

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.,
Petitioner

v.

Sanas.ai Inc.,
Patent Owner

Inter Partes Review Case No. IPR2026-00274
U.S. Patent No. 12,131,745

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 12,131,745**

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
II.	SUMMARY OF THE '745 PATENT.....	1
A.	DESCRIPTION OF THE '745 PATENT	1
B.	SUMMARY OF THE PROSECUTION HISTORY	1
C.	LEVEL OF SKILL OF A PERSON HAVING ORDINARY SKILL IN THE ART.....	2
III.	REQUIREMENTS UNDER 37 C.F.R. § 42.104	2
A.	GROUND FOR STANDING UNDER 37 C.F.R. § 42.104(A).....	2
B.	IDENTIFICATION OF CHALLENGE UNDER 37 C.F.R. § 42.104(B)	2
C.	CLAIM CONSTRUCTION UNDER 37 C.F.R. § 42.104(B)(3).....	3
IV.	SHOWING OF ANALOGOUS, PRIOR ART	3
A.	OVERVIEW OF <i>KESHET</i>	4
B.	OVERVIEW OF <i>KING</i>	5
C.	OVERVIEW OF <i>CUTURI</i>	6
D.	OVERVIEW OF <i>KAMEOKA</i>	6
E.	OVERVIEW OF <i>FAN</i>	7
F.	OVERVIEW OF <i>VYAS</i>	9
G.	OVERVIEW OF <i>CHEN</i>	9
V.	GROUND 1: CLAIMS 1-2, 4, 8-9, 11, AND 15-17 ARE OBVIOUS OVER KESHET, KING, AND CUTURI.....	10
A.	CLAIM 1	10
1.	<i>Claim 1(Pre): “An accent conversion system, comprising an audio interface, memory having instructions stored thereon, and one or more processors coupled to the memory and configured to execute the instructions to:”</i>	<i>10</i>
2.	<i>Claim 1(a): “obtain input audio data via the audio interface;”</i>	<i>11</i>
3.	<i>Claim 1(b): “generate from the input audio data first phonetic embedding vectors for phonetic content representing a source accent;”</i>	<i>12</i>
4.	<i>Claim 1(c): “apply a trained accent conversion neural network to the first phonetic embedding vectors to generate second phonetic embedding vectors</i>	

	<i>corresponding to first phonetic characteristics of speech data in a target accent;”</i>	15
5.	<i>Claim 1(d): “determine a differentiable alignment by jointly maximizing a cosine distance between the first phonetic embedding vectors and the second phonetic embedding vectors; and”</i>	18
	a) <i>Keshet’s Teachings</i>	19
	b) <i>King’s Teachings</i>	19
	c) <i>Motivation to Combine Keshet and King</i>	21
	d) <i>Cuturi’s Teachings</i>	23
	e) <i>Motivation to Modify Keshet-King to Utilize Cuturi’s Differentiable Alignment</i>	24
6.	<i>Claim 1(e): “align the speech data to the phonetic content based on the differentiable alignment to generate and provide output audio data corresponding to the aligned speech data and representing the target accent.”</i>	25
B.	CLAIM 2	27
	1. <i>Claim 2: “...wherein the first phonetic embedding vectors represent second phonetic characteristics of input speech in the input audio data in a numerical format and encode ... phonemes...present in the input speech.”</i>	27
C.	CLAIM 4	27
	1. <i>Claim 4: “...in order to determine the cosine distance... normalize the first and second phonetic embedding vectors by scaling the first and second phonetic embedding vectors to have a magnitude of one and preserving a relative direction of the first and second phonetic embedding vectors; and ...generate a dot product of the normalized first and second phonetic embedding vectors.”</i>	27
D.	CLAIM 8	28
	1. <i>Claim 8(Pre)</i>	28
	2. <i>Claim 8(a)</i>	28
	3. <i>Claim 8(b)</i>	29
	4. <i>Claim 8(c)</i>	29
	5. <i>Claim 8(d)</i>	29
E.	CLAIM 9	29
	1. <i>Claim 9</i>	29
F.	CLAIM 11	29

1.	<i>Claim 11</i>	29
G.	CLAIM 15	29
1.	<i>Claim 15(Pre)</i>	29
2.	<i>Claim 15(a)</i>	29
3.	<i>Claim 15(b)</i>	29
4.	<i>Claim 15(c)</i>	30
5.	<i>Claim 15(d)</i>	30
H.	CLAIM 16	30
1.	<i>Claim 16</i>	30
I.	CLAIM 17	30
1.	<i>Claim 17</i>	30
VI.	GROUND 2: CLAIMS 3 AND 10 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND KAMEOKA.....	31
A.	CLAIM 3	31
1.	<i>Claim 3(a): “...wherein the accent conversion neural network is trained to learn a mapping between the first phonetic embedding vectors and the second phonetic embedding vectors using a labeled dataset comprising paired samples of [source] accent phonetic embedding vectors and corresponding target accent phonetic embedding vectors; and comprises an encoder layer configured to encode the first phonetic embedding vectors into a latent representation and a decoder layer configured to decode the latent representation to generate the second phonetic embedding vectors.”</i>	<i>31</i>
a)	<i>Kameoka’s Teachings</i>	<i>32</i>
b)	<i>Motivation to Combine Kameoka</i>	<i>34</i>
B.	CLAIM 10	35
1.	<i>Claim 10(a)</i>	35
2.	<i>Claim 10(b)</i>	36
VII.	GROUND 3: CLAIMS 5, 12, AND 18 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND FAN	36
A.	CLAIM 5	36
1.	<i>Claim 5: “...in order to generate the output audio data, one or more of align first frames of the speech data with corresponding second frames of the phonetic content, apply one or more of prosody modeling, intonation adjustment, or accent-specific acoustic modeling</i>	

	<i>techniques, or adjust a speech rate, pitch, or gender, wherein the output audio data preserves linguistic content of the input audio data.”</i>	36
	a) <i>Fan’s Teachings</i>	36
	b) <i>Motivation to Combine Fan</i>	37
B.	CLAIM 12	38
	1. <i>Claim 12</i>	38
C.	CLAIM 18	38
	1. <i>Claim 18</i>	38
VIII.	GROUND 4: CLAIMS 6, 13, AND 19 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND CHEN.....	38
A.	CLAIM 6	38
	1. <i>Claim 6: “...apply a gradient-based optimization algorithm to optimize the joint maximization of the cosine distance.”</i>	38
	a) <i>Motivation to Combine Chen</i>	39
B.	CLAIM 13	40
	1. <i>Claim 13</i>	40
C.	CLAIM 19	40
	1. <i>Claim 19</i>	40
IX.	GROUND 5: CLAIMS 7, 14, AND 20 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND VYAS	40
A.	CLAIM 7	40
	1. <i>Claim 7: “...apply one or more dimensionality reduction techniques to preprocess the first and second phonetic embedding vectors.”</i>	40
	a) <i>Motivation to Combine Vyas</i>	41
B.	CLAIM 14	42
	1. <i>Claim 14</i>	42
C.	CLAIM 20	42
	1. <i>Claim 20</i>	42
X.	CONCLUSION.....	42
XI.	MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)	43
A.	REAL PARTY-IN-INTEREST	43
B.	RELATED MATTERS	43
C.	LEAD AND BACK-UP COUNSEL	44
D.	37 C.F.R. § 42.8(B)(4) – SERVICE INFORMATION	44

TABLE OF AUTHORITIES

CASES

Phillips v. AWH Corp., 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc)	3
---	---

STATUTES

35 U.S.C. § 102(a)(1)	passim
35 U.S.C. § 102(a)(2)	passim
35 U.S.C. § 103	2
35 U.S.C. §102(a).....	5

I. INTRODUCTION

Krisp Technologies, Inc. (“Petitioner”), respectfully requests inter partes review (“IPR”) of claims 1-20 (“Challenged Claims”) of U.S. Patent No. 12,131,745 (“’745 Patent”)

II. SUMMARY OF THE ’745 PATENT

A. Description of the ’745 Patent

The ’745 Patent describes a computer-implemented system that converts a speaker’s voice from a source accent to a target accent in real time using machine learning models operating on phonetic embeddings of the input speech signal. EX1001, Abstract, 1:11-14. The system aligns phonetic content between source and target accents by computing and optimizing differentiable alignment (via techniques such as cosine distance) between their respective phonetic embedding vectors to achieve stable, high-quality accent conversion. EX1001, 8:46-55, 10:24-42, Fig. 3.

B. Summary of the Prosecution History

The ’745 Patent faced no office actions and received a Notice of Allowance September 25, 2024. EX1002. The ’745 claims priority to Provisional Application No. 63/510,487 filed June 27, 2023. Because the ’745 claims the benefit of a filing date after March 16, 2013, Petitioner applies AIA versions of §§ 102, 103. For this

Proceeding, Petitioner applies June 27, 2023, as the priority date for the Challenged Claims.¹

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

A PHOSITA at the time of the '745 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, ¶26.

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

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

Petitioner certifies the '745 Patent is eligible for IPR.

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

Petitioner requests the Challenged Claims be found unpatentable on the following grounds under 35 U.S.C. § 103:

Proposed Grounds of Unpatentability	Exhibits
<u>Ground 1:</u> Claims 1-2, 4, 8-9, 11, and 15-17 are obvious over <i>Keshet, King, and Cuturi</i> .	1004, 1005, 1006
<u>Ground 2:</u> Claims 3 and 10 are obvious over <i>Keshet, King, Cuturi, and Kameoka</i> .	1004, 1005, 1006, 1007

¹ Petitioner does not concede any of the Challenged Claims are entitled to the priority of June 27, 2023.

<u>Ground 3:</u> Claims 5, 12, and 18 are obvious over <i>Keshet, King, Cuturi, and Fan</i> .	1004, 1005, 1006, 1008
<u>Ground 4:</u> Claims 6, 13, and 19 are obvious over <i>Keshet, King, Cuturi, and Chen</i> .	1004, 1005, 1006, 1010
<u>Ground 5:</u> Claims 7, 14, and 20 are obvious over <i>Keshet, King, Cuturi, and Vyas</i> .	1004, 1005, 1006, 1009

C. 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 '745 Patent. The prior art is from the same field of endeavor of the '745 Patent, namely speech audio processing. EX1001, Abstract, 1:11-14, 2:51-55, 4:44-5:19; EX1004-Keshet, Abstract, 2, 38; EX1005-King, 1-2, 23-27, 32-38, 41, Fig. 2; EX1006-Cuturi, 2 (Soft-DTW), 6 (3.4 Multistep-ahead prediction), 18-22 (application to phonemes); EX1007-Kameoka, 1:25-53, 2:26-3:52, 6:7-60, 10:35-50, Fig. 4; EX1008-Fan, Abstract, [0134]-[0138], [0201]; EX1009-Vyas, Abstract, 5:15-28, 9:60-67, 11:50-66; EX1010-Chen, 1:5-7, 9:35-10:27, 10:58-11:5, Fig. 4; EX1003, ¶¶56-57.

The prior art is also reasonably pertinent to at least one problem facing the '745 Patent inventors, namely at least one of:

(A) speech modification: EX1001, Title, Abstract, 1:16-19, 10:66-11:10, Fig. 3; EX1004-Keshet, Abstract, 2, 69, 151-153, 159-160, Fig. 7; EX1007-Kameoka, 1:25-43, 2:26-3:52, 6:7-60; EX1008-Fan, Abstract, 3-7, 120, 142, Figs. 7-9;

(B) vector alignment; EX1001, 1:26-48, 2:47-65, 10:24-32, Fig. 3; EX1004-Keshet, 81, 100-101, 114, 155; EX1005-King, Abstract, 1-3, 13, 36-38, 41, Figs. 2-3; EX1006-Cuturi, 2 (Soft-DTW), 2-3 (The DTW and soft-DTW loss functions), 3-4 (Differentiation of soft-DTW); EX1007-Kameoka, 1:25-53, 6:5-57;

(C) efficient and accurate processing: EX1001, 2:47-55, 3:1-17, 7:40-50, 10:10-23, 10:58-11:5; EX1004-Keshet, 69, 95, 165; EX1006-Cuturi, 2 (Soft-DTW), 3 (The case...), 7 (4.3 Time-Series Classification experiments), 8 (Conclusion); EX1007-Kameoka, 1:25-65, 7:1-4; EX1008-Fan, 7, 86; EX1009-Vyas, 1:5-47, 3:14-23, 9:3-12; EX1010-Chen, 1:5-7, 10:58-11:5; EX1003, ¶58-59.

A. Overview of *Keshet*

U.S. Patent Application Publication No. 2024/0304200 to Keshet et al. ("*Keshet*") was published on September 12, 2024, and originates from a PCT

application filed on February 8, 2022. (“*Keshet*”) was published on September 12, 2024, and originates from a PCT application filed on February 8, 2022. *Keshet* claims priority to provisional application No. 63/146,782 filed on February 8, 2021. *Keshet* is thus prior art to the ’745 Patent under at least 35 U.S.C. §102(a) AIA. *Keshet* was not cited or considered during the prosecution of the ’745 Patent. *See generally* EX1001, EX1002.

Keshet teaches systems and methods for modifying speech by identifying and replacing misspoken phonemes. EX1004-Keshet, Abstract, 1, 59. *Keshet* expressly envisions that its system may utilize an embedding vector to account for the “specific accent” of the speaker to be modified. EX1004-Keshet, 38.

B. Overview of *King*

U.S. Patent Application Publication No. 2014/0148933 to King et al. to King et al. (“*King*”) was filed on November 29, 2012, and published on May 29, 2014. *King* is thus prior art to the ’745 Patent under 35 U.S.C. § 102(a)(1)-(2) AIA. *King* was not cited or considered during the prosecution of the ’745 Patent. *See generally* EX1001, EX1002.

King teaches feature alignment techniques used in speech processing. EX1005-King, Abstract, 1-2, 13. *King* utilizes a time warping module to align features by determining the cosine distance between vectors. EX1005-King, 36-38, 41-43.

C. Overview of *Cuturi*

“Soft-DTW: a Differentiable Loss Function for Time-Series” by Cuturi et al. (“*Cuturi*”) was at least publicly available at least by November 8, 2021, and thus qualifies as prior art to the ’745 Patent under at least 35 U.S.C. § 102(a)(1) AIA. EX1016-Munford, 48-58 (declaration of June Munford establishing public availability of *Cuturi*). *Cuturi* was not cited or considered during the prosecution of the ’745 Patent. See generally EX1001, EX1002.

Cuturi teaches techniques for differentiable alignment of processed speech vectors. EX1006-Cuturi, 1 (Abstract, Introduction). *Cuturi* discusses the increase in performance resulting from use of differentiable alignment. EX1006-Cuturi, 2 (lhc).

D. Overview of *Kameoka*

U.S. Patent No. 11,450,332 to Kameoka et al. to Kameoka et al. (“*Kameoka*”) originated from a PCT application filed on February 20, 2019, and was published September 20, 2022. *Kameoka* thus qualifies as prior art to the ’745 Patent under 35 U.S.C. § 102(a)(1)-(2) AIA. *Kameoka* was not cited or considered during the prosecution of the ’745 Patent. See generally EX1001, EX1002.

Kameoka teaches a voice conversion learning system, method, and program that learns a conversion function which converts a voice to a desired attribution. EX1007-Kameoka, 1:17-21, 2:31-35. *Kameoka*’s system utilizes a Convolutional Neural Network (CNN) comprising encoder and decoder layers to convert a time-

dependent input feature vector representing input speech to an output feature vector with the desired attribution. EX1007-Kameoka, Fig. 2, 9:45-67.

E. Overview of *Fan*

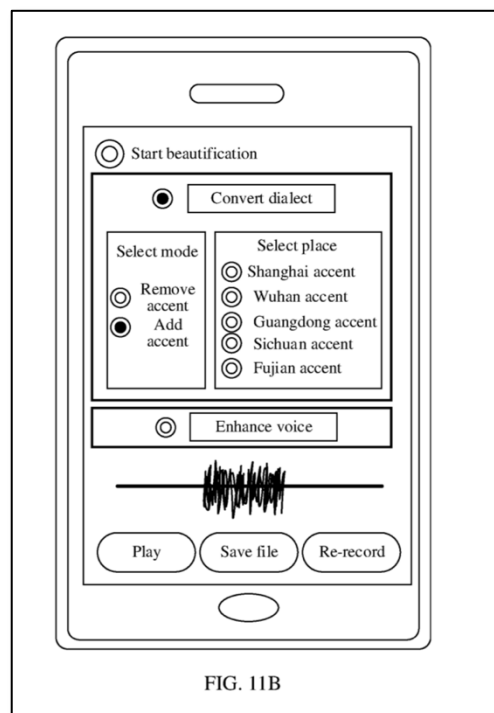
U.S. Patent Application Publication No. 2023/0223006 to Fan et. al. (“*Fan*”)² was filed March 20, 2023, published July 13, 2023, and is a continuation of PCT/CN2021/117945 (EX1015), which was filed on September 13, 2021, and claims priority to Chinese application 202010996501.5 (EX1013) filed September 21, 2020. *See* EX1014-*FanHistory*.

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 PCT 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, ¶¶71-76.

² EX1008 is the U.S. Patent Appl. to Fan. Also included are the Chinese priority application (EX1013, 64), English translation of the Chinese priority application (EX1013, 1), and a certification that the translation is true and correct (EX1013, 63). Also included are the PCT priority application (EX1015), English translation of the PCT priority application (EX1015), and a certification that the translation is true and correct (EX1015, 63). *Fan* made a foreign priority claim within 12 months of the foreign priority application (EX1014, 3-4) and the Office acknowledged same (EX1014, 10).

<i>Fan</i> Publication (EX1008) (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. 11B	58, Fig. 11B	50, Fig. 11B
[0047]	9, ¶2	9, ¶2
[0113]	21, ¶1	20, ¶6
[0120]	22, ¶4	22, ¶3
[0127]	23, ¶5	23, ¶3
[0134]	25, ¶2	25, ¶1
[0135]	25, ¶3	25, ¶2
[0201]	35, ¶3	35, ¶3

Fan teaches “a dialect conversion mode, for adding an accent to or removing an accent from the first voice.” EX1008-Fan, Abstract; EX1003, ¶72.



EX1008-Fan, Fig. 11B; EX1003, 71-76.

F. Overview of *Vyas*

U.S. Patent No. 12,400,078 to Vyas et al. (“*Vyas*”) was filed on March 28, 2022, and claims priority to provisional application No. 63/210,717 filed on June 15, 2021. *Vyas* is thus prior art to the ’745 Patent under at least 35 U.S.C. § 102(a)(1) AIA. *Vyas* was not cited or considered during the prosecution of the ’745 Patent. *See* generally EX1001, EX1002.

Vyas teaches speech preprocessing techniques including a PCA process for dimensionality reduction. EX1009-Vyas, 8:20-33, 5:15-28, 2:29-46. *Vyas* expressly teaches that its techniques generate a latent embedding representing a data sample of speech. *Id.*

G. Overview of *Chen*

U.S. Patent No. 12,217,761 to Chen et al. (“*Chen*”) was filed on October 31, 2021, and was published March 30, 2023. *Chen* is thus prior art to the ’745 Patent under 35 U.S.C. § 102(a)(1)-(2) AIA. *Chen* was not cited or considered during the prosecution of the ’745 Patent. *See* generally EX1001, EX1002.

Chen teaches an ML similarity model used to compute the cosine similarity between frames of audio data comprising parameters such as internal weights. EX1010-Chen, 13:42-60 *see also* 10:18-27, 11:34-47, 14:29-45, 22:37-48. These weights “may be learned through training the similarity model on one or more

training samples, which may comprise updating the parameters based on gradient-based optimization. *Id.*

V. GROUND 1: CLAIMS 1-2, 4, 8-9, 11, AND 15-17 ARE OBVIOUS OVER KESHET, KING, AND CUTURI

A. Claim 1

1. ***Claim 1(Pre): “An accent conversion system, comprising an audio interface, memory having instructions stored thereon, and one or more processors coupled to the memory and configured to execute the instructions to:”***

To the extent the preamble is limiting, *Keshet* teaches an *accent conversion system*³ and *method for automatic alignment of phonetic content for real-time accent conversion* that receives and segments speech to modify selected phoneme segments using a machine learning model. EX1004-Keshet, 38, Figs. 5-6.

Specifically, *Keshet* envisions that the embedding vector may represent a “relevant phoneme of interest P* in a context of a specific linguistic parameters, such as a specific language, a specific dialect, a specific *accent*.” EX1004-Keshet,

³ 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.

159. Notably, the '745 Patent itself envisions its accent conversion system may be used to help speakers learn to improve their pronunciation and intonation, facilitating accent acquisition.” EX1001, 10:66-11:10.

Keshet's system includes memory storing instructions coupled to processors, as well as a “user interface 1120 comprises a recording module which may be configured for digitally recording voice samples,” teaching an *audio interface*. EX1003, ¶¶82-83 *citing* EX1004-Keshet, 72-74, Figs. 1-2.

2. Claim 1(a): “obtain input audio data via the audio interface;”

Keshet obtains (captures) a user's speech via a recording module that “conveys the captured speech signals to the patient application, which analyzes the sounds, words, and sentences to identify speech particles, phonemes and words.” EX1004-Keshet, 73-74, Figs. 1-2 (Step 2200); EX1003, ¶84. *Keshet's* audio interface comprising recording module 110 is highlighted orange in Fig. 6 below.

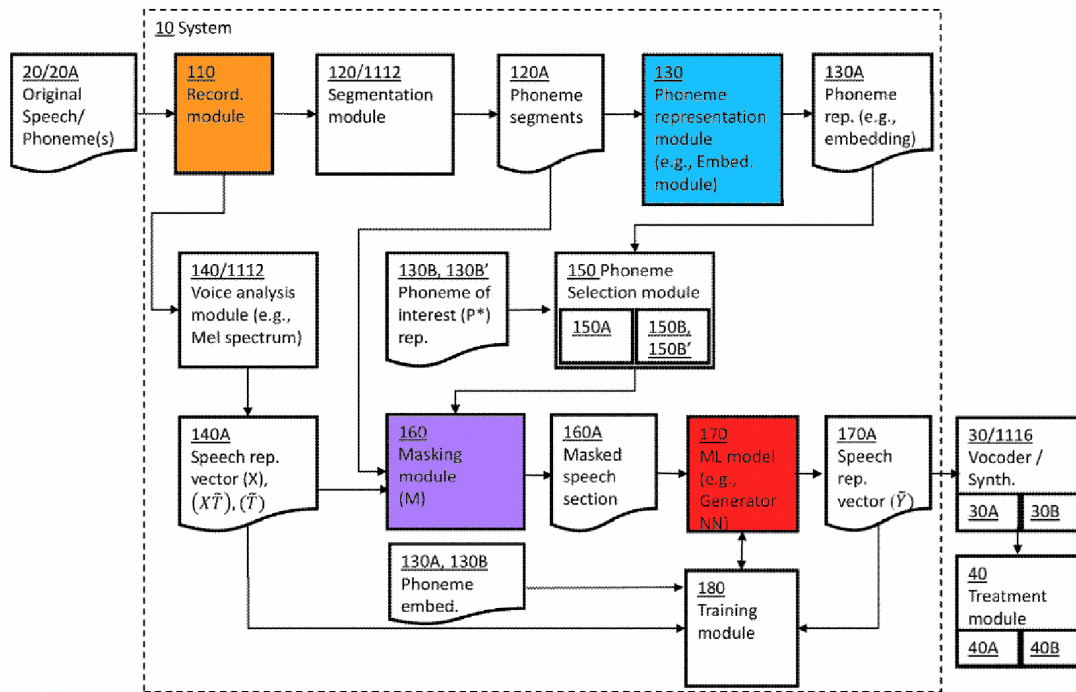


FIG. 6

EX1004-Keshet, Fig. 6 (annotated); EX1003, ¶84.

3. ***Claim 1(b): “generate from the input audio data first phonetic embedding vectors for phonetic content representing a source accent;”***

Keshet’s system analyzes speech received by a speaker “to generate a representation and/or ***embedding of voice characteristics*** of the speaker.” EX1004-Keshet, 60, 18-19, 114-119. As illustrated in Fig. 6, the system includes a “segmentation module” 120 for segmenting the input audio data and a “phoneme embedding (or “representation”) module” 130 for calculating an acoustic phoneme embedding vector from the segmented speech, highlighted blue below.

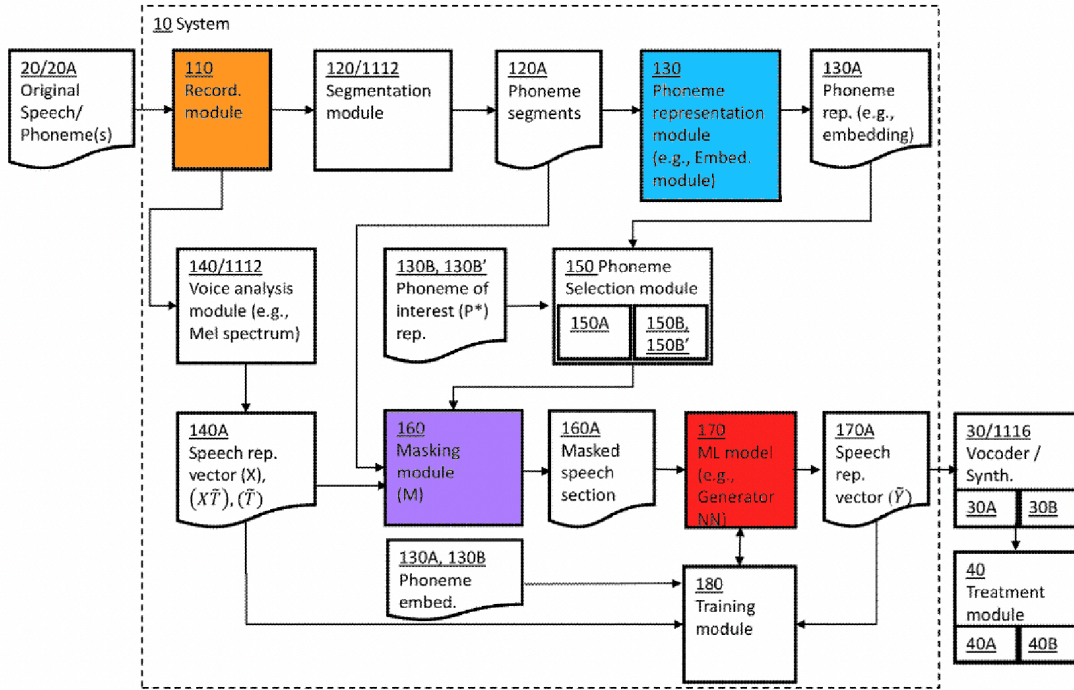


FIG. 6

EX1004-Keshet, Fig. 6 (annotated), 114-119; EX1003, ¶85. *Keshet* specifically teaches input speech is analyzed by module 130 to determine a set of phonemes $p = (p_1, \dots, p_k)$ that represent the input speech. EX1004-Keshet, 77-85, 107. *Keshet* then teaches generating, from the set of phonemes p , a set of vectors $\tilde{x}_k = (x_1, \dots, x_T)$, which are audio features representing the string of phonemes p_k .

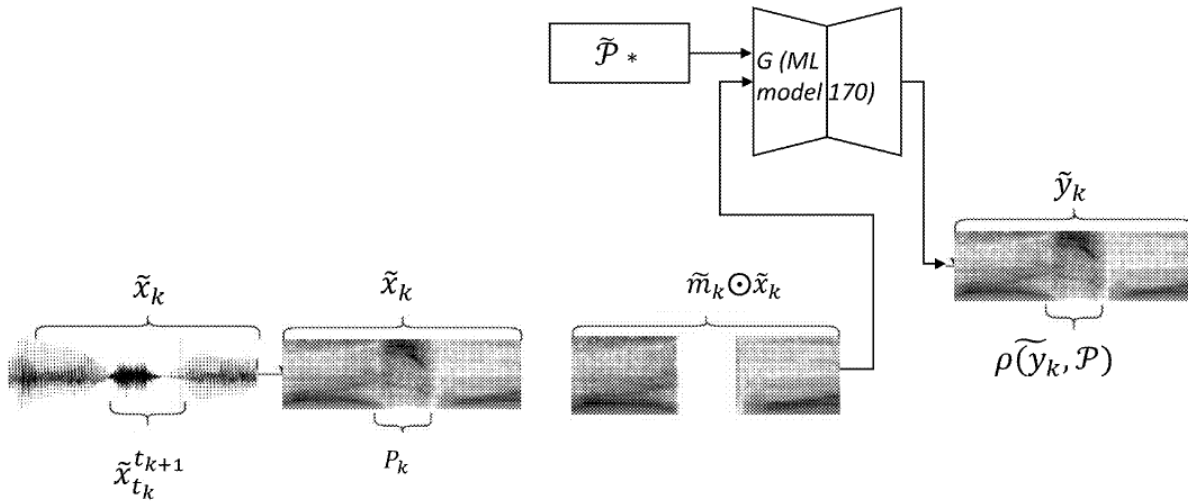


FIG. 5

EX1004-Keshet, Fig. 5. In embodiments, the audio features represented as elements $(x_1, \dots, x_T)$ of $\tilde{x}_k$ are denoted as a set of “D-dimensional feature *vectors*” that “may include or may represent voice characteristics such as frequency bins of the Mel-spectrum or mel-cepstrum of the recorded speech of the speaker.” EX1004-Keshet, 81-82. Thus, Keshet’s vectors $(x_1, \dots, x_T)$ of $\tilde{x}_k$ teach the claimed *phonetic embedding vectors for phonetic content* generated from the *input audio data*. As explained by Mr. Schmandt, each of these vectors $(x_1, \dots, x_T)$ is a vector representing the content present in a particular frame (i.e. frame 1, 2, 3, ..., T) of the input speech data, which, in combination, represent the phonetic content of p . EX1003, ¶¶85-86.

Thus, a PHOSITA would have recognized, because the set of phonemes $(x_1, \dots, x_T)$ form a vector set representing the phonetic content of the source speech, and because Keshet expressly describes the source speech including a “specific

accent,” the vectors $(x_1, \dots, x_T)$ teach *first phonetic embedding vectors*⁴ for phonetic content ***representing a source accent***. EX1003, ¶¶85-86 citing EX1004-Keshet, 159.

4. ***Claim 1(c): “apply a trained accent conversion neural network to the first phonetic embedding vectors to generate second phonetic embedding vectors corresponding to first phonetic characteristics of speech data in a target accent;”***

After generating the first phonetic embedding vectors $(x_1, \dots, x_T)$ from the input speech, Keshet’s masking module 160 performs an unclaimed intervening step in some embodiments of generating a masking vector m to be multiplied to $\tilde{x}_k$ (via an “element-by-element” product) in order to isolate a segment of interest representing a particular phoneme for modification and its immediate vicinity. EX1004-Keshet, 87-89; EX1003, ¶¶87-91. Keshet’s masking module 160 is highlighted purple in Fig. 6 below.

⁴ Hereinafter, the phrase “phonetic embedding vectors” will be abbreviated as “PEVs” unless reciting claim language verbatim.

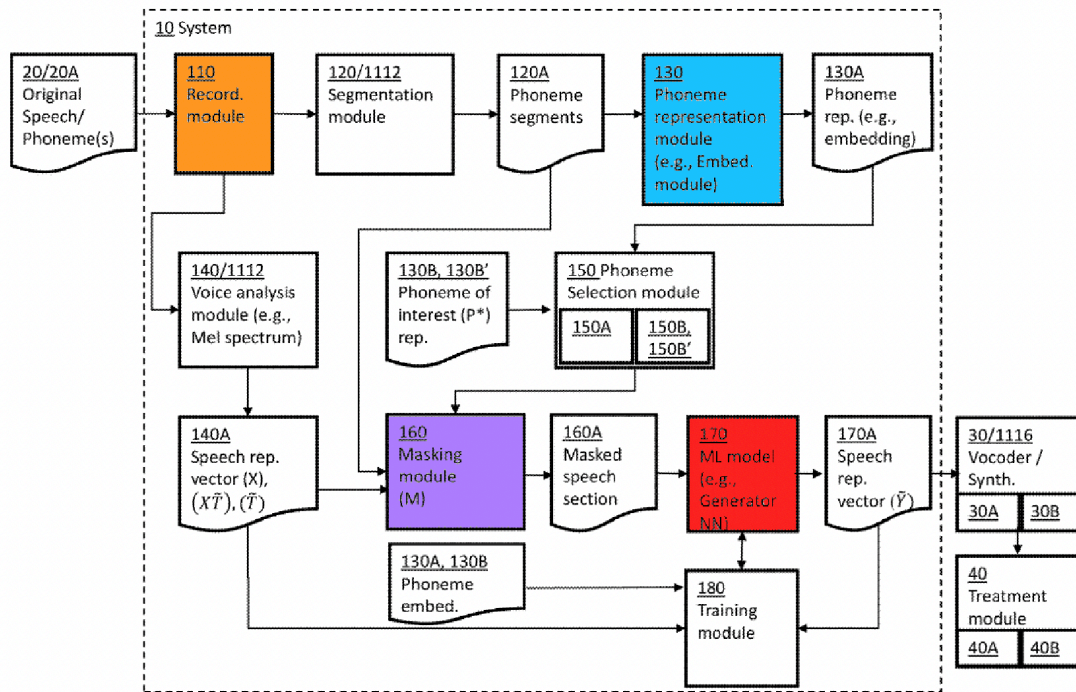


FIG. 6

EX1004-Keshet, Fig. 6 (annotated); EX1003, ¶87. Mr. Schmandt explains this masking vector simply serves to provide a binary 0 or 1 for each frame's vector x_k in $\tilde{x}_k$, identifying which portions of the input speech include an accent needing correction. *Id.* This masking vector is then applied to the first PEV $\tilde{x}_k$ using a simple “element-by-element” product. *Id.*

Thereafter, a trained machine learning model such as ML 170 is applied to the input speech vector to generate a set of second PEVs denoted y_k . EX1004-Keshet, 90-96, *see also* 97-99; EX1003, ¶88. Keshet's ML (Machine Learning) model 170 is highlighted red in Fig. 6 below.

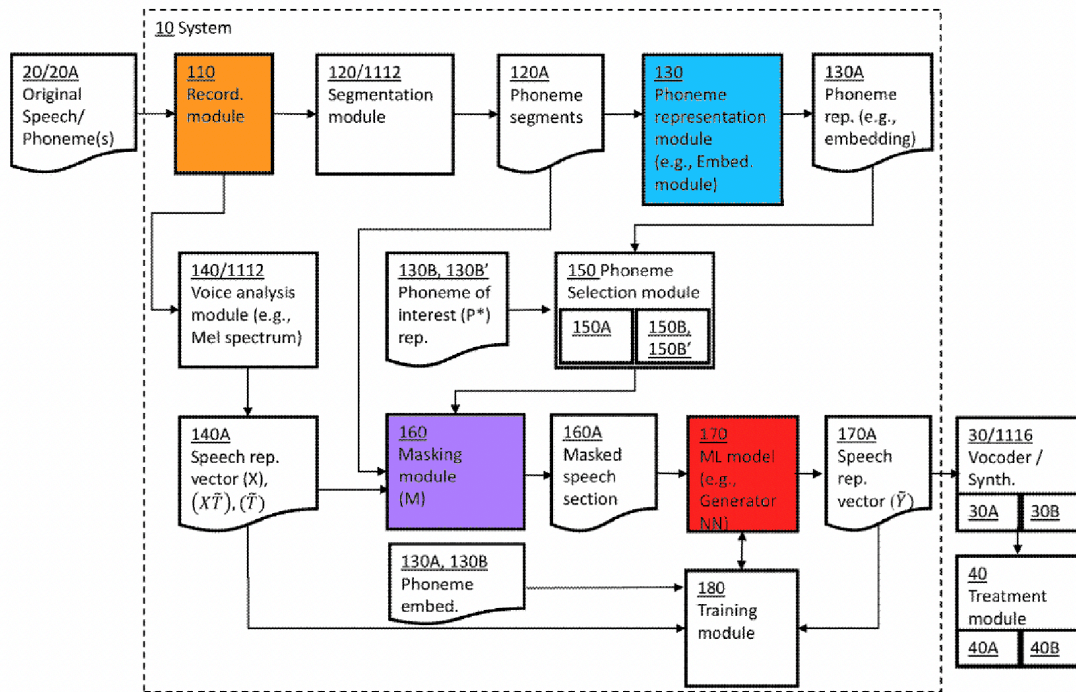


FIG. 6

EX1004-Keshet, Fig. 6 (annotated); EX1003, ¶88. *Keshet* expressly teaches that ML model 170 may comprise a neural network, “trained on unimpaired speech only [] that represents accurate pronunciation.” EX1004-Keshet, 95-98; EX1003, ¶88-90. Thus, *Keshet’s* system *applies a trained accent conversion neural network to the first phonetic embedding vectors. Id.*

As shown in Fig. 5 below, the output of ML model 170 is $\tilde{y}_k$, a string of corrected phonemes corresponding to the identified portions of speech in need of correction. EX1004-Keshet, 93, 97; EX1003, ¶¶89-90.

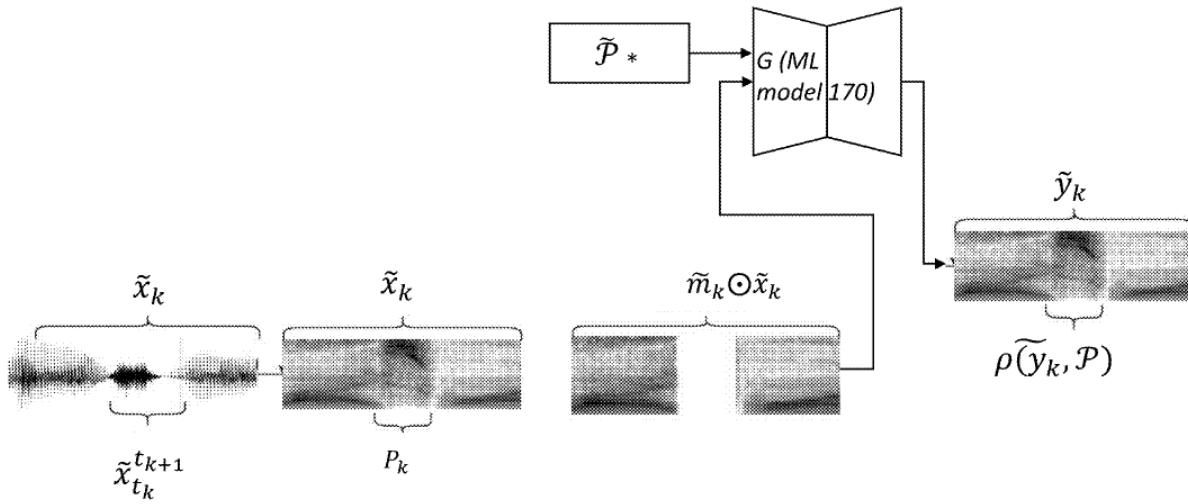


FIG. 5

EX1004-Keshet, Fig. 5. The vector $\tilde{y}_k$ thus teaches the claimed *second phonetic embedding vectors in a target accent. Id.*

5. ***Claim 1(d): “determine a differentiable alignment by jointly maximizing a cosine distance between the first phonetic embedding vectors and the second phonetic embedding vectors; and”***

The '745 Patent describes calculating a cosine difference between the first and second PEVs to identify and align portions of the original accented speech with non-accented, corrected speech. EX1001, 9:45-58, 10:24-32, Fig. 3; EX1003, ¶91. As explained by Mr. Schmandt, because $\cos(0)=1$, maximizing the cosine distance between the PEVs results in alignment such that each vector in the first set of PEVs is aligned with the most similar vector from the second set of PEVs. *Id.* The alignment (step 306) is temporal, on a vector by vector basis performed jointly over time. Each pair of vectors is normalized to magnitude 1 and then the cosine

difference is calculated to determine an alignment. Based on this alignment, the second set of PEVs is adjusted to match the first set of PEVs. *Id.* As will be shown, the combination of *Keshet*, *King*, and *Cuturi* teach claim 1(d). EX1003, ¶¶91-106.

a) *Keshet's Teachings*

Keshet teaches jointly comparing the first PEVs and second PEVs by determining the cosine distance between frames with each of vector sets $\tilde{x}_k$ and $\tilde{y}_k$. EX1004-*Keshet*, 100-103; EX1003, ¶¶92-93. *Keshet* envisions this may be done using a machine learning model “trained with cosine similarity loss applied between the embedding vectors” to jointly compare frames $(x_1, \dots, x_T)$ of $\tilde{x}_k$ with corresponding frames of $\tilde{y}_k$ on a frame-by-frame basis. *Id.* Though *Keshet* envisions that its system may operate even if the originally uttered (x_k) and replacement (y_k) phonemes are of unequal length, *Keshet* omits detailed discussion of how a PHOSITA would have specifically managed this situation, using dynamic alignment. EX1003, ¶92 *citing* EX1004-*Keshet*, 100. *Keshet* does not expressly provide all of the details for *maximizing the cosine distance between the first PEVs and second PEVs* or utilizing a *differentiable* alignment. However, *King* does.

b) *King's Teachings*

King teaches sound feature alignment techniques. EX1005-*King*, Abstract. Like *Keshet*, *King* describes dynamic alignment, which is performed by a Time Warping Module 318, a portion of *King's* Adjustment and Synthesis Module 316.

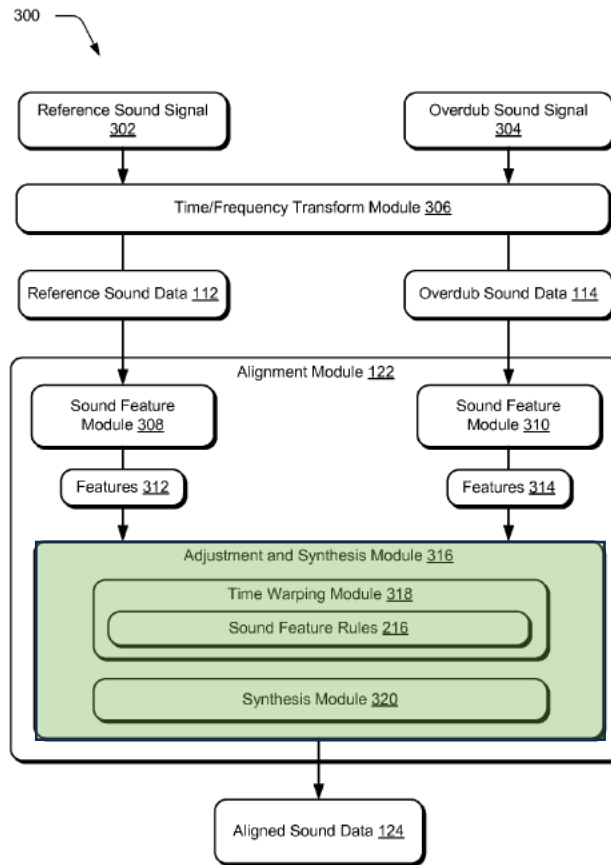


Fig. 3

EX1005-King, 35-41, Fig. 3 (annotated).

King expressly teaches utilizing cosine distance to maximize similarity between features in speech vectors, as described by the '745 Patent. EX1005-King, 36-38; EX1003, ¶¶94-96. “The similarity matrix thus contains a comparison of every frame of the overdub sound data 114 to every frame of the reference sound data...done by a normalized inner product (e.g., cosine distance).” *Id.*

Further, *King* teaches its Adjustment and Synthesis Module 316 may additionally “stretch and compress sound data to match the timing of other sound

data ... by dividing the sound data into a series of short frames and analyzing the spectral content of the frames” using the similarity matrix. “Based on this analysis, a determination may be made as to how the timing of the frames from the overdub sound data 114 is to be changed *such that the spectral characteristics best match the frames in the reference sound data* 112.” In doing so, the resulting overdubbed sound data matches the timing of the reference sound data to remain “consistent with human perception.” EX1005-King, 34, 43, Fig. 3; EX1003, ¶96. Thus, *King* teaches *determining an alignment by maximizing the cosine distance between speech vectors*, which a PHOSITA would have been motivated to employ to jointly align *Keshet’s* first PEVs and second PEVs. *Id.* In the intended mapping, *King’s* Time Warping Module 318 is combined into the modified *Keshet* accent modification system for determining a Similarity Matrix aligning the first and second PEVs. As explained by Mr. Schmandt, corresponding vectors in x_k and y_k are compared jointly to determine the values in *King’s* similarity matrix. EX1003, ¶¶95-96.

c) Motivation to Combine *Keshet* and *King*

A PHOSITA would have been motivated and found it obvious to utilize *King’s* Time Warping Module 318 and sound feature alignment techniques to align the first and second PEVs in *Keshet’s* accent conversion system by calculating a similarity matrix maximizing the cosine distance (similarity) with a reasonable expectation of success. EX1003, ¶¶97-98.

A PHOSITA would have been motivated to utilize *King*'s teachings to better "reflect human perception," desirably improving the system to provide a more "natural sounding" output. EX1003, ¶97 *citing* EX1005-King, 37, 48. A PHOSITA would have recognized the combination constitutes applying a known technique (*King*'s vector alignment by jointly maximizing a cosine distance between corresponding vectors) to a known device (*Keshet*'s hardware and software modules) with the predictable result of better-aligned feature vectors, further motivating the combination. *Id.* A PHOSITA would have had a REOS for the combination at least because the modification would have required only reprogramming of software using hardware already present. EX1003, ¶98. A PHOSITA would have recognized *Keshet* already performs alignment, and *King* merely provides additional implementation details a PHOSITA would have incorporated without undue experimentation, further motivating the combination and leading to a REOS. *Id.*

"differentiable alignment"

The '745 Patent discusses that dynamic time warping (DTW), if performed using a **non-differentiable** approach, may constrain the accuracy and speed of an accent conversion system. EX1001, 1:26-48; Non-differentiability "makes it difficult to train an accent conversion model effectively," leading to "inaccuracies and errors in the accent conversion process and resulting poor-quality audio signals."

Id. As Mr. Schmandt explains, differentiable alignment desirably helps account for alignment during training. EX1003, ¶103.

Mr. Schmandt further explains that traditional (non-differentiable) DTW may, in some cases, provide undesirable output at segment boundaries, such as suddenly changing the value or sign (+/–) of the amount by which the second set of PEVs is stretched/compressed with respect to the first set of PEVs. EX1003, ¶¶99-101. Using *differentiable* alignment, such as that taught by *Cuturi*, ensures that the time ratio between the first and second PEVs remains continuous, avoiding the undesirable effects described by the '745 Patent. EX1001, 10:58-65; EX1003, ¶¶99-101.

d) Cuturi's Teachings

Cuturi describes techniques for determining an alignment between vectors that is differentiable and whose gradient may be computed, as in the description of the '745 Patent. *Compare* EX1001, 1:26-48, 10:58-65 *with* EX1006-*Cuturi*, 1-5 (e.g. at 3, “As the minimum over a finite set of linear functions of Δ , dtw_0 is therefore locally differentiable w.r.t. the cost matrix Δ , with gradient A^* , a fact that has been exploited in all algorithms designed to average time series under the DTW metric.”); EX1003, ¶102. As explained by Mr. Schmandt, a PHOSITA would have had the skill necessary and would have been motivated to employ *Cuturi's* differentiable alignment in the *Keshet-King* system to ensure the ratio between the first and second

PEVs is continuous, eliminating unwanted outcomes at segment boundaries while also improving the training phase of the overall system. EX1003, ¶¶102-103.

e) Motivation to Modify *Keshet-King* to Utilize *Cuturi*'s Differentiable Alignment

A PHOSITA would have been motivated and found it obvious to use *Cuturi*'s known differentiable alignment technique in the *Keshet-King* accent conversion system to increase the speed, efficiency, and accuracy of the conversion. EX1003, ¶¶103-105. A PHOSITA would have been motivated by *Cuturi*'s express statement that “one typically observes an increase in performance” when using *Cuturi*'s differentiable alignment technique. EX1006-Cuturi, 2, 8; EX1003, ¶103. A PHOSITA would have been further motivated by *Cuturi*'s disclosure that its technique, “when paired with a flexible learning architecture such as a neural network, [] allows for a differentiable end-to-end approach to design predictive and generative models for time series.” EX1006-Cuturi, 2; EX1003, ¶103. A PHOSITA would have recognized the similarities between *King*'s similarity matrix and the alignment matrix illustrated in Fig. 2 of *Cuturi*. EX1003, ¶¶103-105 *citing* EX1006-Cuturi, 2-3, Fig. 2. A PHOSITA would have been further motivated to utilize differentiable alignment to enable an accent conversion system that accounts for an alignment function in its training phase, avoiding issues with discontinuous back propagation. EX1003, ¶103 *citing* EX1036-Liu, 2-3.

A PHOSITA would have recognized the combination constitutes applying a known technique (*Cuturi's* differentiable dynamic alignment) to a known device (*Keshet's* hardware and software modules, modified to incorporate *King's* techniques) with the predictable result of differentiable vector alignment accounting for alignment during training and desirably smoothing corner cases while avoiding discontinuities at phoneme boundaries, further motivating the combination. EX1003, ¶¶103-105.

A PHOSITA would have had a REOS for the combination at least because the modification would have required only reprogramming of software using hardware already present, and because *Cuturi* is already intended for use with a neural network. EX1006-Cuturi, 2; EX1003, ¶105.

6. Claim 1(e): “align the speech data to the phonetic content based on the differentiable alignment to generate and provide output audio data corresponding to the aligned speech data and representing the target accent.”

As explained in Claim 1(d), the *Keshet* accent conversion system is modified pursuant to *King* to produce an alignment matrix to determine an amount by which to stretch or compress the second PEVs (*Keshet's* $\tilde{y}_k$ vectors) to match the first set of PEVs (*Keshet's* $\tilde{x}_k$ vectors, with a $\tilde{m}_k$ mask applied via an element-by-element product).

In the combined *Keshet-King-Cuturi* system, output audio data is generated based on *King's* alignment matrix, calculated using *Cuturi's* differentiable

alignment technique. The resulting output audio data is generated synthesized speech that “comprises the spectrogram features of the desired” corrected phonemes, and thus *represents the target accent*, as claimed. EX1004-Keshet, 106-108; EX1003, ¶¶106-108. The synthesized generated speech may be used to generate a new speech utterance, words, stream of words, and/or segment thereof, that represents corrected or desired speech that may be joined into the original input speech to replace a corresponding wrongly pronounced speech segment. *Id.* Per *Keshet*, “the result is a modified version of the original input speech comprising corrected pronunciation of a phoneme of interest, which was mispronounced originally.” EX1004-Keshet, 106-108. In embodiments, *Keshet* teaches a vocoder as part of its voice synthesis module that may be used to provide the audio data to a user in an audible form. EX1004-Keshet, 133-137, Fig. 6. Thus, the modified *Keshet* system teaches *aligning the speech data to the phonetic content based on the differentiable alignment to generate and provide output audio data corresponding to the aligned speech data and representing the target accent*.

Motivations to modify *Keshet*’s system to perform *King*’s alignment techniques and to specifically utilize *Cuturi*’s differentiable alignment are provided above regarding limitation 1(d) and intended for incorporation by reference herein. *Supra*, Section V.A.5.

B. Claim 2

1. ***Claim 2: “...wherein the first phonetic embedding vectors represent second phonetic characteristics of input speech in the input audio data in a numerical format and encode ... phonemes...present in the input speech.”***

Keshet's first PEVs encode phonemes present in the input speech in a numerical format. *Supra*, Section V.A.3 citing EX1004-*Keshet*, 81-82; EX1003, ¶109 (explaining *Keshet's* phonemes are denoted as D-dimensional feature vectors that “may include or may represent voice characteristics such as frequency bins of the Mel-spectrum or mel-cepstrum of the recorded speech of the speaker,” which teaches *encoding phonemes in a numerical format*). Such encoded feature vectors *represent second phonetic characteristics of input speech* and specifically *encode one or more phonemes present in the input speech*. EX1004-*Keshet*, 11, 94-95, 98, 162; EX1003, ¶109.

C. Claim 4

1. ***Claim 4: “...in order to determine the cosine distance... normalize the first and second phonetic embedding vectors by scaling the first and second phonetic embedding vectors to have a magnitude of one and preserving a relative direction of the first and second phonetic embedding vectors; and ...generate a dot product of the normalized first and second phonetic embedding vectors.”***

As discussed above, *Keshet's* system is modified pursuant to *King* to jointly maximize a cosine distance between the first and second PEVs by producing a similarity matrix. *Supra*, Section V.A.5. *King* teaches its similarity matrix “contains

a comparison of every frame of the overdub sound data 114 to every frame of the reference sound data 112. This is typically done by a **normalized inner product** (e.g., cosine distance) as described above although other examples are also contemplated. A value of ‘1’ means that the two frames are identical (**up to a scale factor**) and a value of ‘-1’ means that the two frames are complete opposites.” EX1006-King, 38.

As explained by Mr. Schmandt, *King’s* “inner product” is synonymous with the claimed *dot product*. EX1003, ¶¶110-111 *citing* EX1011-Tanton, 147. Mr. Schmandt also explains “normalization” is a mathematical process by which direction of a vector is retained while the magnitude of the vector is scaled to 1. *Id.* Thus, *King’s* disclosure of “normalizing” the PEVs teach the claimed *scaling the first and second phonetic embedding vectors to have a magnitude of one and preserving [their] relative direction*, while *King’s* “inner product” disclosure teaches *generating a dot product of the normalized first and second phonetic embedding vectors*. *Id.*

D. Claim 8

1. Claim 8(Pre)

Supra, Section V.A.1.

2. Claim 8(a)

Supra, Sections V.A.2.-V.A.3.

3. Claim 8(b)

Supra, Section V.A.4.

4. Claim 8(c)

Supra, Section V.A.5.

5. Claim 8(d)

Supra, Section V.A.6.

E. Claim 9

1. Claim 9

Supra, Section V.B.1.

F. Claim 11

1. Claim 11

Supra, Section V.C.1.

G. Claim 15

1. Claim 15(Pre)

Keshet specifically states it “may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.” EX1004-Keshet, 170, 51, 72; *see also supra*, Section V.A.1.

2. Claim 15(a)

Supra, Sections V.A.2.-V.A.3.

3. Claim 15(b)

Supra, Section V.A.4.

4. Claim 15(c)

Supra, Section V.A.5.

5. Claim 15(d)

Supra, Section V.A.6.

H. Claim 16

1. Claim 16

Supra, Section V.B.1.

I. Claim 17

1. Claim 17

Supra, Section V.C.1.

VI. GROUND 2: CLAIMS 3 AND 10 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND KAMEOKA

A. Claim 3

1. *Claim 3(a): “...wherein the accent conversion neural network is trained to learn a mapping between the first phonetic embedding vectors and the second phonetic embedding vectors using a labeled dataset comprising paired samples of [source]⁵ accent phonetic embedding vectors and corresponding target accent phonetic embedding vectors; and comprises an encoder layer configured to encode the first phonetic embedding vectors into a latent representation and a decoder layer configured to decode the latent representation to generate the second phonetic embedding vectors.”*

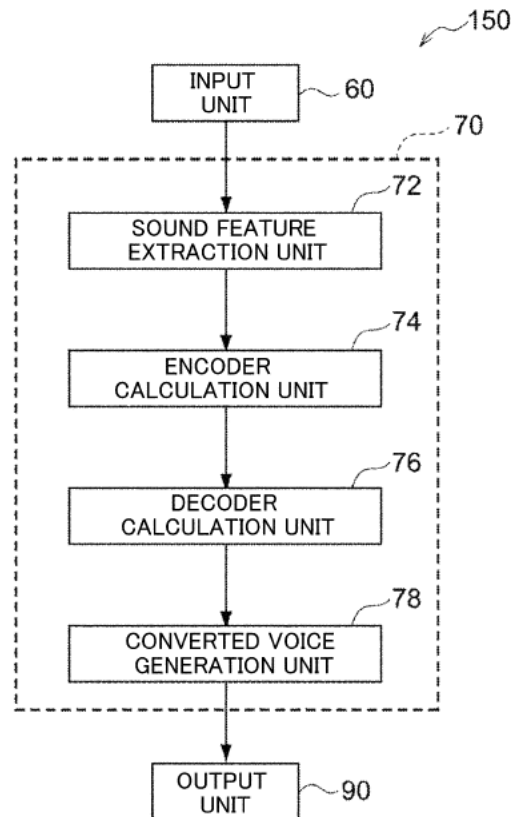
As discussed above regarding Claim 1, *Keshet* teaches an accent conversion neural network to generate second PEVs in a target accent from first PEVs in a source accent. *Supra*, Section V.A.4. While *Keshet*’s discloses its ML model 170 comprises a neural network “trained on unimpaired speech [] that represents accurate pronunciation,” *Keshet* does not expressly provide all of the details regarding how its neural network is trained. EX1004-Keshet, 95-98; EX1003, ¶¶126-131.

⁵ A PHOSITA would have recognized the word “**course**” in claim 3 is clearly a typographical error in which two letters have been transposed, intending to read “comprising paired samples of **source** accent phonetic embedding vectors...” as is recited by the ’745 Specification. EX1003, ¶126 *citing* EX1001, 6:17-26.

a) ***Kameoka's Teachings***

Kameoka teaches a voice conversion learning system, method, and program that learns a conversion function which converts a voice to a desired attribution. EX1007-Kameoka, 1:17-21, 2:31-35. *Kameoka's* system utilizes a Convolutional Neural Network (CNN) to convert a time-dependent input feature vector representing input speech to an output feature vector with the desired attribution. EX1007-Kameoka, 6:28-7:4, 2:26-38. *Kameoka's* neural network comprises an ***encoder unit*** 74 (providing an *encoder layer*) and ***decoder unit*** 76 (providing a *decoder layer*).

Fig. 2



EX1007-Kameoka, Fig. 2, 9:45-67.

Kameoka's encoder calculation unit 74 estimates a latent vector series (a *latent representation*) from an input sound feature vector series in a *source* voice signal (first PEVs) using an attribution label, while its decoder unit 76 estimates a sound feature vector series of a *target* voice signal using the latent vector series and the attribution labels. EX1007-Kameoka, 2:42-3:31, 9:25-67, Fig. 2. “The encoder network is an NN that for a given input data vector x , outputs a parameter of a conditional distribution $q(z|x)$ of a latent variable vector z . The decoder network is an NN that for a given latent variable vector z , outputs a parameter of a conditional distribution $p(x|z)$ of a data vector x .” EX1007-Kameoka, 4:40-58 (*pinpoint* 4:43-48); EX1003, ¶129.

Kameoka's neural network learns (*is trained to learn*), via a learning unit, a function for converting, for each utterance, the sound feature vector to a latent variable vector series according to attribution *labels*. EX1007-Kameoka, 6:44-60; EX1003, ¶130. In an example, *Kameoka* expressly envisions that “the attribution label indicating the attribution of the conversion-source voice signal and the attribution label indicating the attribution of the target voice signal” may be provided manually or “may be extracted automatically from the conversion-source voice signal.” EX1007-Kameoka, 9:33-44, 8:46-67; EX1003, ¶130. Here, a set of data (or *dataset*) is provided to the neural network that links the source and target voice

signals (as *pairs*). *Id.* Thus, *Kameoka* teaches a *labeled dataset comprising paired samples of source and target accent vectors*, as claimed. *Id.*

“After the learning is completed, the sound feature vector series of the utterance of the conversion source attribution may be input to the encoder and target attribution label may be input to the decoder to reconfigure the sound feature vector series.” EX1007-Kameoka, 6:52-56. Thus, *Kameoka* teaches *an accent conversion neural network (CNN) trained to learn a mapping between source vectors and target vectors using a labeled dataset comprising paired samples of vectors; and an encoder layer encoding the source vectors into a latent representation and a decoder layer configured to decode the latent representation to generate output vectors.*

b) Motivation to Combine *Kameoka*

A PHOSITA would have been motivated and found it obvious to utilize *Kameoka's* labeled dataset and paired samples techniques as an operable, expected method of training the neural network in *Keshet's* accent conversion system to accurately generate second (target) PEVs from first (source) PEVs. EX1003, ¶¶132-133. A PHOSITA would have recognized the combination constitutes applying a known technique (training according to *Kameoka's* method) to a known device (*Keshet's* trained conversion neural network, comprising encoder and decoder layers as per *Kameoka*) with the predictable result of a trained accent-conversion neural network, further motivating the combination. *Id.* A PHOSITA would have

incorporated any necessary components, datasets, and/or software provided by *Kameoka* necessary for training *Keshet*'s system, including specifically the encoder and decoder neural networks, learning unit, and label dataset. *Id.* Particularly, a PHOSITA would have been motivated to incorporate *Kameoka*'s encoder unit 74 (teaching *an encoding layer*) and decoder unit 76 (teaching *a decoding layer*) into the modified *Keshet* system to perform *Kameoka*'s training techniques and provide operable hardware and software supporting *Keshet*'s encoder-decoder structure. EX1004-*Keshet*, 11, 94-95, 98, 162; EX1003, ¶¶132-133.

A PHOSITA would have had a REOS for the combination at least because the modification would have required only known software programming techniques common in machine learning applications. EX1003, ¶¶132-133. A PHOSITA would have recognized *Keshet* already trains its accent conversion neural network via training module 180, and *Kameoka* merely provides additional implementation details regarding such training and the structure of training module 180 that a PHOSITA would have incorporated without undue experimentation, further motivating the combination and leading to a REOS. EX1004-*Keshet*, 141-148, Fig.5; EX1003, ¶¶126-133.

B. Claim 10

1. Claim 10(a)

Supra, Section VI.A.1.

2. Claim 10(b)

Supra, Section VI.A.2.

VII. GROUND 3: CLAIMS 5, 12, AND 18 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND FAN

A. Claim 5

- 1. Claim 5: “...in order to generate the output audio data, one or more of align first frames of the speech data with corresponding second frames of the phonetic content, apply one or more of prosody modeling, intonation adjustment, or accent-specific acoustic modeling techniques, or adjust a speech rate, pitch, or gender, wherein the output audio data preserves linguistic content of the input audio data.”**

As discussed above regarding Claim 1, *Keshet-King-Cuturi* teaches aligning frames of the speech data with corresponding frames of phonetic content and generating output audio data preserving linguistic content of the input audio data. *Supra*, Sections V.A.5-V.A.6.

a) Fan’s Teachings

Fan teaches an accent conversion system that adjusts the parameters of input speech to a desired output. EX1008-Fan, 97, 120, 142. *Fan* specifically teaches modifying acoustic features such as the user’s tone, *pitch*, rhythm, and speaking speed (or *speech rate*). EX1008-Fan, 50, 106, 142; EX1003, ¶¶135-139. Using *Fan*’s method, “content of the second voice [the result of its beautification method] is the same as content of the input voice (the first voice),” meaning (like *Keshet*) *Fan*’s

output audio data preserves linguistic content of the input audio data, as claimed.

EX1008-Fan, 127.

b) Motivation to Combine *Fan*

A PHOSITA would have been motivated and found it obvious to utilize *Fan*'s accent modification techniques to transform the pitch (as claimed), rhythm, and/or speaking speed (the claimed "speech rate") in addition to the user's accent while retaining the content of the input speech in *Keshet*'s accent conversion system. EX1003, ¶138. A PHOSITA would have been recognized that *Keshet*'s system would have been desirably improved by the combination, enabling the modified system to provide a better, more customizable example for speech correction, such as in cases where the input speaker speaks too slowly. *Id.* A PHOSITA would have recognized the combination constitutes applying a known technique (*Fan*'s pitch, rhythm, and/or speech rate modification) to a known device (*Keshet*'s hardware and software modules, modified to incorporate *King*'s techniques) with the predictable result of an accent conversion system with pitch, rhythm, and/or speech rate modification capabilities, further motivating the combination. *Id.*

A PHOSITA would have had a REOS for the combination at least because the modification would have required only reprogramming of software using hardware already present. EX1003, ¶139.

B. Claim 12

1. Claim 12

Supra, Section VII.A.1.

C. Claim 18

1. Claim 18

Supra, Section VII.A.1.

VIII. GROUND 4: CLAIMS 6, 13, AND 19 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND CHEN

A. Claim 6

1. Claim 6: “...apply a gradient-based optimization algorithm to optimize the joint maximization of the cosine distance.”

In *Cuturi*’s technique, the gradient of the alignment with respect to all of its variables is computable as a “by-product of the computation of the discrepancy itself.” EX1006-Cuturi, 2-3 (2.2 Differentiation of soft-DTW). *Cuturi* does not expressly teach a *gradient-based optimization algorithm to optimize the joint maximization of the cosine distance*.

However, *Chen* teaches an ML similarity model used to compute the cosine similarity between frames of audio data whose parameters (such as internal weights) “may be learned through training the similarity model 534 on one or more training samples, which may comprise updating the parameters based on gradient-based optimization. EX1010-Chen, 13:42-60 *see also* 10:18-27, 11:34-47, 14:29-45, 22:37-48. As explained by Mr. Schmandt, this optimization technique, sometimes

called “back propagation,” feeds the gradient of the alignment back into the neural network, adjusting its weights to minimize a loss function as desired. EX1003, ¶¶143-145 *citing* EX1012-Tsatsin, 8:28-31. Thus, *Keshet-King-Cuturi-Chen* teaches Claim 6.

a) Motivation to Combine *Chen*

A PHOSITA would have been motivated and found it obvious to utilize *Chen*’s gradient-based optimization algorithm to minimize the loss function of its joint maximization of cosine distance. As explained by Mr. Schmandt, optimization techniques compute a loss function which determines the difference between the currently-generated output and the desired output. EX1003, ¶¶144-145 *citing* EX1012-Tsatsin. Parameters may then be adjusted to reduce the loss function. *Id.* Gradient-based approaches converge the loss function to the desired output quickly, motivating the combination. *Id.* Gradient-based optimization algorithms such as *Chen*’s were well-known techniques prior to the ’745 Patent, and its application to the similar *Keshet* system would have improved its accent modification in the same way as *Chen*’s, quickly reducing the loss function of the cosine distance joint maximization to a desired level. EX1003, ¶¶144-145; EX1010-Chen, 13:42-60.

A PHOSITA would have had a REOS for the combination at least because the modification would have required only reprogramming of software using hardware already present. EX1003, ¶145.

B. Claim 13

1. Claim 13

Supra, Section VIII.A.1.

C. Claim 19

1. Claim 19

Supra, Section VIII.A.1.

IX. GROUND 5: CLAIMS 7, 14, AND 20 ARE OBVIOUS OVER KESHET, KING, CUTURI, AND VYAS

A. Claim 7

1. Claim 7: “...apply one or more dimensionality reduction techniques to preprocess the first and second phonetic embedding vectors.”

The '745 Patent provides examples of well-known dimensionality reduction techniques in neural network architectures that may be used to preprocess vectors including specifically “Principal Component Analysis (PCA) or t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the computational complexity and enhance the alignment accuracy.” EX1001, 1:5-20, 2:61-3:14, 8:20-34, 9:41-44.

Vyas teaches the same. Specifically, *Vyas* teaches “a conventional **dimensionality reduction process** such as an encoder portion of an autoencoder, a **PCA process**, a t-SNE process, etc., that generates a latent embedding” representing a data sample of speech. EX1009-Vyas, 8:20-33; EX1003, ¶¶146-149. *Vyas* provides a number of examples in which its dimensionality reduction techniques

may be utilized in speech applications to “preprocess input data 102 into a form that can be processed by downstream component(s).” EX1009-Vyas, 5:15-28, 2:29-46 (providing speech examples); EX1003, ¶¶146-147.

a) Motivation to Combine *Vyas*

A PHOSITA would have been motivated and found it obvious to utilize *Vyas*’s dimensionality reduction techniques (such as PCA) to “preprocess input data 102 into a form that can be processed by downstream components” in *Keshet*’s accent conversion system, reducing computing overhead. EX1003, ¶148. A PHOSITA would have recognized this would have increased the accuracy and reduced the computational load necessary for *Keshet*’s processor, motivating the combination. *Id.* A POSITA would have been further motivated by *Vyas*’s express teaching that its techniques desirably “may allow for improved feature selection and/or weighting when interpretable embeddings are used.” EX1009-Vyas, 3:14-23; EX1003, ¶148. A PHOSITA would have recognized the combination constitutes applying a known technique (*Vyas*’s dimensionality reduction preprocessing techniques) to a known device (*Keshet*’s hardware and software modules) with the predictable result of accent conversion requiring less computational overhead, further motivating the combination. *Id.* EX1003, ¶148.

A PHOSITA would have had a REOS for the combination at least because the modification would have required only reprogramming of software using hardware already present. EX1003, ¶149.

B. Claim 14

1. Claim 14

Supra, Section IX.A.1.

C. Claim 20

1. Claim 20

Supra, Section IX.A.1.

X. CONCLUSION

For the foregoing reasons, Petitioner respectfully requests *inter partes* 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

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

A. Real Party-In-Interest

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

B. 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.*, IPR2026-00275 (PTAB), challenging U.S. Patent No. 11,715,457
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, PGR2026-00032 (PTAB), challenging U.S. Patent No. 12,412,561
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, PGR2026-00033 (PTAB), challenging U.S. Patent No. 12,417,756

C. Lead and Back-Up Counsel

Petitioner provides 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	Chris R. Schmidt (Reg. No. 63,982) chris.schmidt@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	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

D. 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. 12,131,745 for Claims 1-20

Claim Designation	Claim Language
Claim 1(P)	An accent conversion system, comprising an audio interface, memory having instructions stored thereon, and one or more processors coupled to the memory and configured to execute the instructions to:
Claim 1(a)	obtain input audio data via the audio interface;
Claim 1(b)	generate from the input audio data first phonetic embedding vectors for phonetic content representing a source accent;
Claim 1(c)	apply a trained accent conversion neural network to the first phonetic embedding vectors to generate second phonetic embedding vectors corresponding to first phonetic characteristics of speech data in a target accent;
Claim 1(d)	determine a differentiable alignment by jointly maximizing a cosine distance between the first phonetic embedding vectors and the second phonetic embedding vectors; and
Claim 1(e)	align the speech data to the phonetic content based on the differentiable alignment to generate and provide output audio data corresponding to the aligned speech data and representing the target accent.
Claim 2	The accent conversion system of claim 1, wherein the first phonetic embedding vectors represent second phonetic characteristics of input speech in the input audio data in a numerical format and encode one or more of phonetic features, patterns, phonemes, pronunciation, intonation, speech sounds, or phonetic units present in the input speech.
Claim 3(P)	The accent conversion system of claim 1, wherein the accent conversion neural network:
Claim 3(a)	is trained to learn a mapping between the first phonetic embedding vectors and the second phonetic embedding vectors using a labeled dataset comprising paired samples of source accent phonetic embedding vectors and corresponding target accent phonetic embedding vectors; and

Claim Designation	Claim Language
Claim 3(b)	comprises an encoder layer configured to encode the first phonetic embedding vectors into a latent representation and a decoder layer configured to decode the latent representation to generate the second phonetic embedding vectors.
Claim 4(P)	The accent conversion system of claim 1, wherein the one or more processors are further configured to execute the instructions to, in order to determine the cosine distance:
Claim 4(a)	normalize the first and second phonetic embedding vectors by scaling the first and second phonetic embedding vectors to have a magnitude of one and preserving a relative direction of the first and second phonetic embedding vectors; and
Claim 4(b)	generate a dot product of the normalized first and second phonetic embedding vectors.
Claim 5	The accent conversion system of claim 1, wherein the one or more processors are further configured to execute the instructions to, in order to generate the output audio data, one or more of align first frames of the speech data with corresponding second frames of the phonetic content, apply one or more of prosody modeling, intonation adjustment, or accent-specific acoustic modeling techniques, or adjust a speech rate, pitch, or gender, wherein the output audio data preserves linguistic content of the input audio data.
Claim 6	The accent conversion system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a gradient-based optimization algorithm to optimize the joint maximization of the cosine distance.
Claim 7	The accent conversion system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply one or more dimensionality reduction techniques to preprocess the first and second phonetic embedding vectors.
Claim 8(P)	A method for automatic alignment of phonetic content for real-time accent conversion, the method implemented by an accent conversion system and comprising:
Claim 8(a)	obtaining a set of phonetic embedding vectors for phonetic content representing a source accent and obtained from input audio data;

Claim Designation	Claim Language
Claim 8(b)	applying a trained machine learning model to the set of phonetic embedding vectors to generate a set of transformed phonetic embedding vectors corresponding to phonetic characteristics of speech data in a target accent;
Claim 8(c)	determining an alignment by maximizing a cosine distance between the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors; and
Claim 8(d)	aligning the speech data to the phonetic content based on the determined alignment to generate output audio data representing the target accent.
Claim 9	The method of claim 8, wherein the first phonetic embedding vectors represent second phonetic characteristics of input speech in the input audio data in a numerical format and encode one or more of phonetic features, patterns, phonemes, pronunciation, intonation, speech sounds, or phonetic units present in the input speech.
Claim 10(P)	The method of claim 8, wherein the machine learning model:
Claim 10(a)	is trained to learn a mapping between the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors using a labeled dataset comprising paired samples of course accent phonetic embedding vectors and corresponding target accent phonetic embedding vectors; and
Claim 10(b)	comprises an encoder layer configured to encode the set of phonetic embedding vectors into a latent representation and a decoder layer configured to decode the latent representation to generate the set of transformed phonetic embedding vectors.
Claim 11(P)	The method of claim 8, further comprising, in order to determine the cosine distance:
Claim 11(a)	normalizing the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors by scaling the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors to have a same magnitude and preserving a relative direction of the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors; and

Claim Designation	Claim Language
Claim 11(b)	generate a dot product of the normalized the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors.
Claim 12	The method of claim 8, further comprising, in order to generate the output audio data, one or more of aligning first frames of the speech data with corresponding second frames of the phonetic content, applying one or more of prosody modeling, intonation adjustment, or accent-specific acoustic modeling techniques, or adjusting a speech rate, pitch, or gender, wherein the output audio data preserves linguistic content of the input audio data.
Claim 13	The method of claim 8, further comprising applying a gradient-based optimization algorithm to optimize the joint maximization of the cosine distance.
Claim 14	The method of claim 8, further comprising applying one or more dimensionality reduction techniques to preprocess the set of phonetic embedding vectors and the set of transformed phonetic embedding vectors.
Claim 15(P)	A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:
Claim 15(a)	obtain first phonetic embedding vectors for phonetic content representing a source accent and determined from obtained input audio data;
Claim 15(b)	apply a trained neural network to the first phonetic embedding vectors to generate second phonetic embedding vectors corresponding to phonetic characteristics of speech data in a target accent;
Claim 15(c)	determine an alignment by maximizing a cosine distance between the first phonetic embedding vectors and the second phonetic embedding vectors; and
Claim 15(d)	align the speech data to the phonetic content based on the determined alignment to generate output audio data representing the target accent.

Claim Designation	Claim Language
Claim 16	The non-transitory computer-readable medium of claim 15, wherein the first phonetic embedding vectors represent second phonetic characteristics of input speech in the input audio data in a numerical format and encode one or more of phonetic features, patterns, phonemes, pronunciation, intonation, speech sounds, or phonetic units present in the input speech.
Claim 17(P)	The non-transitory computer-readable medium of claim 15, wherein the instructions, when executed by the at least one processor further causes the at least one processor to, in order to determine the cosine distance:
Claim 17(a)	normalize the first and second phonetic embedding vectors by scaling the first and second phonetic embedding vectors to have a magnitude of one and preserving a relative direction of the first and second phonetic embedding vectors; and
Claim 17(b)	generate a dot product of the normalized first and second phonetic embedding vectors.
Claim 18	The non-transitory computer-readable medium of claim 15, wherein the instructions, when executed by the at least one processor further causes the at least one processor to, in order to generate the output audio data, one or more of align first frames of the speech data with corresponding second frames of the phonetic content, apply one or more of prosody modeling, intonation adjustment, or accent-specific acoustic modeling techniques, or adjust a speech rate, pitch, or gender, wherein the output audio data preserves linguistic content of the input audio data.
Claim 19	The non-transitory computer-readable medium of claim 15, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a gradient-based optimization algorithm to optimize the joint maximization of the cosine distance.
Claim 20	The non-transitory computer-readable medium of claim 15, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply one or more dimensionality reduction techniques to preprocess the first and second phonetic embedding vectors.

APPENDIX OF EXHIBITS

Exhibit 1001	U.S. Patent No. 12,131,745 (“’745 Patent”)
Exhibit 1002	Prosecution History for the 12,131,745 Patent (“’745 File History”)
Exhibit 1003	Declaraton of Christopher Schmandt
Exhibit 1004	U.S. Patent Application Publication No. 2024/0304200 to Keshet et al. (“ <i>Keshet</i> ”)
Exhibit 1005	U.S. Patent Application Publication No. 2014/0148933 to King et al. (“ <i>King</i> ”)
Exhibit 1006	Marco Cuturi and Mathieu Blondel, Soft-DTW: a Differentiable Loss Function for Time-Series (“ <i>Cuturi</i> ”)
Exhibit 1007	U.S. Patent No. 11,450,332 to Kameoka et al. (“ <i>Kameoka</i> ”)
Exhibit 1008	U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>Fan</i> ”)
Exhibit 1009	U.S. Patent No. 12,400,078 to Vyas et al. (“ <i>Vyas</i> ”)
Exhibit 1010	U.S. Patent No. 12,217,761 to Chen et al. (“ <i>Chen</i> ”)
Exhibit 1011	Tanton, James, <i>Encyclopedia Of Mathematics</i> (Facts On File, Inc. 2005).
Exhibit 1012	U.S. Patent No. 10,909,459 to Tsatsin et al. (“ <i>Tsatsin</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	<i>Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning</i> to Zhao et al. (“ <i>Accentron</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, Voice Communication with Computers, Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)
Exhibit 1022	C. O’Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallag

	han.com/research/papers/Pitch.pdf) September 30, 2015 (“ <i>O’Callaghan</i> ”)
Exhibit 1023	B. Sisman et al., Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, June 2019, 1085-1097 (“ <i>Sisman</i> ”)
Exhibit 1024	J. Holmes, Speech Synthesis and Recognition, Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003 (“ <i>Holmes</i> ”)
Exhibit 1025	H. Fletcher, Loudness, Its Definition, Measurement and Calculation, Bell Telephone Laboratories, October 1933 (“ <i>Fletcher</i> ”)
Exhibit 1026	S. Ding, Personal VAD: Speaker-Conditioned Voice Activity Detection, arXiv, April 8, 2020 (“ <i>Chang</i> ”)
Exhibit 1027	C. Wiltshire, Chapter 13 English Sounds in Context: The Pronunciation of Phonemes and Morphemes, in Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners, 303-324, 2019 (“ <i>Wiltshire</i> ”)
Exhibit 1028	N. Kazanina, et al., Phonemes: Lexical access and beyond, Psychon Bull Rev 25, 560–585, September 5, 2017 (“ <i>Kazanina</i> ”)
Exhibit 1029	D. Jurafsky et al., Speech and Language Processing, Chapter 27, October 16, 2019, (“ <i>Jurafsky</i> ”)
Exhibit 1030	L. Sun, Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training, IEEE on Multimedia and Expo, August 29, 2016, (“ <i>Sun</i> ”)
Exhibit 1031	U.S. Patent No. 11,664,037 to Lim et al. (“ <i>Lim</i> ”)
Exhibit 1032	U.S. Patent Application Publication No. 2022/0122579 to Biadsky et al. (“ <i>Biadsky</i> ”)
Exhibit 1033	U.S. Patent Publication No. 2018/0174595 to Dirac et al. (“ <i>Dirac</i> ”)
Exhibit 1034	L. Roberts, Understanding the Mel Spectrogram, (https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53), March 5, 2020 (“ <i>Roberts</i> ”)
Exhibit 1035	N. Umeda, <i>Consonant duration in American English</i> , The Journal of the Acoustical Society of America, March 1, 1977 (“ <i>Umeda</i> ”)
Exhibit 1036	S. Liu et al., <i>End-to-End Accent Conversion Without Using Native Utterances</i> , 2020 IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP) 6289-6293 (“ <i>Liu</i> ”)
Exhibit 1037	U.S. Patent Appl. No. 2022/0246132 to Zhang et al. (“ <i>Zhang</i> ”)

Exhibit 1038	Y. Wang, Tacotron: Towards End-to End Speech Synthesis, arXiv, April 6, 2017 (“ <i>Wang</i> ”)
Exhibit 1039	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 (“ <i>Zhao</i> ”)
Exhibit 1040	S. Khorram et al, Trainable Time Warping: Aligning Time Series in the Continuous-Time Doman, arXiv, March 21, 2019 (“ <i>Khorram</i> ”)

CERTIFICATION OF WORD COUNT

The undersigned certifies pursuant to 37 C.F.R. §42.24 that the foregoing Petition for *Inter Partes* 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 7040 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
Adam M. Sandwell, Reg. No. 72,484
adam.sandwell@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 *Inter Partes* Review and Exhibits were provided via Federal Express to the Patent Owner by serving the correspondence address of record for the '745 Patent:

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

Further, a courtesy copy of this Petition for *Inter Partes* 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