1007-Sargsyan, [0027], [0036]. A PHOSITA would have recognized that applying *Sargsyan's* sampling techniques to *Feinauer's* system would have only required a reprogramming according to known techniques to produce the reasonable result of being able to save network traffic in *Feinauer's* accent conversion method when processing a digital signal, resulting in a REOS. EX1003, ¶173. A PHOSITA would have been motivated by *Sargsyan's* expressly-discussed benefits of improved output sound quality and reduced network traffic, further motivating the combination. EX1007-Sargsyan, [0027]; EX1003, ¶173.

Claim 8

2. *Claim 8: "The computing system for real-time correction of claim 6, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate."*

See Section VI.A.1.

Claim 13

1. ***Claim 13: “The non-transitory computer-readable storage medium of claim 11, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”***

See Section VI.A.1.

VII. GROUND 3: CLAIMS 4, 9, AND 14 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND KALINLI*

Claim 4

1. ***Claim 4: “The method of claim 1, further comprising processing, in parallel via at least two processing units of the computing system, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.”***

As discussed regarding limitations 1(a)(2)-1(a)(3), *Feinauer-Prabhavalkar-Fan* teach extracting acoustic features and linguistic features. See Sections V.A.3-V.A.4. Further, *Feinauer* teaches “multiple processors may provide functionality for ***parallel***, simultaneous execution of instructions or for ***parallel***, simultaneous execution of one instruction **on more than one piece of data.”** EX1004-Feinauer, [0032]. However, *Feinauer* does not describe performing such parallel processing *via at least two processing units of the computing system*.

a) ***Kalinli’s Teachings***

Kalinli teaches a speech processing system that extracts features using a plurality of processors. EX1009-Kalinli, [0017] (speech processing), [0021]-[0023]

(feature extraction), [0048]-[0049] (multiple processors). Specifically, *Kalinli* envisions “multiple processors could extract different features from an input sound window in parallel.” EX1009-Kalinli, [0048]. *Kalinli* provides an example in which a parallel speech processing system has 8 “synergistic” processing elements (SPE) operating under a single power processing element (PPE). EX1009-Kalinli, [0048]-[0051], Fig. 4. Different processors are dedicated to extraction of particular features such as intensity or frequency features. EX1009-Kalinli, [0048]-[0049]; EX1003, ¶¶177-179. Thus, *Kalinli* provides express disclosure of performing parallel processing of different speech features *via at least two processing units of the computing system*. As explained below, a PHOSITA would have been motivated to modify the *Feinauer* system according to *Kalinli*’s disclosure to extract the different linguistic and acoustic features using a plurality of parallel processing units.

Thus, in combination, *Feinauer-Prabhavalkar-Fan* in view of *Kalinli* teaches *processing, in parallel via at least two processing units of the computing system, extracting the acoustic features, extracting the linguistic features*.

b) Motivation to Combine

A PHOSITA would have been motivated to modify *Feinauer-Prabhavalkar-Fan* parallel processing teachings to include *Kalini*’s teachings of multiple processors to extract features. EX1003, ¶179. A PHOSITA would have been motivated to apply *Kalini*’s teachings so that one of *Feinauer*’s processor extracts

the acoustic features as discussed in Claim 1(a)(2) and *Feinauer's* second processor extracts the linguistic features as discussed in Claim 1(a)(3). EX1003, ¶180. A PHOSITA would have been motivated to perform the Claim 1(a)(2)-1(a)(3) process in parallel via at least two processing units of the computing system, dedicating particular processors to particular feature extraction as in *Kalinli*, improving the accuracy and speed of the system. EX1003, ¶180 *citing* EX1009-Kalinli, [0039]-[0042]. A PHOSITA would have recognized that applying *Kalinli's* techniques to *Feinauer's* system would have only required software reprogramming according to known techniques to produce the reasonable result of increased synchronization in *Feinauer-Prabhavalkar-Fan* accent conversion method, resulting in a REOS. EX1003, ¶180. The combination constitutes use of a known technique (*Kalinli's* delegation of particular feature extraction responsibilities to particular processors) to improve similar devices (the modified *Feinauer* and *Kalinli* ML speech processing systems) in the same way. *Id.*

Claim 9

2. Claim 9: “The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.”

See Section VII.A.1.

Claim 14

3. Claim 14

See Section VII.A.1.

VIII. GROUND 4: CLAIMS 5, 10, 15 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND WATTS*

Claim 5

1. Claim 5: “The method of claim 1, further comprising: extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, and generating the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and providing, as an input to at least one of the modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.”

The '561 describes a “computational graph” as including a “pipeline.” EX1001-7:43-44, 16:31-34, 17:19-20. Though *Feinauer* envisions parallel processing, it does not teach the recited computational graph time-shifting details of Claim 5. EX1004-Feinauer, [0032]. Further, the '561 describes time-shifts as being “selected to find a trade-off for quality of output feature vs provided delay. Higher shift values (for example, 100 ms) can be selected for linguistic features and lower shift values can be selected for acoustic features (for example, 20 ms) and input for vocoding (for example, 40 ms).” EX1001-17:50-56.

In the combination discussed above, *Fan's* style separation model, voice feature extraction model, and vocoder are combined to *Feinauer* to *extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk* as recited in claim 5. See Sections V.A.3, V.A.4, and V.A.6. Further, each of these modules in the system receives input data for processing and subsequently outputs data, as claimed. The modified *Feinauer* system does not teach computational graphs (pipelines) using time-shifts.

However, *Watts* teaches a “computational pipeline” for phoneme and speech sound classification. EX1008-Watts, 5:7-47. As explained by Mr. Schmandt, computational graphs (pipelines) were well-known and widely-used approaches used in speech processing for improving speed and efficiency by coordinating the timing of the processing. EX1003, ¶185. *Watts* provides details of such timing in a speech processing system. EX1008-Watts, 4:43-47, 7:5-20. Thus, *Watts* teaches the claimed *computational graph*.

Watts teaches details of numbers suggesting that some speech events may be classified at different time scales in a speech processing system. EX1008-Watts, 4:43-47, 7:5-20. Thus, *Watts* teaches the claimed *computational graph*. *Watts* teaches time parameters for acoustic and linguistic feature extraction. *Watts's* phoneme classifier extracts linguistic features on a “150 ms context” EX1008-Watts, 7:5-20. Further, *Watts* teaches extraction of acoustic features such as pitch “on a

much finer time-scale (e.g., 1 ms) so as to be able to track fast formant changes associated with consonant closures.” EX1008-Watts, 4:43-47. Thus, a PHOSITA would have recognized that *Watts* teaches the *time-shift of acoustic features* in its computational graph are *lower* than the *time-shift of linguistic features*. EX1003, ¶186. A PHOSITA would have expected these time-shift parameters would have been provided to *Feinauer-Prabhavalkar-Fan*’s style separation model and voice feature extraction model respectively in the combined system, since those are the components receiving voice audio data for feature extraction. (*providing, as an input to at least one of the modules, a time-shift parameter*) EX1003, ¶¶187-188.

a) Motivation to Combine *Watts*

A PHOSITA would have been motivated to modify *Feinauer-Prabhavalkar-Fan* in view of *Watts* to teach *generating the output chunk* using *Fan*’s respective *modules* (style separation model, voice feature extraction model, and vocoder) of a *computational graph*, as demonstrated by *Watts*. EX1003, ¶189. A PHOSITA would have appreciated that such a modification would have been the use of known technique (*Watts*’s computational graph architecture) to improve similar devices (*Feinauer-Prabhavalkar-Fan*’s accent conversion system) in the same way, thus resulting in increased processing speed and efficiency. EX1003, ¶189.

A PHOSITA would have been further motivated to specifically configure the combined system to utilize *a time-shift parameter indicating a difference in*

timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features as demonstrated by Watts. EX1003, ¶190. A PHOSITA would have appreciated that such a modification would have been the use of known technique (Watts's time shift parameters) to improve similar devices (Feinauer-Prabhavalkar-Fan's accent conversion system) in the same way, thus resulting in increased processing speed and accuracy, allowing Feinauer-Prabhavalkar-Fan's accent conversion system to extract even small nuances of pitch change. EX1003, ¶190. A PHOSITA would have had a reasonable expectation of success in making such a modification because the combination required only reprogramming of Feinauer-Prabhavalkar-Fan's software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶190.

Claim 10

2. Claim 10

See Section VIII.A.1.

Claim 15

3. Claim 15

See Section VIII.A.1.

IX. CONCLUSION

For the foregoing reasons, Petitioner respectfully requests *Post Grant* review of the Challenged Claims.

Respectfully submitted,

ERISE IP, P.A.

BY: /s/ Eric A. Buresh
Eric A. Buresh, Reg. No. 50,394
eric.buresh@eriseip.com
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601

COUNSEL FOR PETITIONER

X. MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)

Real Party-In-Interest

Petitioner is the real party-in-interest. 37 C.F.R. § 42.8(b)(1).

Related Matters

Pursuant to 37 C.F.R. § 42.8(b)(2), Petitioner is aware of the following matters involving the '756 Patent or related patents:

- *Sanas.ai Inc. v. Krisp Technologies, Inc.*, Case No. 3:25-cv-05666-TSH (N.D. Cal.)
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00272 (PTAB), challenging U.S. Patent No. 11,948,550
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00273 (PTAB), challenging U.S. Patent No. 12,125,496
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00274 (PTAB), challenging U.S. Patent No. 12,131,745
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00275 (PTAB), challenging U.S. Patent No. 11,715,457
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, PGR2026-00033 (PTAB), challenging U.S. Patent No. 12,417,756

Lead and Back-Up Counsel

Petitioners provide the following designation and service information for lead and back-up counsel. 37 C.F.R. § 42.8(b)(3) and (b)(4).

Lead Counsel	Back-Up Counsel
Eric A. Buresh (Reg. No. 50,394) eric.buresh@eriseip.com PTAB@eriseip.com <u>Postal and Hand-Delivery Address:</u> ERISE IP, P.A. 7015 College Blvd., Suite 700 Overland Park, Kansas 66211 Telephone: (913) 777-5600 Fax: (913) 777-5601	Chris R. Schmidt (Reg. No. 63,982) chris.schmidt@eriseip.com Nick R. Apel (Reg. No. 84,017) nick.apel@eriseip.com Adam M. Sandwell (Reg. No. 72,484) adam.sandwell@eriseip.com <u>Postal and Hand-Delivery Address:</u> ERISE IP, P.A. 7015 College Blvd., Suite 700 Overland Park, Kansas 66211 Telephone: (913) 777-5600 Fax: (913) 777-5601

37 C.F.R. § 42.8(b)(4) – Service Information

Please address all correspondence to the lead and back-up counsel as shown above. Petitioner consents to electronic service by e-mail at the e-mail addresses provided above.

CLAIM LISTING APPENDIX
U.S. Patent No. 12,412,561 for Claims 1-18

Claim Designation	Claim Language
Claim 1(P)	A method for real-time accent correction, the method implemented by a computing system and comprising:
Claim 1(a)	extracting, from a chunk of a stream of chunks of an input audio signal, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;
Claim 1(b)	synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and
Claim 1(c)	generating, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal.
Claim 2(P)	The method of claim 1, wherein:
Claim 2(a)	the spectrum representation includes a melspectrogram with the reduced accent; and
Claim 2(b)	the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.
Claim 3(P)	The method of claim 1, wherein:
Claim 3(a)	the input audio signal is digitized with a first sample rate; and
Claim 3(b)	the output audio signal is digitized with a second sample rate different from the first sample rate.
Claim 4	The method of claim 1, further comprising processing, in parallel via at least two processing units of the computing system, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.
Claim 5(P)	The method of claim 1, further comprising:

Claim Designation	Claim Language
Claim 5(a)	extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, and generating the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and
Claim 5(b)	providing, as an input to at least one of the modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.
Claim 6(P)	A computing system, comprising an input device, a vocoder, memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:
Claim 6(a)	extract, from a chunk of a stream of chunks of an input audio signal obtained via the input device, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;
Claim 6(b)	synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and
Claim 6(c)	generate, via the vocoder and based on the spectrum representation, an output chunk of an output audio signal.
Claim 7(P)	The computing system for real-time correction of claim 6, wherein:
Claim 7(a)	the spectrum representation includes a melspectrogram with the reduced accent; and
Claim 7(b)	the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.

Claim Designation	Claim Language
Claim 8(P)	The computing system for real-time correction of claim 6, wherein:
Claim 8(a)	the input audio signal is digitized with a first sample rate; and
Claim 8(b)	the output audio signal is digitized with a second sample rate different from the first sample rate.
Claim 9	The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.
Claim 10(P)	The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to:
Claim 10(a)	extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and
Claim 10(b)	provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.
Claim 11(P)	A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:
Claim 11(a)	extract, from a chunk of a stream of chunks of an input audio signal, linguistic features and acoustic features, wherein the linguistic features have a reduced accent;
Claim 11(b)	synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;
Claim 11(c)	generate, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal; and

Claim Designation	Claim Language
Claim 11(d)	provide the output audio signal for output via an auditory channel of an output device.
Claim 12(P)	The non-transitory computer-readable storage medium of claim 11, wherein:
Claim 12(a)	the spectrum representation includes a melspectrogram with the reduced accent; and
Claim 12(b)	the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.
Claim 13(P)	The non-transitory computer-readable storage medium of claim 11, wherein:
Claim 13(a)	the input audio signal is digitized with a first sample rate; and
Claim 13(b)	the output audio signal is digitized with a second sample rate different from the first sample rate.
Claim 14	The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.
Claim 15(P)	The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to:
Claim 15(a)	extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and

Claim Designation	Claim Language
Claim 15(b)	provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.
Claim 16(P)	A method for real-time accent correction, the method implemented by a computing system and comprising:
Claim 16(a)	dividing a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;
Claim 16(b)	extracting, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;
Claim 16(c)	obtaining a speaker embedding for a human speaker; and
Claim 16(d)	generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.
Claim 17(P)	A computing system, comprising memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:
Claim 17(a)	divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;
Claim 17(b)	extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;
Claim 17(c)	obtain a speaker embedding for a human speaker; and
Claim 17(d)	generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.
Claim 18(P)	A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:

Claim Designation	Claim Language
Claim 18(a)	divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;
Claim 18(b)	extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;
Claim 18(c)	obtain a speaker embedding for a human speaker; and
Claim 18(d)	generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.

APPENDIX OF EXHIBITS

Exhibit 1001	U.S. Patent No. 12,412,561 (“’561 Patent”)
Exhibit 1002	Prosecution History of U.S. Patent No. 12,412,561 (“’561 FH”)
Exhibit 1003	Declaration of Chris Schmandt
Exhibit 1004	U.S. Patent Application Publication No. 2020/0193971 to Feinauer et al. (“ <i>Feinauer</i> ”)
Exhibit 1005	U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“ <i>Prabhavalkar</i> ”)
Exhibit 1006	U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>Fan</i> ”)
Exhibit 1007	U.S. Patent Application Publication No. 2020/0066296 to Sargsyan et al. (“ <i>Sargsyan</i> ”)
Exhibit 1008	U.S. Patent No. 7,319,959 to Watts (“ <i>Watts</i> ”)
Exhibit 1009	U.S. Patent Application Publicaiton No. 2012/0116756 to Kalinli (“ <i>Kalinli</i> ”)
Exhibit 1010	<i>Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning</i> to Zhao et al. (“ <i>Accentron</i> ”)
Exhibit 1011	<i>Intentionally left blank</i>
Exhibit 1012	S. Ding, et al., Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition (“ <i>Ding</i> ”)
Exhibit 1013	Priority Chinese Patent Publication No. 202010996501 to Fan et al. and Certified English Translation
Exhibit 1014	Excerpts of the Prosecution History for the U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>FanHistory</i> ”)
Exhibit 1015	Priority PCT 2022/057759 to Fan et al. and Certified English Translation
Exhibit 1016	Munford Librarian Declaration
Exhibit 1017	C. Xydeas, Differential Encoding Techniques Applied to Speech Signals, 1978 (“ <i>Xydeas</i> ”)
Exhibit 1018	U.S. Patent Application Publication No. 2013/0246064 to Wasserblat et al. (“ <i>Wasserblat</i> ”)
Exhibit 1019	U.S. Patent Application Publication No. 2021/0272571 to Balasubramaniam et al. (“ <i>Balasubramaniam</i> ”)
Exhibit 1020	U.S. Patent Application Publication No. 2020/0294492 to Sypniewski et al. (“ <i>Sypniewski</i> ”)
Exhibit 1021	C. Schmandt, <i>Voice Communication with Computers</i> , Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)

Exhibit 1022	C. O’Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallaghan.com/research/papers/Pitch.pdf) September 30, 2015, (“O’Callaghan”)
Exhibit 1023	B. Sisman et al., <i>Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion</i> , IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, June 2019, 1085-1097 (“Sisman”)
Exhibit 1024	J. Holmes, <i>Speech Synthesis and Recognition</i> , Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003 (“Holmes”)
Exhibit 1025	H. Fletcher, Loudness, Its Definition, Measurement and Calculation, Bell Telephone Laboratories, October 1933 (“Fletcher”)
Exhibit 1026	S. Ding et al., Personal VAD: Speaker-Conditioned Voice Activity Detection, arXiv, April 8, 2020 (“Chang”)
Exhibit 1027	C. Wiltshire, <i>Chapter 13 English Sounds in Context: The Pronunciation of Phonemes and Morphemes</i> , in <i>Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners</i> , 303-324, 2019 (“Wiltshire”)
Exhibit 1028	N. Kazanina, et al., <i>Phonemes: Lexical access and beyond</i> , Psychon Bull Rev 25, 560–585, September 5, 2017 (“Kazanina”)
Exhibit 1029	D. Jurafsky et al., <i>Speech and Language Processing</i> , Chapter 27, October 16, 2019 (“Jurafsky”)
Exhibit 1030	L. Sun, <i>Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training</i> , IEEE on Multimedia and Expo, August 29, 2016 (“Sun”)
Exhibit 1031	G. Zhao et. al, <i>Converting Foreign Accent Speech Without a Reference</i> , IEEE/ACM Transactions On Audio, Speech, and Language Processing, Vol. 29, 2021, 2367-2381 (“Zhao”)
Exhibit 1032	U.S. Patent Application Publication No. 2022/0122579 to Biadsy et al. (“Biadsy”)
Exhibit 1033	U.S. Patent Appl. No. 2018/0174595 to Dirac et al. (“Dirac”)
Exhibit 1034	L. Roberts, <i>Understanding the Mel Spectrogram</i> , (https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53), March 5, 2020 (“Roberts”)
Exhibit 1035	N. Umeda, <i>Consonant duration in American English</i> , The Journal of the Acoustical Society of America, March 1, 1977 (“Umeda”)
Exhibit 1036	U.S. Patent Application No. 2010/0274554 to Orr et al. (“Orr”)

Exhibit 1037	U.S. Patent No. 11,410,658 to Ma et al. (“ <i>Ma</i> ”)
Exhibit 1038	U.S. Patent Application No. 2019/0392798 to Danjyo et al. (“ <i>Danjyo</i> ”)

CERTIFICATION OF WORD COUNT

The undersigned certifies pursuant to 37 C.F.R. §42.24 that the foregoing Petition for *Post Grant* Review, excluding any table of contents, mandatory notices under 37 C.F.R. §42.8, certificates of service or word count, or appendix of exhibits, contains 9567 words according to the word-processing program used to prepare this document (Microsoft Word).

Dated: March 18, 2026

ERISE IP, P.A.

BY: /s/ Eric A. Buresh
Eric A. Buresh, Reg. No. 50,394
eric.buresh@eriseip.com
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601

COUNSEL FOR PETITIONER

**CERTIFICATE OF SERVICE ON PATENT OWNER
UNDER 37 C.F.R. § 42.105**

Pursuant to 37 C.F.R. §§ 42.6(e) and 42.105(b), the undersigned certifies that on March 18, 2026, a complete and entire copy of this Petition for *Post Grant* Review and Exhibits were provided via Federal Express to the Patent Owner by serving the correspondence address of record for the '561 Patent:

Troutman Pepper Locke LLP
80 Linden Oaks
Suite 320
Rochester, NY 14625

Further, a courtesy copy of this Petition for *Post Grant* Review was sent via email to Patent Owner's litigation counsel:

Michael Ng (Michael.Ng@kobrekim.com)
Daniel Zabeer (Daniel.Zaheer@kobrekim.com)
Jessica Fender (Jessica.Fender@kobrekim.com)
Victoria Fordin (Victoria.Fordin@kobrekim.com)
Zachary Ritz (Zachary.Ritz@kobrekim.com)

Respectfully submitted,

ERISE IP, P.A.

BY: /s/ Eric A. Buresh
Eric A. Buresh, Reg. No. 50,394
eric.buresh@eriseip.com
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601

COUNSEL FOR PETITIONER