

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. IPR2026-00275
U.S. Patent No. 11,715,457

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 11,715,457**

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	SUMMARY OF THE '457 PATENT.....	1
A.	DESCRIPTION OF THE '457 PATENT	1
B.	SUMMARY OF THE PROSECUTION HISTORY	2
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)	3
C.	CLAIM CONSTRUCTION UNDER 37 C.F.R. § 42.104(B)(3)	3
IV.	SHOWING OF ANALOGOUS, PRIOR ART.....	3
A.	OVERVIEW OF <i>FEINAUER</i>	5
B.	OVERVIEW OF <i>FAN</i>	6
C.	OVERVIEW OF <i>PRABHAVALKAR</i>	9
D.	OVERVIEW OF <i>ACCENTRON</i>	10
E.	OVERVIEW OF <i>SINGH</i>	11
F.	OVERVIEW OF <i>BONGIOVI</i>	12
G.	OVERVIEW OF <i>WU</i>	12
A.	OVERVIEW OF <i>OSUNA</i>	12
V.	GROUND 1: CLAIMS 1, 3-5, 8-10, 14, 16-17, AND 20 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, AND FAN.....	13
A.	CLAIM 1	13
1.	<i>Claim 1(Pre): “A method for real-time correction of an accent in a speech audio signal, the method being implemented by a computing device and comprising:”</i>	<i>13</i>
2.	<i>Claim 1(a): “dividing the 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>	<i>15</i>
a)	<i>Feinauer’s Teachings.....</i>	<i>15</i>
b)	<i>Prabhavalkar’s Teachings</i>	<i>15</i>
c)	<i>Motivation to Combine Feinauer and Prabhavalkar</i>	<i>17</i>
3.	<i>Claim 1(b): “extracting, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”</i>	<i>18</i>
a)	<i>Feinauer’s Teachings.....</i>	<i>19</i>
b)	<i>Fan’s Teachings.....</i>	<i>19</i>
c)	<i>Motivation to Combine Feinauer-Prabhavalkar and Fan.....</i>	<i>20</i>
d)	<i>Prabhavalkar’s Teachings</i>	<i>21</i>
e)	<i>Motivation to Combine Prabhavalkar’s Context</i>	<i>22</i>

4.	<i>Claim 1(c): “extracting, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”</i>	23
	a) <i>Fan’s Teachings</i>	24
	b) <i>Motivation to Combine Feinauer-Prabhavalkar and Fan</i>	25
5.	<i>Claim 1(d): “receiving a speaker embedding for a human speaker; and”</i>	26
	a) <i>Fan’s Teachings</i>	27
	b) <i>Motivation to Combine Fan’s speaker embedding</i>	29
6.	<i>Claim 1(e): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”</i>	31
	a) <i>Feinauer’s Teachings</i>	32
B.	CLAIM 2	34
1.	<i>Claim 2(a): “The method for real-time correction of claim 1, wherein: the speech audio signal is recorded from a voice of a user by the computing device via an acoustic sensor; and:”</i>	34
2.	<i>Claim 2(b): “a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.”</i>	35
	a) <i>Prabhavalkar’s Teachings</i>	35
	b) <i>Feinauer’s Teachings</i>	36
	a) <i>Motivation to Combine Prabhavalkar</i>	36
C.	CLAIM 3	37
1.	<i>Claim 3: “The method for real-time correction of claim 1, wherein the speaker embedding is generated based on the speech audio signal.”</i>	37
	a) <i>Motivation to Combine Fan’s speaker embedding</i>	39
D.	CLAIM 4	40
1.	<i>Claim 4: “The method for real-time correction of claim 1, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.”</i>	40
E.	CLAIM 8	40
1.	<i>Claim 8: “The method for real-time correction of claim 1, wherein the context is a pre-determined number of the frames belonging to chunks preceding the input chunk in the stream of input chunks.”</i>	40
F.	CLAIM 14	40
1.	<i>Claim 14[Pre]: “A computing apparatus comprising: a processor; and a memory storing instructions that, when executed by the processor, configure the apparatus to:”</i>	40
2.	<i>Claim 14(a): “divide a speech audio signal including an accent into a stream of input chunks, an input chunk from the stream of</i>	

	<i>input chunks including a pre-defined number of frames of the speech audio signal;”</i>	41
3.	<i>Claim 14(b): “extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”</i>	41
4.	<i>Claim 14(c): “extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”</i>	41
5.	<i>Claim 14(d): “receive a speaker embedding for a human speaker; and;”</i>	41
6.	<i>Claim 14(e): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”</i>	41
G.	CLAIM 15	41
1.	<i>Claim 15(a): “The computing apparatus of claim 14, wherein: the speech audio signal is recorded from a voice of a user by the computing apparatus via an acoustic sensor; and”</i>	41
2.	<i>Claim 15(b): “a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.”</i>	42
H.	CLAIM 16	42
1.	<i>Claim 16: “The computing apparatus of claim 14, wherein the speaker embedding is generated based on the speech audio signal.”</i>	42
I.	CLAIM 17	42
1.	<i>Claim 17: “The computing apparatus of claim 14, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.”</i>	42
J.	CLAIM 20	42
1.	<i>Claim 20[Pre]: “A non-transitory computer-readable storage medium, the computer-readable storage medium including instructions that when executed by a computer, cause the computer to:”</i>	42
2.	<i>Claim 20(a): “divide a speech audio signal including an accent 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>	43
3.	<i>Claim 20(b): “extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”</i>	43
4.	<i>Claim 20(c): “extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”</i>	43

- 5. Claim 20(d): “receive a speaker embedding for a human speaker; and” 43
- 6. Claim 20(e): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.” 43

VI. GROUND 2: CLAIMS 9-11 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND ACCENTRON..... 44

- A. CLAIM 9 44
 - 1. Claim 9: “The method for real-time correction of claim 1, wherein the generating the output chunk includes: providing the speaker embedding, the acoustic features, and the linguistic features to a synthesis module to generate a melspectrogram with the reduced accent; and providing the melspectrogram to a vocoder to generate the output chunk of the output audio signal.” 44
 - a) Accentron’s Teachings 44
 - a) Motivation to Combine the modified Feinauer with Accentron and Fan 46
- B. CLAIM 10 48
 - 1. Claim 10(a): “The method for real-time correction of claim 9, wherein: the synthesis module includes an encoder, a decoder, and a post-net module designed to improve an output of the decoder; and;” 48
 - a) Accentron’s Teachings 48
 - b) Motivation to Combine Accentron 50
 - 2. Claim 10(b): “generating the melspectrogram includes: processing the linguistic features with the reduced accent by the encoder to generate hidden features; combining the hidden features, the acoustic features, and the speaker embeddings to generate further features; processing the further features by the decoder and the post-net module to generate the melspectrogram.” 51
 - a) Motivation to Combine Accentron 53
- C. CLAIM 11 56
 - 1. Claim 11: “The method for real-time correction of claim 10, wherein: the acoustic features are split into a stream of acoustic features chunks corresponding to the chunks in the stream of input chunks; the linguistic features are split into a stream of linguistic features chunks corresponding to the chunks in the stream of input chunks; the melspectrogram is split into a stream of melspectrogram chunks corresponding to the chunks in the stream of input chunks; 56
and a melspectrogram chunk of the stream of melspectrogram chunks is generated based on the following: an acoustic features chunk of the stream of acoustic features chunks and acoustic features context including the pre-determined number of acoustic features frames belonging to acoustic features chunks preceding the acoustic features chunk in the stream of acoustic features chunks;

and; a linguistic features chunk of the stream of melspectrogram chunks and linguistic features context including the pre-determined number of linguistic features frames belonging to linguistic features chunks preceding the linguistic features chunk in the stream of acoustic features chunks.” 56

VII. GROUND 3: CLAIMS 6 AND 19 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND SINGH 57

- A. CLAIM 6 57
1. Claim 6: “The method for real-time correction of claim 1, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.” 57
- a) Singh’s Teachings 58
- b) Motivation to Combine Singh 60
- B. CLAIM 19 60
1. Claim 19: “The computing apparatus of claim 14, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.” 60

VIII. GROUND 4: CLAIMS 7 AND 18 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND WU..... 61

- A. CLAIM 7 61
1. Claim 7: “The method for real-time correction of claim 1, wherein the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.” 61
- a) Wu’s Teachings 61
- b) Motivation to Combine Wu 63

IX. GROUND 5: CLAIMS 12-13 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND BONGIOVI 64

- A. CLAIM 12 64
1. Claim 12: “The method for real-time correction of claim 1, further comprising, prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust one or more characteristics of the speech audio signal to improve extraction of the linguistic features and the acoustic features.” 64
- a) Feinauer-Prabhavalkar-Fan’s Teachings 64

	b)	<i>Bongiovi’s Teachings</i>	65
	c)	Motivation to Combine <i>Bongiovi</i>	67
B.	CLAIM 13		69
	1.	<i>Claim 13: “The method for real-time correction of claim 1, further comprising: prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust a loudness of the speech audio signal from a first level to a second level; after generating the output chunk of the output audio signal, processing the output chunk by the digital signal processing module, to adjust the loudness of the output audio signal to the first level.”</i>	69
	a)	Motivation to Combine <i>Bongiovi</i>	70
X.	GROUND 6: CLAIM 5 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND OSUNA		72
A.	CLAIM 5		72
	1.	<i>Claim 5: “The method for real-time correction of claim 1, wherein the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library.”</i>	72
	a)	Osuna’s Teachings	72
	b)	Motivation to Combine <i>Osuna</i>	73
XI.	GROUND 7: CLAIM 18 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, WU, AND OSUNA		74
A.	CLAIM 18		74
	1.	<i>Claim 18: “The computing apparatus of claim 14, wherein: the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library; and the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.”</i>	74
XII.	CONCLUSION		74
XIII.	MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)		76
	A.	REAL PARTY-IN-INTEREST.....	76
	B.	RELATED MATTERS	76
	C.	LEAD AND BACK-UP COUNSEL	76
	D.	37 C.F.R. § 42.8(B)(4) – SERVICE INFORMATION.....	77

TABLE OF AUTHORITIES

CASES

Penumbra, Inc. v. RapidPulse, Inc., IPR2021-01466, Paper 34, at 29-35
(PTAB March 10, 2023)7
Phillips v. AWH Corp., 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc)3

STATUTES

35 U.S.C. § 102(a)(1)5, 11, 12, 13
35 U.S.C. § 102(a)(2)7,10
35 U.S.C. § 102(d)(2)7
35 U.S.C. § 1033

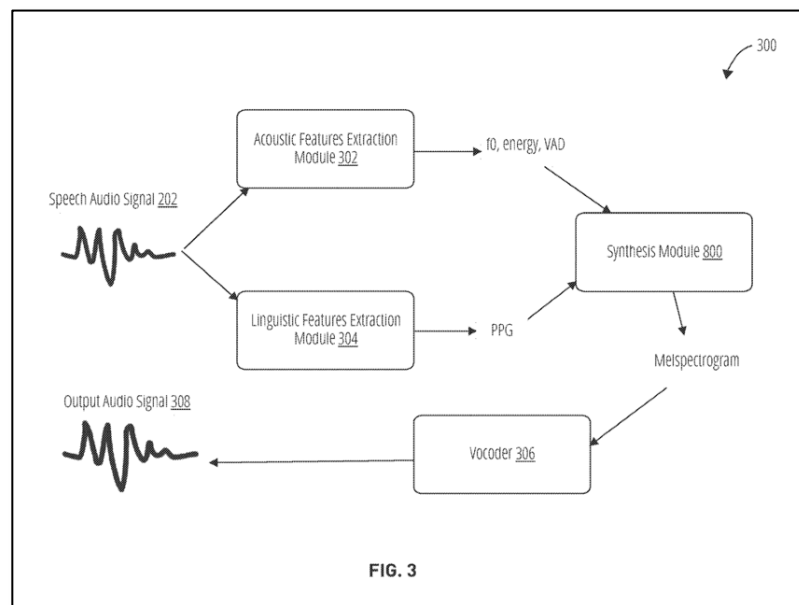
I. INTRODUCTION

Krisp Technologies, Inc. (“Petitioner”), respectfully requests inter partes review (“IPR”) of claims 1-20 (“Challenged Claims”) of U.S. Patent No. 11,715,457 (“’457 Patent”).

II. SUMMARY OF THE ’457 PATENT

A. Description of the ’457 Patent

The ’457 Patent describes “real-time correction of an accent in a speech audio signal.” EX1001, Abstract. “FIG. 3 is a block diagram showing a pipeline 300 for real-time correction of an accent in speech audio signals, according to an example embodiment. The pipeline 300 may include an acoustic features extraction module 302, a linguistic features extraction module 304, a synthesis module 800, and a vocoder 306.” EX1001, 6:6-11.



EX1001, Fig. 3.

B. Summary of the Prosecution History

The '457 Patent faced no office actions and received a notice of allowance March 15, 2023. Ex.1002. The '457 claims priority to Provisional Application No. 63/297,901 filed January 10, 2022. Because the '457 claims the benefit of a filing date after March 16, 2013, Petitioner applies AIA versions of §§ 102, 103. For this Proceeding, petitioner applies January 10, 2022, 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 '457 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

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

Petitioner certifies the '457 Patent is eligible for IPR.

¹ Petitioner does not concede any of the Challenged Claims are entitled to the priority of January 10, 2022.

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
Ground 1: Claims 1-4, 8, 14-17, and 20 are obvious over <i>Feinauer, Prabhavalkar, and Fan</i> .	1004, 1005, 1006
Ground 2: Claims 9-11 are obvious over <i>Feinauer, Prabhavalkar, Fan, and Accentron</i> .	1004, 1005, 1006, 1008
Ground 3: Claims 6 and 19 are obvious over <i>Feinauer, Prabhavalkar, Fan, and Singh</i> .	1004, 1005, 1006, 1007
Ground 4: Claim 7 is obvious over <i>Feinauer, Prabhavalkar, Fan, and Wu</i> .	1004, 1005, 1006, 1009
Ground 5: Claims 12-13 are obvious over <i>Feinauer, Prabhavalkar, Fan, and Bongiovi</i> .	1004, 1005, 1006, 1010
Ground 6: Claim 5 is obvious over <i>Feinauer, Prabhavalkar, Fan, and Osuna</i> .	1004, 1005, 1006, 1011
Ground 7: Claim 18 is obvious over <i>Feinauer, Prabhavalkar, Fan, Osuna, and Wu</i> .	1004, 1005, 1006, 1009, 1011

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 '457 Patent. The prior art is from the same field of endeavor of the '457 Patent, namely speech audio

processing. EX1001, Abstract, 1:16-19, 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]-[0138], [0201]; EX1008-Accentron, 1, 33; EX1007-Singh, [0003]-[0004], [0007]; EX1009-Wu, 1:6-8, 5:46-47, 5:55-56, Fig. 2; EX1010-Bongiovi, [0046], Fig. 3; EX1011-Osuna, Abstract; EX1003, ¶¶56-57, 64, 70, 75, 81, 86, 90, 93, 97.

The prior art is also reasonably pertinent to at least one problem facing the '457 Patent inventors, namely at least one of:

(A) real-time audio modification: EX1001, Title, Abstract, 1:16-19, 1:23-29, 4:23-27, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086], EX1008-Accentron, 1, 33, EX1005-Prabhavalkar, 17:46-51, 18:33-37, 20:16-21; 22:16-27;

(B) accent conversion; EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086]; EX1006-Fan, Abstract, [0120], [0135], [0201]; EX1008-Accentron, 1, EX1011-Osuna, Abstract;

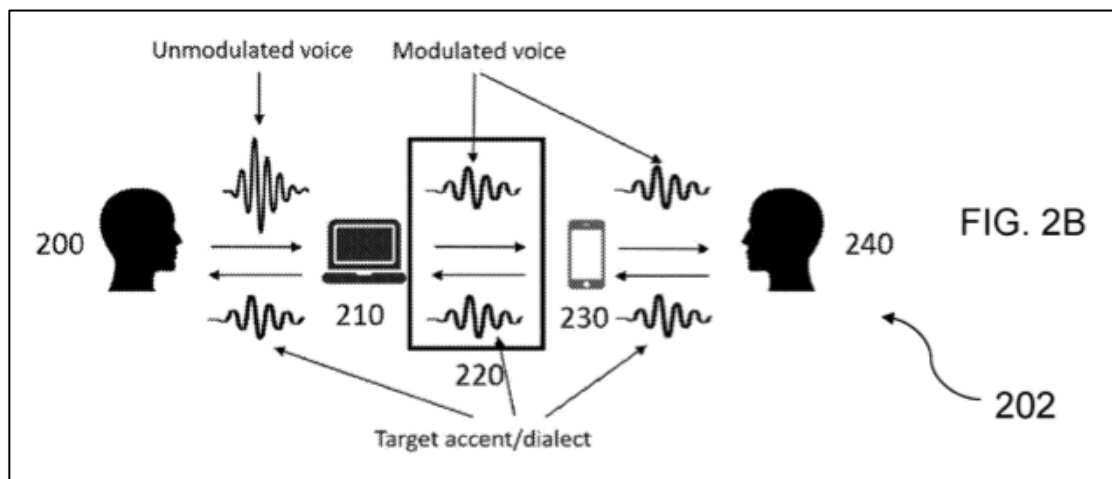
(C) feature extraction from audio data; EX1001, 4:16-18, 14:15-19, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086], [0095], [0099].; EX1005-Prabhavalkar, 2:44-50, 11:36-41, Fig. 2; EX1006-Fan, Abstract, [0113], [0120], [0127], [0135], [0201]; EX1007-Singh, Abstract, [0003]-[0004],

[0007], [0022]; EX1008-Accentron, 1, 8, Figs. 1-2; Figs. 1-2; EX1011-Osuna, Abstract; and/or

(D) audio processing quality improvement; EX1001, 11:54-67, 14:20-27; EX1005-Prabhavalkar, 1:29-35; EX1007-Singh, [0009], [0022]; EX1010-Bongiovi, [0002]-[0003], [0034]-[0035]; EX1009-Wu, Abstract, 2:43-57, 3:43-44, Fig. 2; EX1003, ¶58-59, 65-67, 71-72, 76-78, 82-84, 87-88, 91, 94, 98-99.

A. 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.

B. 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*.

Fan is entitled to an effective filing date of September 21, 2020 (and therefore qualifies as prior art to the ’457 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

² EX1006 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).

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 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, ¶¶68, 73.

<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. 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
[0003]	1, ¶3	1, ¶2
[0007]	2, ¶2	2, ¶2
[0010]	3, ¶2	3, ¶1
[0019]	5, ¶3	5, ¶2
[0020]	5, ¶4	5, ¶3
[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
[0056]	10, ¶7	10, ¶6
[0075]	13, ¶5	13, ¶4
[0081]	14, ¶1	14, ¶1
[0082]	14, ¶2	14, ¶2
[0085]	14, ¶5	14, ¶5
[0086]	15, ¶1	2, ¶2, 15, ¶1
[0088]	15, ¶3	15, ¶3
[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
[0100]	18, ¶3	18, ¶3
[0106]	19, ¶6	19, ¶5
[0113]	21, ¶1	20, ¶6
[0120]	22, ¶4	22, ¶3
[0124]	23, ¶2	22, ¶7
[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
[0150]	28, ¶4	28, ¶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; EX1003, ¶¶69-70.

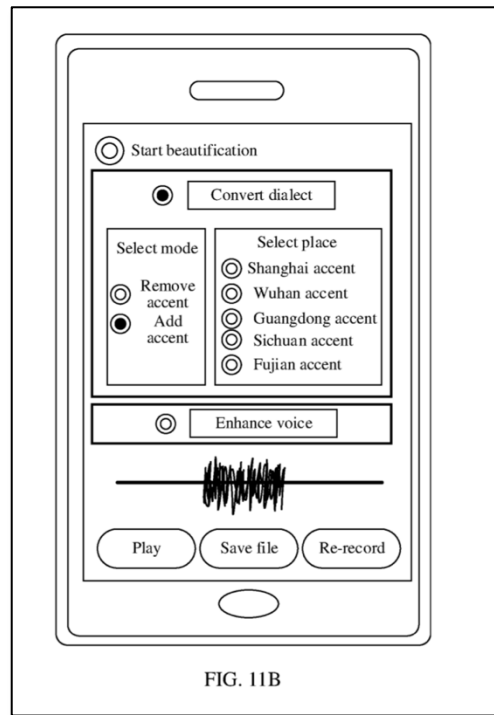


FIG. 11B

EX1006-Fan, Fig. 11B.

C. Overview of *Prabhavalkar*

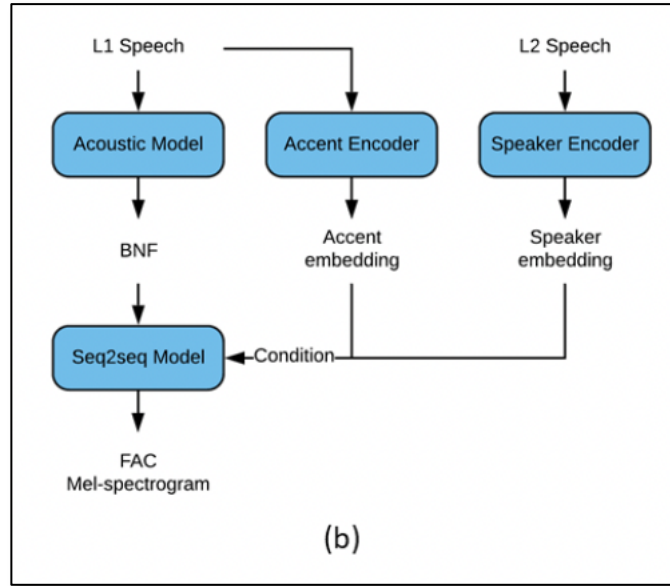
U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“*Prabhavalkar*”) was filed on September 20, 2021, and published October 1, 2024, qualifying as prior art under 35 U.S.C § 102(a)(2) AIA.

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.

D. Overview of *Accentron*

“*Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning*” to Zhao et al. (“*Accentron*”) was publicly available online on ScienceDirect December 31, 2021, and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA. EX1016-Munford, 7-35 (declaration of June Munford establishing public availability of *Accentron*), EX1046 (affidavit of Mina Ching at the Internet Archive authenticating the support for *Accentron*’s public availability.)

Accentron teaches “an approach that can generate accent-converted speech for arbitrary L2 speakers.” EX1008-*Accentron*, 1.



EX1008-Accentron, Fig. 1. (excerpt).

E. Overview of *Singh*

U.S. Patent Appl. No. 2019/0325861 to Singh et al. (“*Singh*”) was filed April 18, 2019, and published October 24, 2019, and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA.

Singh teaches “automatic speech recognition using domain adaptation techniques.” EX1007-Singh, [0002]. *Singh* teaches “[t]he domain can be accent[.]” EX1007-Singh, [0034]. *Singh* teaches to optimize the neural network “by using a special Gradient Reversal Layer ... at the end of feature extractor 6 and at the beginning of domain classifier 8.” EX1007-Singh, [0022].

F. Overview of *Bongiovi*

U.S. Patent Appl. No. 2008/0219459 to Bongiovi et al. (“*Bongiovi*”) was filed March 14, 2008, and published September 11, 2008, and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA.

Bongiovi teaches “a loudness adjusting step to compensate for loudness variations caused by the volume control of the output device.” EX1010-Bongiovi, [0035]. Specifically, “method 300 can be used to generate an output signal with a predetermined loudness regardless of the volume of the input audio signal.” EX1010-Bongiovi, [0046].

G. Overview of *Wu*

U.S. Patent No. 11,164,592 to Wu et al. (“*Wu*”) was filed on May 9, 2019, and granted on November 2, 2021, and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA.

Wu teaches a value of a voice activity detector, “that may determine whether voice activity (e.g., speech) is detected.” EX1009-Wu, 5:46-47, 5:55-56.

H. Overview of *Osuna*

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion” to Zhao et al. (“*Osuna*”) was publicly available at least as early as October 2019 in IEEE and therefore qualifies as prior art under 35 U.S.C § 102(a)(1) AIA. EX1016-Munford, 44-47 (declaration of June Munford establishing public availability of *Osuna*).

Osuna teaches “accent conversion” using phonetic postetiograms (PPGs) “to measure the similarity of speech frames across speakers. A phonetic posteriorgram is computed by segmenting speech into frames and computing the posterior probability that each frame belongs to a set of pre-defined phonetic units.” EX1011-Osuna, Abstract, 3.

V. GROUND 1: CLAIMS 1-4, 8, 14-17, AND 20 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, AND FAN.

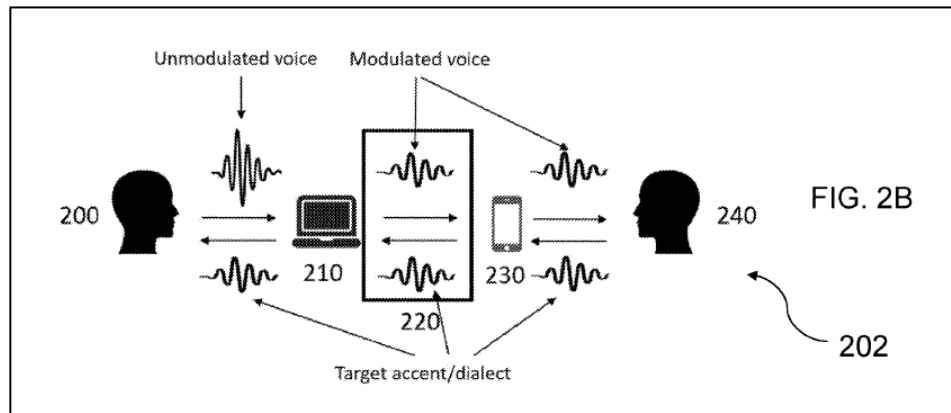
A. Claim 1

1. Claim 1(Pre): “A method for real-time correction of an accent in a speech audio signal, the method being implemented by a computing device and comprising:”

To the extent the preamble is limiting, *Feinauer* teaches a *method for real-time correction of an accent in a speech audio signal*³ that receives an unmodulated voice signal (a *speech audio 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

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.

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 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 a speech audio signal. EX1003, ¶102.

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 device*. EX1003, ¶¶103-104 *citing* EX1004-Feinauer, [0031]-[0032], [0041], [0043], [0121], Fig. 1.

2. Claim 1(a): “dividing the 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;”

a) Feinauer’s Teachings

Feinauer teaches receiving an unmodulated voice signal (*speech audio signal*) via a microphone. EX1004-Feinauer, [0031]-[0032], [0042], [0044]-[0046], [0056], Fig. 1; EX1003, ¶106-107. *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, [0057]; EX1003, ¶108. 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, ¶109.

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, ¶110. *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, ¶110.

Prabhavalkar characterizes its approach as “streaming,” wherein speech is 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, ¶111 *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, ¶112-113 *citing* EX1004-*Feinauer*, [0057]; EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22, 17:33-51, 18:18-22, 18:28-41, 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, ¶114 *citing* EX1005-*Prabhavalkar*, 17:33-51. *Feinauer* is targeted at real-time applications, such as talking over a phone, so a PHOSITA would have found it advantageous to apply *Prabhavalkar*'s temporal architecture for chunk processing to reduce latency. *Id.*

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, ¶115 *citing* EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22, 17:33-51, 22:1-38, 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, ¶¶115-116 *citing* EX1005-Prabhavalkar, 25:55-62.

3. *Claim 1(b): “extracting, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”*

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, ¶¶68, 73.⁴ *Prabhavalkar* teaches “attention” may comprise a pre-determined number of chunks preceding a current chunk, where the pre-determined number of chunks include a pre-determined number of frames. (*context*) EX1005-Prabhavalkar, 6:56-7:11, 15:16-22.

“extracting...from the input chunk...acoustic features”

⁴ In addition to the citations to *Fan*, Petitioner relies on parallel citations to the certified translation of *Fan*'s priority application (Ex. 1013) as set forth in the chart found in Section IV.B above and in Schmandt's declaration (Ex. 1003, ¶73).

The '457 Patent provides examples of “acoustic features” including “pitch 206 (or main frequency (F0)), energy 208 (signal amplitude), and voice activity detection.” EX1001, 5:60-62, 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, ¶¶119-120. As discussed above regarding Claim 1(a), 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 characteristic such as ***pitch, tone, volume, etc.***” EX1004-Feinauer, [0057]; EX1003, ¶120. 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, ¶121. Like *Feinauer*, *Fan* teaches modifying acoustic features such as the user’s tone, pitch, rhythm, and speaking speed. EX1006-Fan, [0050], [0106]; EX1003, ¶121. *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, ¶122.

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, ¶123. 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, [0045]-[0050], [0090]-[0094], [0127], [0152]-[0155], Figs. 5, 7; EX1003, ¶123. 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.* A PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s modules described functionality in *Fan*’s style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. *Id.*

“a context associated with the input chunk”

The '457 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, 1:64-66, 6:55-63, 17:1-4 (Claim 8).

As discussed regarding Claim 1(a), 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, 6:56-7:11, 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, ¶125. *Prabhavalkar* explains that “attention” may be computed using chunks immediately preceding each input chunk. EX1003, ¶125 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.

d) 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, ¶126 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*, 6:56-7:12. In a specific example, *Prabhavalkar* “computes 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, ¶127. 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, 1:64-66, 6:55-63, 17:1-4 (Claim 8); EX1003, ¶127.

e) Motivation to Combine *Prabhavalkar*’s Context

A PHOSITA would have been motivated to extract acoustic features as taught by *Feinauer-Prabhavalkar-Fan* by considering the context of the frames of the 10 previous chunks as taught by *Prabhavalkar*. EX1003, ¶128 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, ¶129 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, ¶130. 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.*

4. Claim 1(c): “extracting, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”

As discussed above regarding Claim 1(b), *Feinauer-Prabhavalkar-Fan* renders obvious extracting acoustic features from an input chunk and context. *See* Section V.A.3. 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, ¶132. 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. EX1003, ¶132.

The ’457 Patent states “linguistic features 210 can be represented as Phonetic PosteriorGrams (PPGs).” EX1001, 6:18-38, Fig. 3.

“extracting...from the input chunk...linguistic features with a reduced accent”

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, ¶134-135.

Fan teaches a voice feature extraction model that extracts phoneme posteriorgrams (PPGs) from the user’s speech representing the words spoken, as their phonemic constituents. EX1006-Fan, [0007], [0085]-[0086], Fig. 5; Fig. 9, [0126]-[0127], [0150], EX1003, ¶135. 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, ¶135. *Fan’s* PPGs are linguistic features because they represent phoneme probabilities and phonemes are a linguistic feature of any language. EX1006-Fan, [0085]. EX1003, ¶135. 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, ¶136.

As supported by the prior art and the ’457 specification, a PHOSITA would have recognized that PPGs extracted from a user’s speech carry a reduced accent.

EX1003, ¶137 *citing* EX1001, 6:24-35, 13:22-32; EX1012-Ding, Abstract. Specifically, as explained by Mr. Schmandt and supported by *Ding*, PPGs are mostly speaker-independent, but do carry some reduced amount of accent that may leak through to subsequently-generated speech. *Id.* Thus, because PPGs satisfy *linguistic features with a reduced accent*, *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, ¶138. 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, [0045]-[0050], [0085], [0090]-[0094], [0127], [0152]-[0155], Figs. 5, 7; EX1003, ¶138. 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.* 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. *Id.*

“a context associated with the input chunk”

As discussed regarding Claim 1(b), in the *Feinauer-Prabhavalkar-Fan* method, acoustic features are extracted from *input chunks and a context associated with an input chunk*, the frames of preceding chunks taught by *Prabhavalkar*. See Section V.A.3 citing EX1005-*Prabhavalkar*, 6:56-7:12, 15:16-22, 22:1-38, 17:33-51, 18:1-41. For the same reasons and motivations, a PHOSITA would have found it obvious to *extract linguistic features from an input chunk and a context associated with the input chunk* as well. EX1003, ¶139. The MTC and REOS provided for Claim 1(b) is incorporated by reference for this limitation.

5. Claim 1(d): “receiving a speaker embedding for a human speaker; and”

The '457 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, 13:33-38. The '457 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, 9:55-57. 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, 9:58-67. 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 to be matched. EX1004-Feinauer, [0081], [0105], [0110], Fig. 7; EX1003, ¶¶140-142. *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, ¶¶141-142. *Feinauer* does not provide further detail regarding the contents of its translation model.

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, ¶143. *Fan* envisions a “style embedding vector” that a PHOSITA would have recognized includes a representation of a target accent. EX1003, ¶¶144-146 *citing* EX1008-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].

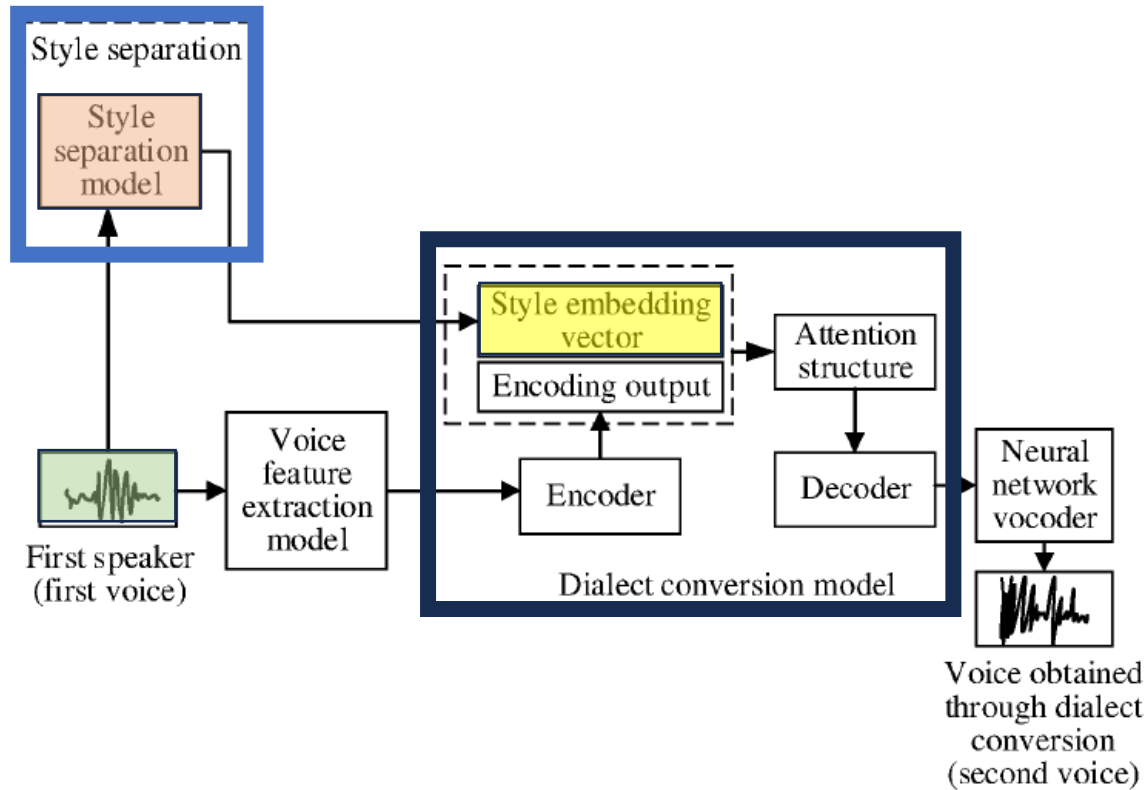


FIG. 9

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, ¶¶144-145 (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, ¶146 *citing* EX1008-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 timbre feature vector (*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, ¶147.

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, ¶¶148-149. 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, ¶150. *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, ¶¶149-150.

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(b)) and a speaker embedding (style embedding vector) to modify the accent (dialect) of the output audio signal.

EX1003, ¶151. 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*. EX1003, ¶150.

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. EX1003, ¶151. 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. EX1003, ¶151.

6. Claim 1(e): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

The '457 Patent describes output chunks, like input chunks, comprising a plurality of frames. EX1001, 11:49-53, 12:13-17. Specifically, the '457 Patent describes output chunks as corresponding in size to that of the input chunk from which they were generated. *Id.*

As discussed above, the modified *Feinauer* system combines a target accent *speaker embedding*, *acoustic features* such as pitch, tone, and/or volume per Claim 1(b), and *linguistic features* including PPGs of spoken words into an output audio

signal having a modified accent. EX1004-Feinauer, [0045], [0057]; EX1006-Fan, [0127], Fig. 9. 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(b) 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, ¶155 (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* translation model, modified to include *Fan's* style embedding vector template per Claim 1(d)), *acoustic features* (*Feinauer's* pitch, tone, and/or volume per Claim 1(b)), and *linguistic features* (*Feinauer's* individual words, modified to be represented by *Fan's* PPGs per Claim 1(c)). EX1004-Feinauer, [0086], [0110], [0117], Figs. 5-7; EX1003, ¶155. *Feinauer* teaches providing a modified output, envisioning its system may “modif[y] the voice of user 200 into an accent/dialect more familiar to call participant 240 (such as modifying the voice from an American

accent to an Australian accent).” EX1004-Feinauer, [0046]; EX1003, ¶156. *Fan* further provides detail, stating, “[t]hen, 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].

“generating an output chunk”

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, ¶¶157-158. Thus, *Prabhavalkar*’s generates output from a given chunk before moving on to processing succeeding chunks. *Id.* *Prabhavalkar*’s “chunk-based approach,” which consumes input as chunks, would apply the same sequential processing when applied to *Feinauer*’s “portions” of user’s speech to produce output chunks of processed audio frames as an “output chunk.” *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). EX1003, ¶159 *citing* EX1005-*Prabhavalkar*, 17:33-51. A MTC to utilize *Prabhavalkar*’s chunk-based approach is provided above regarding Claim 1(a) and is incorporated by reference for this limitation. *See* Section V.A.2.c. 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, ¶160. 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. Claim 2

1. Claim 2(a): “The method for real-time correction of claim 1, wherein: the speech audio signal is recorded from a voice of a user by the computing device via an acoustic sensor; and:”

Feinauer teaches the *speech audio signal is recorded from a voice of a user by the computing device via a microphone (acoustic sensor)*. Specifically, “communication devices 210 and 230 may include any type and form of hardware and software for receiving, detecting or being provided voice or audio as input, such as via a microphone.” EX1004-*Feinauer*, [0043]; *see also* [0031]. A PHOSITA

would have known that a microphone is an acoustic sensor as further supported by *Accentron*. EX1003, ¶161. EX1008-Accentron, 5.

2. Claim 2(b): “a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.”

Feinauer-Prabhavalkar-Fan in view of *Prabhavalkar* teaches a delay (100ms) between a first timestamp (150ms) and a second timestamp (100ms) is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.

a) Prabhavalkar’s Teachings

As discussed in Claim 1(a), *Prabhavalkar* teaches each chunk of a speech audio signal contains 5 frames. *Prabhavalkar*’s frames capture recorded speech audio signal. EX1005-*Prabhavalkar*, 26:57-65. Each frame is “downsampled to a 30ms frame rate.” EX1005-*Prabhavalkar*, 15:16-22; *see also* 20:14-20. Section V.A.1. Therefore, there are 150ms of recorded audio data in 5 frames. EX1003, ¶163. Thus, if a *first timestamp corresponding to a time when the chunk of the speech audio signal is recorded per Feinauer* is at an initial time, *Prabhavalkar*’s 150 milliseconds of recorded speech audio per chunk teaches that the first input chunk is complete after the first 150ms. EX1003, ¶163.

b) *Feinauer's Teachings*

As discussed in Claim 1(Pre), *Feinauer* teaches that the accent conversion process happens in 100ms. EX1004-Feinauer, [0086]. *See* Section V.A.1. Specially, *Feinauer* teaches “the conversation can be broadcast on a lag of which the duration must be long enough for a word to be expressed and analyzed for its accent and dialect strength and then modified to be expressed by the modification software.” EX1004-Feinauer, [0086]. This “lag duration can be a duration that sounds like real-time to a user (e.g. 100ms).” EX1004-Feinauer, [0086]. A PHOSITA would have understood that the claimed *second timestamp* corresponds to *Feinauer's* 100 milliseconds after completion of the first input chunk, at 250ms since the speech began. EX1003, ¶¶164-165. Thus, *Feinauer-Prabhavalkar-Fan* 250ms is *between 40 milliseconds and 300 milliseconds*, teaching Claim 2(b). EX1003, ¶¶164-165.

a) *Motivation to Combine Prabhavalkar*

As discussed in Claim 1(a), a PHOSITA would have been motivated to modify *Feinauer* to teach chunks of an audio signal. *See* Section V.A.2. A PHOSITA would have understood and been further motivated to modify *Feinauer* so that the chunks of the audio signal correspond to frames of a certain length of recorded speech audio, such as *Prabhavalkar's* 30ms frame length. EX1003, ¶166. A PHOSITA would have understood this is simply a design choice and in other circumstances a PHOSITA may have wanted the frame length to be longer or

shorter, balancing the resultant accuracy with the required computational overhead. EX1003, ¶167. A PHOSITA would have found it obvious and would have been motivated to use *Prabhavalkar's* teachings of dividing a speech utterance into chunks with each chunk having 5 frames each 30ms long as an operable resolution well-known in the field. EX1003, ¶168. A PHOSITA would have been motivated to look towards *Prabhavalkar's* teachings since it was well known that frames of an audio signal correspond to a time. EX1003, ¶169. A PHOSITA would have appreciated that such a combination would have been a combination of known elements yielding predictable results, including the ability to operate *Feinauer's* accent conversion software at increased speed and efficiency. EX1003, ¶169. A PHOSITA would have appreciated a REOS for this modification because the combination required only reprogramming of *Feinauer's* software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶169.

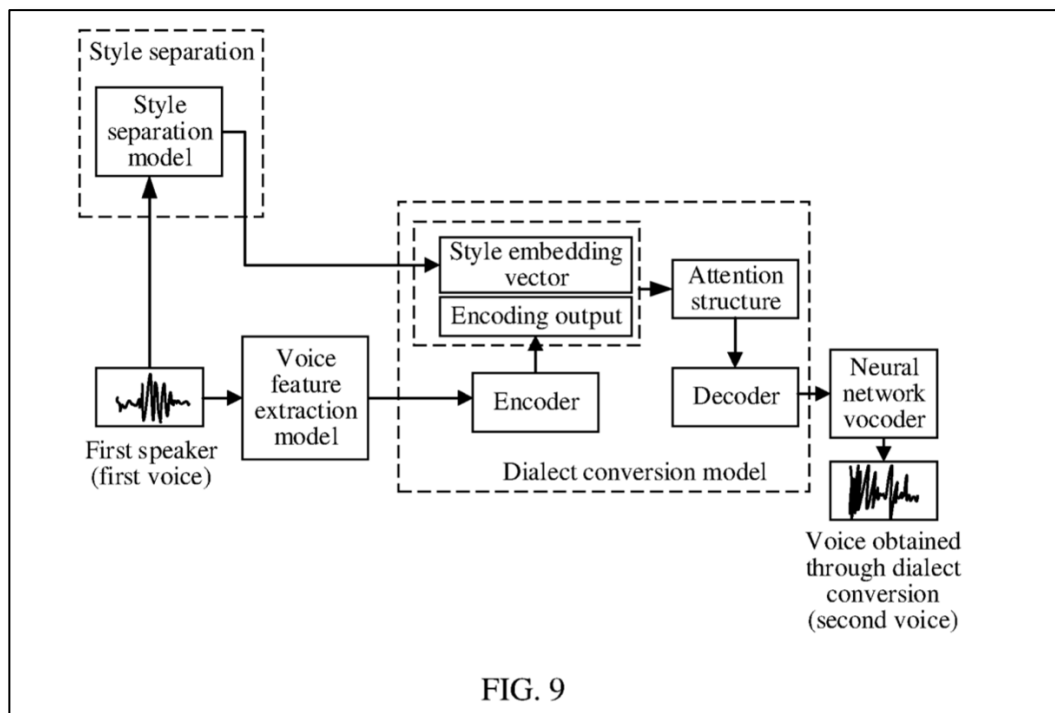
C. Claim 3

1. Claim 3: "The method for real-time correction of claim 1, wherein the speaker embedding is generated based on the speech audio signal."

As discussed in Claim 1(d), *Fan* discusses an embodiment a user may wish to sound like a different target accent user. *Fan* also teaches an embodiment where a user may wish to sound like themselves. EX1006-Fan, [0050], [0127], Fig. 9; EX1003, ¶171. In this case, *Fan's* system and method are being used for both accent

conversion and style enhancement. As Mr. Schmandt explains, *Fan's* style enhancement is an optional additional stage that may be activated based on user interface controls, and should not be understood as a distinct, exclusive embodiment from *Fan's* embodiments performing accent conversion only. EX1003, ¶¶171, 149.

Specifically, *Fan* teaches embodiments in which a style separation feature separates a style of the first voice to obtain a timbre feature vector of the first voice. EX1006-Fan, [0127]; EX1003, ¶¶171-172.



EX1006-Fan, Fig. 9.

Fan envisions that a user's voice's timbre may be represented as a "style embedding *vector*[" EX1006-Fan, [0094], Figs. 9, 12; EX1003, ¶173. A PHOSITA would have recognized this "style embedding vector" teaches a speaker embedding

vector consistent with the '457 specification. EX1003, ¶173 *citing* EX1008-Accentron, 21. Thus, in *Fan*, a *speaker embedding* vector at least reflecting the timbre of a human speaker is received by the processor of the device for modifying the accent of the input speech. *Id.*

a) 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 source user in *Feinauer* as an operable method of modifying the source voice’s timbre to match the timbre of the source user, modifying the resulting accent. EX1004-Feinauer, [0081], [0105], Fig. 7; EX1006-Fan, [0010], [0081]-[0082], [0094], [0160]-[0162], Figs. 12-13; EX1003, ¶174. A POSITA would have realized that such a combination in Claim 1(e), would still operate the same with the output chunk being generated based upon the speaker embedding. EX1003, ¶175. A PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer's* translation model and *Fan's* style embedding vector) 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.*

D. Claim 4

1. Claim 4: “The method for real-time correction of claim 1, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.”

As discussed regarding Claim 1(d), *Feinauer*’s system loads a relevant “translation model” *speaker embedding* corresponding to the particular target accent. See Section V.A.5 citing EX1004-Feinauer, [0081], [0105], Fig. 7. *Feinauer* further teaches its “target accent model can be trained on prior conversation data of people with differing accents[.]” EX1004-Feinauer, [0037], [0039], [0065].

E. Claim 8

1. Claim 8: “The method for real-time correction of claim 1, wherein the context is a pre-determined number of the frames belonging to chunks preceding the input chunk in the stream of input chunks.”

See Claim 1(b).

F. Claim 14

1. Claim 14[Pre]: “A computing apparatus comprising: a processor; and a memory storing instructions that, when executed by the processor, configure the apparatus to:”

As discussed in Claim 1(Pre), *Feinauer* teaches a computing device that includes a computer apparatus, processor, and memory storing instructions. EX1004-Feinauer, [0081], [0118]-[0121]; See also Section V.A.1.

2. Claim 14(a): “divide a speech audio signal including an accent 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;”

See Claim 1(a).

3. Claim 14(b): “extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”

See Claim 1(b).

4. Claim 14(c): “extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”

See Claim 1(c).

5. Claim 14(d): “receive a speaker embedding for a human speaker; and;”

See Claim 1(d).

6. Claim 14(e): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

See Claim 1(e).

G. Claim 15

1. Claim 15(a): “The computing apparatus of claim 14, wherein: the speech audio signal is recorded from a voice of a user by the computing apparatus via an acoustic sensor; and”

See Claim 2(a).

2. Claim 15(b): “a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.”

See Claim 2(b).

H. Claim 16

1. Claim 16: “The computing apparatus of claim 14, wherein the speaker embedding is generated based on the speech audio signal.”

See Claim 3.

I. Claim 17

1. Claim 17: “The computing apparatus of claim 14, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.”

See Claim 4.

J. Claim 20

1. Claim 20[Pre]: “A non-transitory computer-readable storage medium, the computer-readable storage medium including instructions that when executed by a computer, cause the computer 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], [0081], [0118]-[0121]; *See also* Section V.A.1.

2. Claim 20(a): “divide a speech audio signal including an accent 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;”

See Claim 1(a).

3. Claim 20(b): “extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;”

See Claim 1(b).

4. Claim 20(c): “extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”

See Claim 1(c).

5. Claim 20(d): “receive a speaker embedding for a human speaker; and”

See Claim 1(d).

6. Claim 20(e): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

See Claim 1(e).

VI. GROUND 2: CLAIMS 9-11 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND ACCENTRON

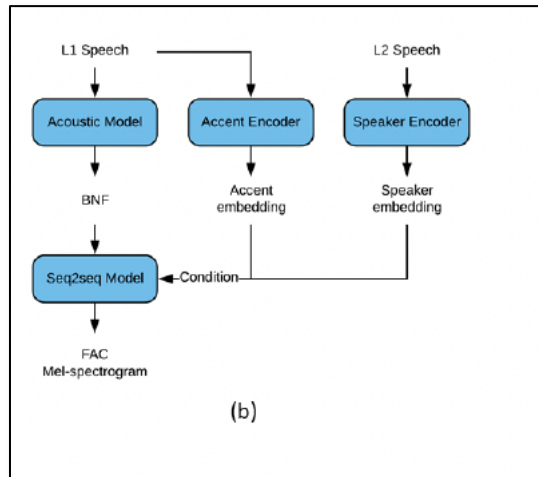
A. Claim 9

1. Claim 9: “The method for real-time correction of claim 1, wherein the generating the output chunk includes: providing the speaker embedding, the acoustic features, and the linguistic features to a synthesis module to generate a melspectrogram with the reduced accent; and providing the melspectrogram to a vocoder to generate the output chunk of the output audio signal.”

As taught in Claim 1(e), *Feinauer-Prabhavalkar-Fan* teaches generating an output chunk of an output audio signal with a reduced accent based on speaker embedding, acoustic features, and linguistic features. See Section V.A.6. While *Feinauer* does not discuss providing these values to a *synthesis module* within accent modifier 318 to generate a melspectrogram with the reduced accent; and providing the melspectrogram to a vocoder to generate the output chunk of the output audio signal, *Accentron* teaches such.

a) Accentron’s Teachings

Accentron teaches a seq2seq *synthesis module*. EX1008-Accentron, 7.



EX1008-Accentron, Fig. 1. (excerpt). *Accentron's* seq2seq model consumes bottleneck features, accent embedding, and speaker embedding “to synthesize Mel-spectrogram for an arbitrary L2 speaker.” EX1008-Accentron, 7, Table 1. Therefore, *Accentron* teaches providing the speaker embedding, the acoustic features, and the linguistic features to a synthesis module to generate a melspectrogram. As discussed in Claim 1(c), *Feinauer-Prabhavalkar-Fan* teach the PPGs contain a reduced accent. EX1003, ¶195; See Section V.A.4. Thus, in the combination with *Accentron*, a PHOSITA would have appreciated that the input PPGs (in place of BNFs) would have a *reduced accent*.

Accentron's seq2seq model is a *synthesis module* which includes a decoder. EX1008-Accentron, 7; Table 1. *Accentron's* “Mel-spectrogram is converted back to a waveform” using a WaveRNN vocoder. EX1008-Accentron, 9; 18; EX1003, ¶¶196-198. Therefore, *Accentron* teaches providing the melspectrogram to a *vocoder*.

a) Motivation to Combine the modified *Feinauer* with *Accentron* and *Fan*

In the proposed modified method, a PHOSITA would have been motivated and found it obvious to provide *Feinauer-Prabhavalkar-Fan*'s speaker embedding, acoustic features, and linguistic features to *Accentron*'s synthesis module to generate a melspectrogram with the reduced accent; and providing the melspectrogram to a vocoder to generate the output chunk of the output audio signal as taught by *Accentron* as an operable method of producing output modified speech. EX1003, ¶200. A PHOSITA would have appreciated this modification would have been part of the “accent modifier 318” corresponding to modifying a user's accent in *Feinauer* as a well-known method of modifying the source voices to match the accent of the target, achieving *Feinauer*'s stated goal of selective voice accent modification. EX1004-Feinauer, [0001], [0081], [0105], Fig. 7; EX1003, ¶201. A PHOSITA would have appreciated *Feinauer* already teaches being capable to produce a spectrogram, and would have found it simple to specifically modify the system to produce melspectrograms as taught by *Fan* and *Accentron*. EX1004-Feinauer, [0064], [0080], [0095], [0099], [0110], Fig. 7; EX1003, ¶201. A PHOSITA would have been motivated to modify *Feinauer*'s “accent modifier 318” to utilize a neural network as taught by *Accentron*'s neural network seq2seq module. EX1008-Accentron, 14. EX1003, ¶¶201-204. *Feinauer* already teaches that “features of the vocal characteristics in each category can be analyzed by a supervised neural

network” within the accent modifier 318 on the analysis side. EX1004-Feinauer, [0098]. A PHOSITA would have understood that the linguistic features would have a reduced accent and therefore the mel spectrogram will have a reduced accent. *See* Claim 1(b). A PHOSITA would have appreciated that *Fan*, already in the combination teaches using mel features when generating an output. EX1006-Fan, [0099]. A PHOSITA would have recognized descriptions related to Fan’s style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. EX1003, ¶204.

A PHOSITA would have been motivated to train the modified *Feinauer* system to operate in the manner taught by *Accentron*’s neural network seq2seq model as a *module* for the reasons presented above regarding Claim 1(b)-1(c), the discussion and motivations of which are herein incorporated by reference. *See* Section V.A.3-V.A.4; EX1003, ¶202. A PHOSITA would have been motivated to make the combination since it was well known mel-spectrograms provide a convenient representation of audio for the vocoder to reconstruct and that neural networks are commonly used in accent conversion systems and implementing a neural network in *Feinauer*’s accent modifier 318 which would have increased efficiency. EX1003, ¶202. A PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer*’s accent modifier model and *Accentron*’s neural network seq2seq module producing a melspectrogram and

vocoder) according to known speech processing methods with predictable results, further motivating the combination. EX1003, ¶¶203-204 (discussing EX1006-Fan, [0047]-[0049], Fig. 12). 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.*

B. Claim 10

1. Claim 10(a): “The method for real-time correction of claim 9, wherein: the synthesis module includes an encoder, a decoder, and a post-net module designed to improve an output of the decoder; and;”

Feinauer-Prabhavalkar-Fan in view of *Accentron* teaches the synthesis module includes an encoder, a decoder, and a post-net module designed to improve an output of the decoder.

a) Accentron's Teachings

Accentron's synthesis module (seq2seq model) includes an encoder, a decoder, and a post-net module and that the post-net module is designed to improve an output (Mel-spectrogram) of the decoder.

Table 1: Hyperparameters of the seq2seq foreign accent conversion model.

Block	Component	Parameters
Inputs	BNF	256-dimensional
	Speaker d-vector	256-dimensional
	Accent d-vector	256-dimensional
Encoder	3 × Conv layers	512 5×1 kernel with 1×1 stride; ReLU; batch norm
	2 × p-Bi-LSTM	256 cells per direction;
Attention	Location-sensitive attention	128-dim attention context; 32 31×1 attention conv kernels
Decoder	PreNet	2 × Fully-connected layer ; 256 neurons; ReLU
	2 × LSTM	1024 cells
	Linear (Mel)	1 × Fully-connect layer; 80 units; no activation
	Linear (stop token)	1 × Fully-connect layer; 1 unit; no activation
	PostNet	5 × Conv layers; 512 5×1 kernel with 1×1 stride; tanh; batch norm
Outputs	Mel-spectrogram	80-dimensional
	Stop token	2-dimensional

EX1008-Accentron, Table 1 (annotated).

During accent conversion and “to incorporate the spectral residual and improve synthesis quality,” *Accentron*’s “predicted Mel-spectrogram is passed through a post-net.” EX1008-Accentron, 15. The seq2seq model is “optimized by minimizing the Euclidean distance between the target Mel-spectrogram and the prediction before/after the post-net.” EX1008-Accentron, 15. Therefore, the *post-net module is designed to improve (optimize) an output (Mel-spectrogram) of the decoder.*

b) Motivation to Combine *Accentron*

A PHOSITA would have been motivated and found it obvious to modify *Feinauer-Prabhavalkar-Fan* with *Accentron* so that *Feinauer's* accent modifier 318 as modified by *Accentron's* seq2seq module (synthesis module) in Claim 9, includes *an encoder, a decoder, and a post-net module designed to improve an output of the decoder*. See Section VI.A.1. EX1003, ¶209.

In the proposed modified method, a PHOSITA would have been motivated and found it obvious to modify *Feinauer's* accent modifier 318 to include *an encoder, a decoder, and a post-net module designed to improve an output of the decoder*, as taught by *Accentron*. EX1003, ¶210. A PHOSITA would have appreciated this modification would have been part of the “accent modifier 318” corresponding to modifying a user’s accent in *Feinauer* as an operable method of modifying the source voices to match the accent of the target. EX1004-Feinauer, [0081], [0105], [0110], Fig. 7; EX1003, ¶¶210-211.

A PHOSITA would have appreciated that adding a step for optimizing the accent modifier 318 post net encoder by “minimizing the Euclidean distance between the target Mel-spectrogram and the prediction before/after the post-net” would have increased overall accuracy to improve an output of the accent modifier 318’s decoder as modified and taught by *Accentron*. EX1008-Accentron, 15; EX1003, ¶212.

As discussed in Claim 9, a PHOSITA would have been motivated to modify *Feinauer's* “accent modifier 318” to be a neural network. EX1003, ¶214. A PHOISTA would have been motivated to make the combination since it was well known to convert an accent to include an encoder, decoder, and a post-net module. EX1003, ¶¶213-214. A PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer's* accent modifier model and *Accentron's* seq2seq module) according to known speech processing methods with predictable results, further motivating the combination. EX1003, ¶215. (discussing EX1006-Fan, [0047], Fig. 12). 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.*

2. Claim 10(b): “generating the melspectrogram includes: processing the linguistic features with the reduced accent by the encoder to generate hidden features; combining the hidden features, the acoustic features, and the speaker embeddings to generate further features; processing the further features by the decoder and the post-net module to generate the melspectrogram.”

Feinauer-Prabhavalkar-Fan in view of *Accentron* teach generating the melspectrogram includes: processing the linguistic features with the reduced accent by the encoder to generate hidden features; combining the hidden features, the acoustic features, and the speaker embeddings to generate further features;

processing the further features by the decoder and the post-net module to generate the melspectrogram.

As discussed in Claims 1(b)-(c), *Feinauer-Prabhavalkar-Fan* teach linguistic features with the reduced accent, acoustic features, and speaker embeddings. *See* Sections V.A.4, V.A.5. *Accentron* further teaches *generating hidden features to generate further features then processing the further features by the decoder and the post-net module to generate the melspectrogram.*

The linguistic features (BNF) is input into the encoder, where the “encoder converts a BNF sequence to a hidden representation sequence.” EX1008-Accentron, 8, 14. Therefore, *Accentron* teaches *processing the linguistic features ... by the encoder to generate hidden features.* As discussed in Claim 1(c), *Feinauer-Prabhavalkar-Fan* teach the PPGs contain a reduced accent, therefore in the combination with *Accentron* a PHOSITA would have appreciated that the input PPGs (in place of BNFs) would be input *with the reduced accent.* EX1003, ¶218; *See* Section V.A.4.

Then, “to condition the decoder on the voice identity and the accent of the target speaker, we concatenate the target speaker’s speaker embedding and accent embedding to the hidden representation.” EX1008-Accentron, 12, 10. Therefore, *Accentron* teaches *combining the hidden features, the acoustic features, and the speaker embeddings to generate further features.*

After concatenation, the hidden representations from the bottleneck features, accent embedding, and speaker embedding is fed “to the decoder to produce the predicted FAC Mel-spectrogram.” EX1008-Accentron, 12-13, 10. Therefore, *Accentron* teaches *processing the further features by the decoder*.

The seq2seq model is “optimized by minimizing the Euclidean distance between the target Mel-spectrogram and the prediction before/after the post-net.” EX1008-Accentron, 15. The decoder “produces an 80-dimensional Mel-spectrogram as the prediction of the L2 speech” with a “25ms window and 10ms shift.” EX1008-Accentron, 14, 18. Therefore, *Accentron* teaches *processing the further features by the decoder and the post-net module to generate the melspectrogram*.

a) Motivation to Combine *Accentron*

A PHOSITA would have been motivated and found it obvious to modify *Feinauer-Prabhavalkar-Fan* with *Accentron* so that *Feinauer’s* accent modifier 318 as modified by *Accentron’s* seq2seq module (synthesis module) in Claim 9, includes processing the linguistic features with the reduced accent by the encoder to generate hidden features; combining the hidden features, the acoustic features, and the speaker embeddings to generate further features; processing the further features by the decoder and the post-net module to generate the melspectrogram. *See* Section VI.A.1. EX1003, ¶222. A PHOSITA would have appreciated this modification

would have been part of the “accent modifier 318” corresponding to modifying a user’s accent in *Feinauer* as an operable method of modifying the source voices to match the accent of the target. EX1004-Feinauer, [0081], [0105], [0110], Fig. 7; EX1003, ¶222.

In the proposed modified method, a PHOSITA would have been motivated and found it obvious to modify *Feinauer*’s accent modifier 318 to process the linguistic features with the reduced accent by the encoder to generate hidden features as taught by *Accentron*. EX1003, ¶222. As discussed in Claim 9, a PHOSITA would have been motivated to modify *Feinauer*’s “accent modifier 318” to be a neural network. EX1003, ¶222. It was well known in the prior art that neural networks generate hidden features. EX1003, ¶222. A PHOSITA would have appreciated that by generating hidden features, this would improve the overall accent conversion time in *Feinauer*. EX1003, ¶222.

A PHOSITA would have been motivated to modify *Feinauer*’s accent modifier 318 to combining the hidden features, acoustic features, and speaker embeddings to generate further features as taught by *Accentron*, improving the overall accent conversion accuracy in *Feinauer*. EX1003, ¶222.

After the concertation, a PHOSITA would have been motivated to modify *Feinauer*’s accent modifier 318 to process the further features by the decoder and the post-net module to generate the melspectrogram as taught by *Accentron*.

EX1003, ¶223. As discussed in Claim 9, a PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer's* accent modifier model and *Accentron's* seq2seq module producing a melspectrogram and vocoder) according to known speech processing methods with predictable results, further motivating the combination. *Id.* See Section VI.A.1.

A PHOSITA would have been motivated to make the combination since these were all well-known concepts in the accent conversion space. EX1003, ¶224. A PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer's* accent modifier model and *Accentron's* seq2seq module) according to known speech processing 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.*

C. Claim 11

1. Claim 11: “The method for real-time correction of claim 10, wherein: the acoustic features are split into a stream of acoustic features chunks corresponding to the chunks in the stream of input chunks; the linguistic features are split into a stream of linguistic features chunks corresponding to the chunks in the stream of input chunks; the melspectrogram is split into a stream of melspectrogram chunks corresponding to the chunks in the stream of input chunks; and a melspectrogram chunk of the stream of melspectrogram chunks is generated based on the following: an acoustic features chunk of the stream of acoustic features chunks and acoustic features context including the pre-determined number of acoustic features frames belonging to acoustic features chunks preceding the acoustic features chunk in the stream of acoustic features chunks; and; a linguistic features chunk of the stream of melspectrogram chunks and linguistic features context including the pre-determined number of linguistic features frames belonging to linguistic features chunks preceding the linguistic features chunk in the stream of acoustic features chunks.”

The combination of *Feinauer-Prabhavalkar-Fan-Accentron* teaches claim 11. Specifically, as discussed regarding Claim 1(b), *Feinauer-Prabhavalkar-Fan* teaches extracting acoustic features from input chunks and context. *Supra*, Section V.A.3. Further, as discussed regarding Claim 1(c), *Feinauer-Prabhavalkar-Fan* teaches extracting linguistic features from input chunks and context. *Supra*, Section V.A.4. Claim 1(d-e) explains that acoustic features in chunks with context – output chunks. *Supra*, Section V.A.5-V.A.6. As explained regarding Claims 9-10, *Accentron* further teaches generating output chunks comprising melspectrograms. *Supra*, Section VI.A.1-VI.B.2.

Specifically, in the combination, *Prabhavalkar's* “chunk-based” processing technique is utilized, consuming all frames in a first input chunk, processing them and outputting results “before it consumes the next chunk.” EX1005-*Prabhavalkar*, 18:20-23, 17:33-46; *See* Section V.A.6. Here, a “linguistic features chunk” is mapped as the linguistic features derived from the input chunk in the combined *Feinauer-Prabhavalkar* system, while the a “acoustic features chunk” is mapped as the acoustic features derived from the same chunk. EX1003, ¶226. Thus, the acoustic features of the first chunk are consumed with the corresponding linguistic features of the first chunk to produce the first output chunk, which comprises melspectrograms per Claim 9. *See* Section VI.A.1; EX1003, ¶226.

VII. GROUND 3: CLAIMS 6 AND 19 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND SINGH

A. Claim 6

1. Claim 6: “The method for real-time correction of claim 1, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.”

As discussed in Claim 1(c), 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 module extracting linguistic features including PPGs for *Feinauer's* real-time accent conversion.

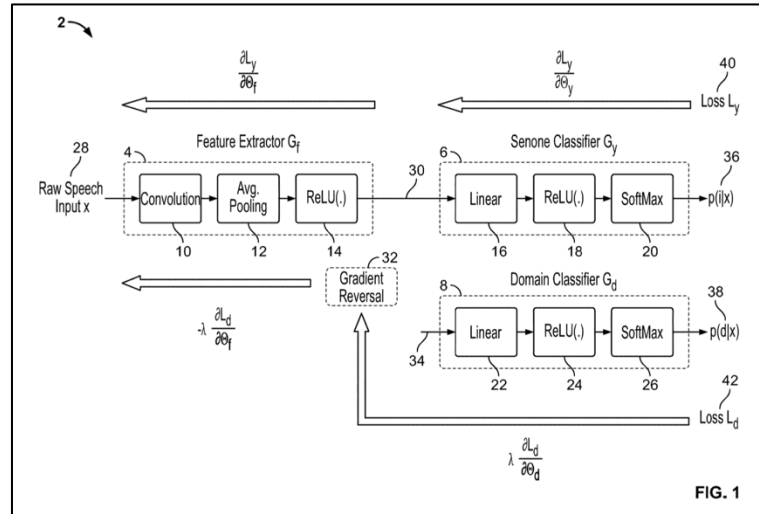
EX1003, ¶227. *Fan*'s voice feature extraction model is a "deep neural network" and outputs the linguistic features with the reduced accent. EX1006-Fan, [0085]; EX1003, ¶227; Claim 1(c).

Feinauer-Prabhavalkar-Fan in view of *Singh* further teach *the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.*

a) *Singh's* Teachings

Singh teaches methods for training neural networks to learn features from raw speech. EX1007-Singh, Abstract, [0001], [0009]. *Singh* envisions training these networks using a loss function. EX1007-Singh, [0018], [0020]-[0023]. Thus, when used to train the linguistic feature extraction module (LFEM) of claim 1, *Feinauer-Prabhavalkar-Fan-Singh* teaches the *linguistic feature "neural network being trained using a loss function[.]"* EX1003, ¶¶228-230, *See* Section V.A.4. *Singh* additionally describes embodiments recognizing English spoken in a variety of accents, including a British, Australian, or American accent. EX1007-Singh, [0009], [0023], [0034], EX1003, ¶¶228-229. Within the American accent, *Singh's* system is trained to account for at a plurality of American sub-accent (corresponding to "8 major dialect regions of the United States of America"), thus further teaching a *"neural network being trained... to reduce... contributions due to a further accent present in the audio data."* *Id.*

Specifically, *Singh's* neural network architecture includes a “feature extractor 4, a label (or senone) classifier 6, and a domain classifier 8.” EX1007-Singh, [0018].



EX1007-Singh, [0008], [0020], Fig. 1. *Singh's* feature extraction neural network is trained using a multi-accent speech corpus (Voxforge dataset) comprising “6,300 sentences, 10 sentences spoken by each of the 630 speakers for 8 major dialect regions” of the USA. EX1007-Singh, [0023]; EX1003, ¶229. In examples, an utterance spoken with an American source accent may be adapted to a target British accent. EX1003, ¶¶228-231. *Singh's* feature extraction neural network is trained to reduce contributions due to the speaker's American accent. EX1003, ¶¶228-231. Thus, when *Singh's* approach is used to train the *Feinauer-Prabhavalkar-Fan* LFEM, *Feinauer-Prabhavalkar-Fan-Singh* teaches linguistic feature “neural network being trained... to reduce... contributions due to a further accent present in the audio data.” EX1003, ¶231.

b) Motivation to Combine *Singh*

A PHOSITA would have been motivated to utilize *Singh*'s teachings regarding neural network training to train *Feinauer-Prabhavalkar-Fan*'s LFEM in order to increase the accuracy and efficiency of the accent modification. EX1003, ¶232. It was well known to use a domain loss function in order to train neural networks. EX1003, ¶232.

A PHOSITA would have appreciated that such a combination would have been a combination of known elements that would yield predictable results, including a neural network trained to a desired threshold of accuracy for reducing contributions due to a further spoken accent. EX1003, ¶233. A PHOSITA would have appreciated a REOS for this modification because the combination required only reprogramming of *Feinauer*'s software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶233.

B. Claim 19

1. Claim 19: "The computing apparatus of claim 14, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data."

See Claim 6.

VIII. GROUND 4: CLAIM 7 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND WU

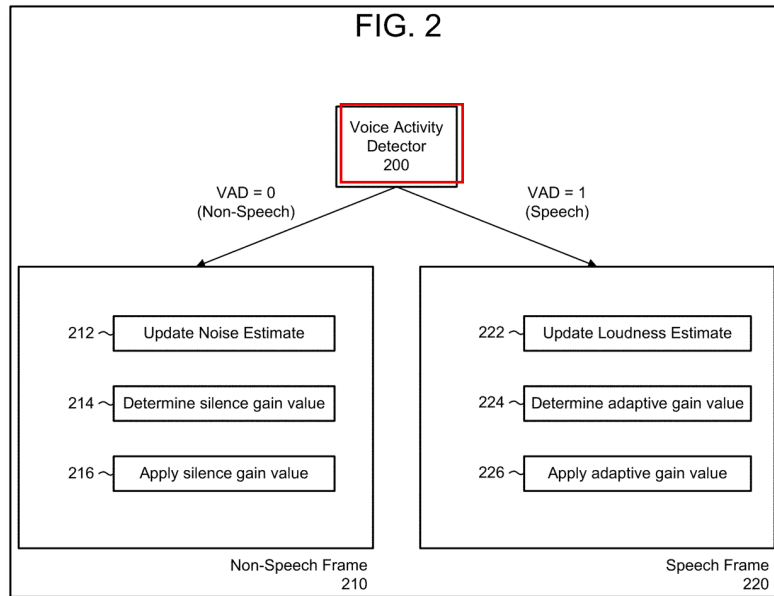
A. Claim 7

1. Claim 7: “The method for real-time correction of claim 1, wherein the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.”

As discussed in Claim 1(b), *Feinauer-Prabhavalkar-Fan* teaches acoustic features including tone, *pitch*, volume, rhythm, and speaking speed. A PHOSITA would have understood that the volume of a speech audio signal corresponds to energy of the speech audio signal. EX1003, ¶235 (Mr. Schmandt explaining the direct mathematical correlation between volume and energy). See Section V.A.3. *Feinauer-Prabhavalkar-Fan* in view of *Wu* further teach *a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.*

a) Wu’s Teachings

Wu teaches a voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal, namely “voice activity detector (VAD) 200 that may determine whether voice activity (e.g., speech) is detected.” EX1009-Wu, 5:46-47, 5:55-56, Fig. 2. VAD 200 is illustrated below in Fig. 2:



EX1009-Wu, Fig. 2.

Wu teaches that a VAD *value* of 1 demonstrates that “voice activity is detected” (*indicating presence of the human voice in the speech audio signal*) so “device 110 may classify an audio frame as corresponding to speech[.]” *Wu*, 7:20-25. Conversely, a VAD *value* of 0 demonstrates that “voice activity is not detected” (*indicating absence of a human voice in the speech audio signal*) and the current audio frame is classified as “non-speech” instead. EX1009-Wu, 7:20-25, 7:39-40; 6:60-65; EX1003, ¶237. *Wu* further teaches that VAD data generated by VAD 200 “may classify a **first audio frame that is below an energy threshold as non-speech**” “and/or may classify a **second audio frame that is above the energy threshold** as corresponding to speech....” EX1009-Wu, 7:31-38, Fig. 3.

b) Motivation to Combine Wu

A PHOSITA would have found it obvious and been motivated to further modify *Feinauer-Prabhavalkar-Fan* to include *Wu*'s VAD to “not only improve the performance of downstream components by discarding non-speech signals, but also significantly reduces the overall computational cost due to its relatively small size.” EX11026-Chang, 1; EX1003, ¶238. A PHOSITA would have also appreciated including a VAD to “eliminate dead space” and “improve utilization efficiency of...storage resources.” See EX1017-Garrod, 10:30-33; EX1003, ¶238. A PHOSITA would have appreciated that such a modification merely entails using a known technique (voice activity detection to determine speech versus non-speech in audio data, per *Wu*) to improve similar devices (voice conversion systems). EX1003, ¶238.

It was understood prior to the '457 Patent that “VAD algorithms are well known in the art.” EX1003, ¶238. A PHOSITA would have appreciated a REOS for this modification because the combination required only reprogramming of *Feinauer's* software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶238.

IX. GROUND 5: CLAIMS 12-13 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND BONGIOVI

A. Claim 12

1. Claim 12: “The method for real-time correction of claim 1, further comprising, prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust one or more characteristics of the speech audio signal to improve extraction of the linguistic features and the acoustic features.”

The '457 Patent describes a digital signal processing module 1002 “used to control loudness of output audio signal 308. For example, digital signal processing module 1002 may *auto-gain loudness of speech audio signal 202 before pipeline 300 processing and then, based on a user setting, restore or not to restore level of loudness of output audio signal 308* to corresponding level of loudness of speech audio signal 202.” EX1001, 12:52-59. *Feinauer-Prabhavalkar-Fan* modified to include *Bongiovi* teach *further comprising, prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust one or more characteristics of the speech audio signal to improve extraction of the linguistic features and the acoustic features.*

a) *Feinauer-Prabhavalkar-Fan’s Teachings*

Feinauer-Prabhavalkar-Fan in view of *Bongiovi* teach prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing

module to adjust one or more characteristics (the loudness) of the speech audio signal.

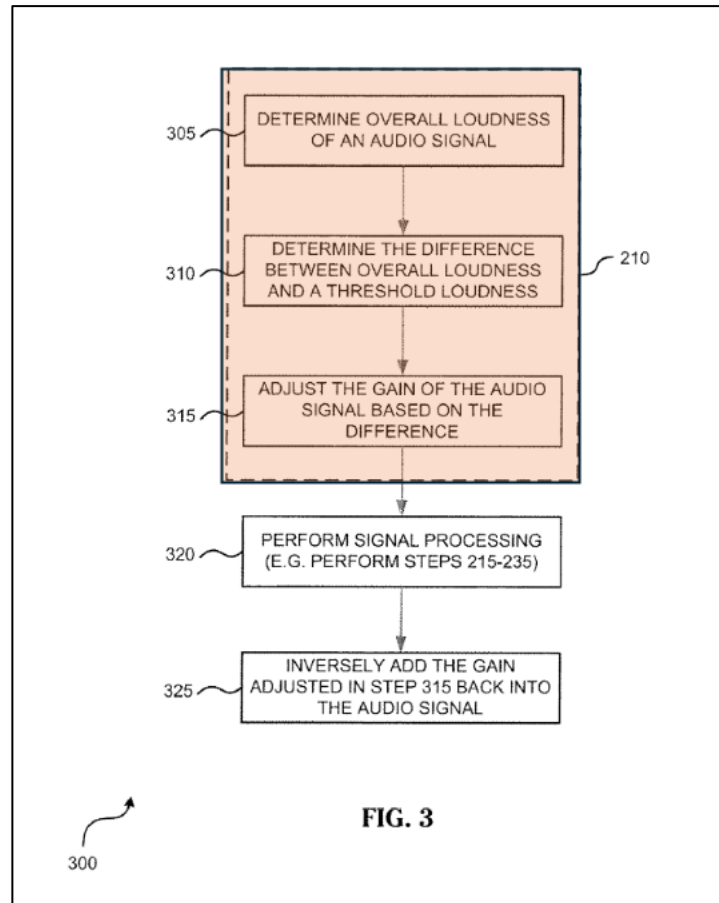
“prior to dividing the speech audio signal”

As discussed in Claim 1(a), *Feinauer-Prabhavalkar* teaches dividing a speech utterance into chunks of 5 frames each. *See* Section V.A.2 *citing* EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. As will be discussed, a PHOSITA would have implemented *Bongiovi*’s teachings of pre-processing the audio’s loudness prior to *Feinauer-Prabhavalkar* dividing a speech utterance into “chunks.”

“processing the speech audio signal...”

b) *Bongiovi*’s Teachings

Bongiovi teaches digitally processing an audio signal using a computer chip, such as...digital signal processor, or DSP. EX1010-Bongiovi, Abstract, [0052]. A PHOSITA would have recognized a DSP satisfies the claimed *digital signal processing module*. EX1003, ¶243.



EX1010-Bongiovi, Fig. 3. (annotated).

When the audio comes into *Bongiovi's* device, the overall loudness of the received signal is determined. EX1010-Bongiovi, [0047] (Step 305); EX1003, ¶244. After this, “the determined loudness is compared with a threshold loudness,” and based on the determination result, “the gain of the raw audio signal is *adjusted* based on the comparison done in step 310. If the loudness is above the threshold, the gain of the audio signal is reduced...If the loudness is below the threshold, the [g]ain of the audio signal is increased.” EX1010-Bongiovi, [0047]-[0049]. A PHOSITA would have understood that loudness is a *speech characteristic of an audio signal*.

EX1003, ¶244. *Bongiovi* then teaches “[a]fter the gain is adjusted in step 315, the signal can be further processed in step 320.” EX1010-Bongiovi, [0050].

“to improve extraction of the linguistic features and the acoustic features”

Bongiovi teaches performing steps 215-235 adjusting the volume of the input audio signal to reduce distortion. EX1010-Bongiovi, [0046]. A PHOSITA would have appreciated that reduced distortion would have *improved* the quality for *extraction of the linguistic features and the acoustic features*. EX1003, ¶245.

c) Motivation to Combine *Bongiovi*

A PHOSITA would have found it obvious and would have been motivated to use *Bongiovi*’s teachings of using a digital signal processor (“DSP”) to adjust one or more characteristics (the loudness) of the speech audio signal prior to *Feinauer-Prabhavalkar*’s processing steps of dividing speech utterances. EX1003, ¶246; EX1010-Bongiovi, [0046], [0050], [0052]; EX1005-Prabhavalkar, 6:56-7:4, 15:16-22. A PHOSITA would have appreciated that reducing distortion of *Feinauer*’s speech audio signal before further processing would have resulted in improved accent modification quality, speed, and computational efficiency by reducing the amount of unwanted signal being processed. EX1003, ¶247. A PHOSITA would have appreciated that such a combination would have been a combination of known elements (*Bongiovi*’s DSP into the modified *Feinauer* accent modification system) that would yield the predictable results of utilizing a DSP to adjust characteristics of

a speech audio signal. EX1003, ¶247. A PHOSITA would have appreciated a REOS for the modification because the combination required only reprogramming of *Feinauer's* software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶247.

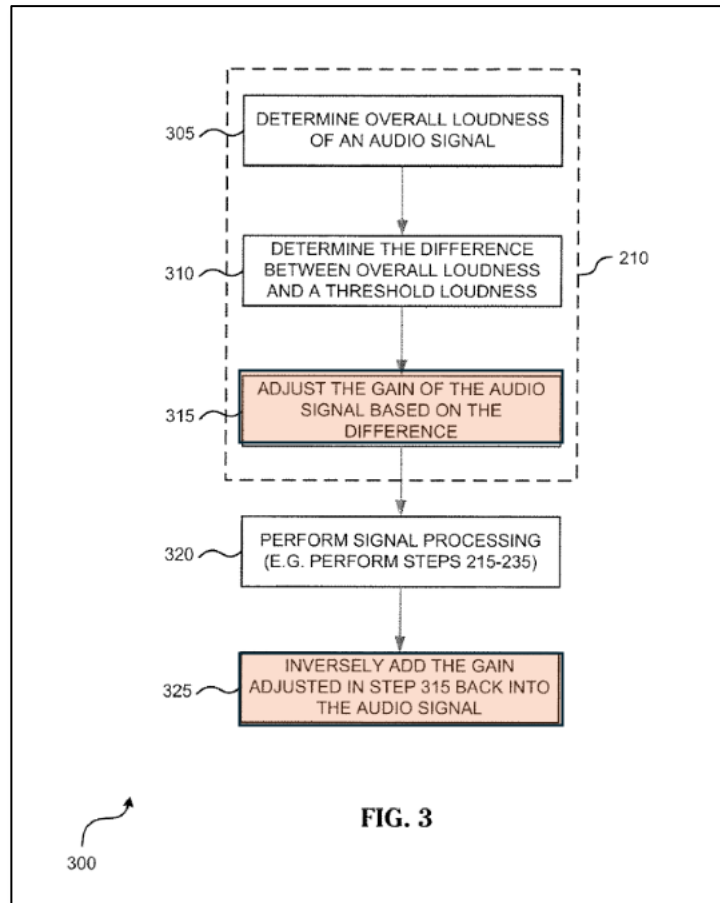
A PHOSITA would have been motivated to modify *Feinauer* such that *Bongiovi's* teachings of pre-processing the audio's loudness would be implemented prior to *Feinauer-Prabhavalkar's* dividing a speech utterance into chunks. EX1005-Prabhavalkar, 6:56-7:4, 15:16-22; EX1004-Feinauer, [0057]; EX1003, ¶248. *Bongiovi* already teaches this. EX1010-Bongiovi, [0050]. ("After the gain is adjusted in step 315, the signal can be further processed in step 320."); EX1003, ¶248. A PHOSITA would have appreciated that such a combination would have been obvious to implement in *Feinauer's* accent conversion system, since *Bongiovi* already teaches such. EX1003, ¶249. A PHOSITA would have appreciated that such a combination would have been a combination of known elements (pre-processing audio prior to dividing the audio into chunks) that would yield predictable results, including improved signal quality. EX1003, ¶250. It was well known to pre-process audio before further speech processing to desirably improve performance. EX1003, ¶251. A PHOSITA would have appreciated a REOS for this modification because the combination required only reprogramming of *Feinauer's* software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶251.

B. Claim 13

1. Claim 13: “The method for real-time correction of claim 1, further comprising: prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust a loudness of the speech audio signal from a first level to a second level; after generating the output chunk of the output audio signal, processing the output chunk by the digital signal processing module, to adjust the loudness of the output audio signal to the first level.”

Feinauer-Prabhavalkar-Fan in view of *Bongiovi* teaches *prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust a loudness of the speech audio signal from a first level to a second level. See Section IX.A.1. Further, the combination teaches after generating the output chunk of the output audio signal, processing the output chunk by the digital signal processing module, to adjust the loudness of the output audio signal to the first level.*

Bongiovi teaches “[a]fter the gain is adjusted in step 315, the signal can be further processed in step 320.” EX1010-Bongiovi, [0050]. After this, in “the gain add[ed] or subtracted from the audio signal in step 315 is subtracted or added back in by an amount that is exactly inverse to the amount taken or added in step 315. In this way, **the original volume control is mimicked** at the end of method 300.” *Id.* Therefore, the *digital signal processing module is adjust[ing] the loudness of the output audio signal to the first level.* EX1003, ¶¶253-254.



EX1010-Bongiovi, Fig. 3. (annotated).

a) Motivation to Combine *Bongiovi*

A PHOSITA would have found it obvious and would have been motivated to use *Bongiovi*'s teachings of using a digital signal processor ("DSP") to adjust a loudness of the speech audio signal from a first level to a second level prior to *Feinauer-Prabhavalkar-Fan*'s audio processing steps and to adjust to the first loudness after *Feinauer-Prabhavalkar-Fan* produces the output chunk. EX1003, ¶255; EX1010-Bongiovi, [0046], [0050]-[0052], Fig. 3; EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22, 17:33-51; EX1004-*Feinauer*, [0057].

A PHOSITA would have been motivated to modify *Feinauer-Prabhavalkar-Fan* such that *Bongiovi*'s teachings to adjust the loudness back to the first level after *Feinauer-Prabhavalkar-Fan* produces the output chunk. EX1003, ¶¶255-256. *Bongiovi* already teaches this step is done after signal processing. EX1010-Bongiovi, Fig. 3, [0050]. As discussed in Claim 1(a), each output chunk is 5 frames, therefore, a PHOSITA would have appreciated that processing signal loudness after generating the chunk would have been simpler. *See* Section V.A.1. EX1003, ¶256. A PHOSITA would have appreciated that such a combination would have been a combination of known elements (adjusting volume after generating an output chunk) that would yield predictable results, including improved efficiency and user experience. EX1003, ¶257. A PHOSITA would have appreciated a REOS for this modification because the combination required only reprogramming of *Feinauer*'s software using ordinary skill of a PHOSITA and hardware already present, including the use of a DSP to adjust characteristics of a speech audio signal. EX1003, ¶257. A PHOSITA would have appreciated that such a combination would have been a combination of known elements (processing audio loudness after generating output chunks) that would yield predictable results. EX1003, ¶257.

X. GROUND 6: CLAIM 5 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND OSUNA

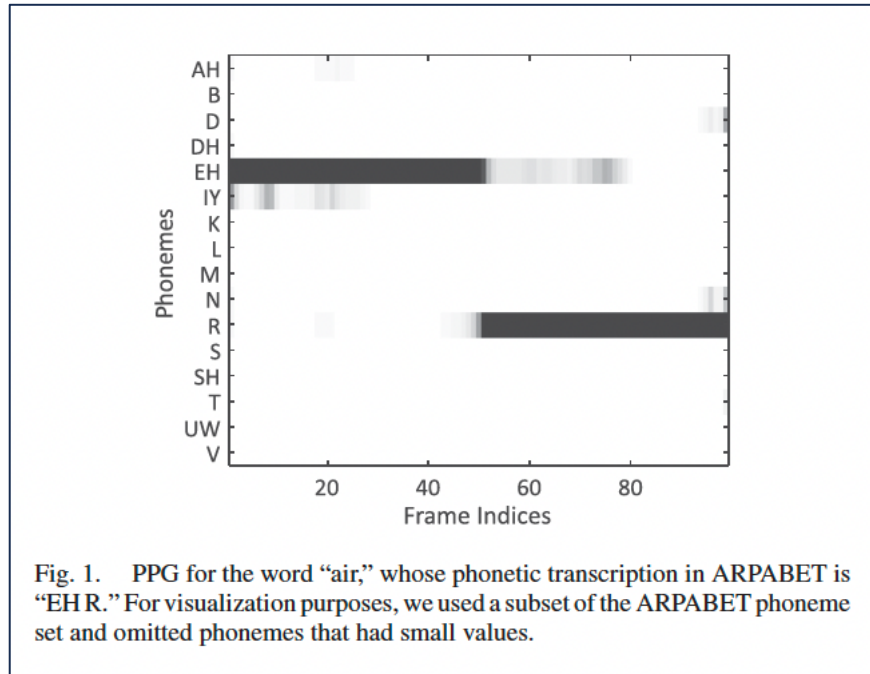
A. Claim 5

1. Claim 5: “The method for real-time correction of claim 1, wherein the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library.”

As discussed regarding Claim 1(c), *Feinauer-Prabhavalkar-Fan* teaches extracting linguistic features using models trained from speakers (per *Feinauer*) including PPGs (per *Fan*). See Section V.A.4 citing EX1004-Feinauer, [0083], [0094], [0099]; EX1006-Fan, [0007], [0085]-[0086], Fig. 5; EX1003, ¶259. Further, *Feinauer* explains its models use a “dictionary” that identifies a replacement from one accent for a word to another accent.” EX1004-Feinauer, [0009], [0038], [0064], [0066], [0081], [0113]. *Feinauer* does not expressly teach using a standardized phonetic dictionary.

a) Osuna’s Teachings

Osuna teaches using the ARPAbet standard phoneme representation for its PPG, as shown in the PPG that is depicted in Fig 2.



EX1011-Osuna, Fig. 1. EX1003, ¶260. A PHOSITA would understand that any PPG will have a representation of phonemes because the first P in PPG refers to phonemes. EX1003, ¶260. A PHOSITA would realize that ARPAbet was a well-known standard way of representing phonemes. EX1003, ¶260.

b) Motivation to Combine *Osuna*

A PHOSITA would have found it obvious and would have been motivated to use *Osuna*’s teachings of using an ARPAbet dictionary to *Feinauer*’s processing steps of determining PPGs. EX1003, ¶261. A PHOSITA would have appreciated that *Osuna*’s dictionary would have desirably improved the system by efficiently allowing the modified *Feinauer-Prabhavalkar-Fan* system’s PPG to have a representation of phonemes by using an ARPAbet dictionary. EX1003, ¶261. A PHOSITA would have appreciated that such a combination would have been a

combination of known elements (*Osuna's* phonetic ARPAbet dictionary into the modified *Feinauer* accent modification system) that would yield predictable results, including the use of a phonetic dictionary to determine linguistic content. EX1003, ¶262. A PHOSITA would have appreciated a REOS for the modification because the combination required only reprogramming of *Feinauer's* software using ordinary skill of a PHOSITA and hardware already present. EX1003, ¶262.

XI. GROUND 7: CLAIM 18 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, WU, AND OSUNA

A. Claim 18

1. Claim 18: "The computing apparatus of claim 14, wherein: the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library; and the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal."

See Claims 5 (discussing a *standardized phonetic dictionary*) and 7 (discussing a *VAD*).

XII. 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

XIII. 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

The '457 Patent is presently the subject of a patent infringement lawsuit filed against Petitioner in the United States District Court for the Northern District of California. SANAS.AI INC. v. KRISP TECHNOLOGIES, INC., Case No. 3:25-cv-05666-TSH.

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 <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 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

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. 11,715,457 for Claims 1-20

Claim Designation	Claim Language
Claim 1(P)	A method for real-time correction of an accent in a speech audio signal, the method being implemented by a computing device and comprising:
Claim 1(a)	dividing the 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 1(b)	extracting, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;
Claim 1(c)	extracting, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;
Claim 1(d)	receiving a speaker embedding for a human speaker; and
Claim 1(e)	generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.
Claim 2(P)	The method for real-time correction of claim 1, wherein:
Claim 2(a)	the speech audio signal is recorded from a voice of a user by the computing device via an acoustic sensor; and
Claim 2(b)	a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.
Claim 3	The method for real-time correction of claim 1, wherein the speaker embedding is generated based on the speech audio signal.
Claim 4	The method for real-time correction of claim 1, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.
Claim 5	The method for real-time correction of claim 1, wherein the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library.

Claim Designation	Claim Language
Claim 6	The method for real-time correction of claim 1, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.
Claim 7	The method for real-time correction of claim 1, wherein the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.
Claim 8	The method for real-time correction of claim 1, wherein the context is a pre-determined number of the frames belonging to chunks preceding the input chunk in the stream of input chunks.
Claim 9(P)	The method for real-time correction of claim 1, wherein the generating the output chunk includes:
Claim 9(a)	providing the speaker embedding, the acoustic features, and the linguistic features to a synthesis module to generate a melspectrogram with the reduced accent; and
Claim 9(b)	providing the melspectrogram to a vocoder to generate the output chunk of the output audio signal.
Claim 10(P)	The method for real-time correction of claim 9, wherein:
Claim 10(a)	the synthesis module includes an encoder, a decoder, and a post-net module designed to improve an output of the decoder; and
Claim 10(b)	generating the melspectrogram includes:
Claim 10(c)	processing the linguistic features with the reduced accent by the encoder to generate hidden features;
Claim 10(d)	combining the hidden features, the acoustic features, and the speaker embeddings to generate further features;
Claim 10(e)	processing the further features by the decoder and the post-net module to generate the melspectrogram.
Claim 11(P)	The method for real-time correction of claim 10, wherein:
Claim 11(a)	the acoustic features are split into a stream of acoustic features chunks corresponding to the chunks in the stream of input chunks;

Claim Designation	Claim Language
Claim 11(b)	the linguistic features are split into a stream of linguistic features chunks corresponding to the chunks in the stream of input chunks;
Claim 11(c)	the melspectrogram is split into a stream of melspectrogram chunks corresponding to the chunks in the stream of input chunks; and
Claim 11(d)	a melspectrogram chunk of the stream of melspectrogram chunks is generated based on the following:
Claim 11(e)	an acoustic features chunk of the stream of acoustic features chunks and acoustic features context including the pre-determined number of acoustic features frames belonging to acoustic features chunks preceding the acoustic features chunk in the stream of acoustic features chunks; and
Claim 11(f)	a linguistic features chunk of the stream of melspectrogram chunks and linguistic features context including the pre-determined number of linguistic features frames belonging to linguistic features chunks preceding the linguistic features chunk in the stream of acoustic features chunks.
Claim 12	The method for real-time correction of claim 1, further comprising, prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust one or more characteristics of the speech audio signal to improve extraction of the linguistic features and the acoustic features.
Claim 13(P)	The method for real-time correction of claim 1, further comprising:
Claim 13(a)	prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust a loudness of the speech audio signal from a first level to a second level;
Claim 13(b)	after generating the output chunk of the output audio signal, processing the output chunk by the digital signal processing module, to adjust the loudness of the output audio signal to the first level.
Claim 14(P)	A computing apparatus comprising:
Claim 14(a)	a processor; and

Claim Designation	Claim Language
Claim 14(b)	a memory storing instructions that, when executed by the processor, configure the apparatus to:
Claim 14(c)	divide a speech audio signal including an accent 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 14(d)	extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;
Claim 14(e)	extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;
Claim 14(f)	receive a speaker embedding for a human speaker; and
Claim 14(g)	generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.
Claim 15(P)	The computing apparatus of claim 14, wherein:
Claim 15(a)	the speech audio signal is recorded from a voice of a user by the computing apparatus via an acoustic sensor; and
Claim 15(b)	a delay between a first timestamp and a second timestamp is between 40 milliseconds and 300 milliseconds, wherein the first timestamp corresponds to a time when the chunk of the speech audio signal is recorded and the second timestamp corresponds to a time when the output chunk is generated.
Claim 16	The computing apparatus of claim 14, wherein the speaker embedding is generated based on the speech audio signal.
Claim 17	The computing apparatus of claim 14, wherein the speaker embedding is pretrained based on audio data including a recorded speech of a target speaker having a further accent.
Claim 18(P)	The computing apparatus of claim 14, wherein:
Claim 18(a)	the linguistic features include one of the following: phonetic posteriorgrams with a standardized phonetic dictionary and phonetic posteriorgrams with a data-driven phonetic library; and

Claim Designation	Claim Language
Claim 18(b)	the acoustic features include a pitch of the speech audio signal, an energy of the speech audio signal, and a value of a voice activity detector, the voice activity detector indicating absence of a human voice in the speech audio signal or presence of the human voice in the speech audio signal.
Claim 19	The computing apparatus of claim 14, wherein the linguistic features extraction module includes a neural network trained based on audio data to output the linguistic features with the reduced accent, the neural network being trained using a loss function to reduce, in the linguistic features, contributions due to a further accent present in the audio data.
Claim 20(P)	A non-transitory computer-readable storage medium, the computer-readable storage medium including instructions that when executed by a computer, cause the computer to:
Claim 20(a)	divide a speech audio signal including an accent 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 20(b)	extract, by an acoustic features extraction module from the input chunk and a context associated with the input chunk, acoustic features;
Claim 20(c)	extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;
Claim 20(d)	receive a speaker embedding for a human speaker; and
Claim 20(e)	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. 11,715,457 (“457 Patent”)
Exhibit 1002	Prosecution History of U.S. Patent No. 11,715,457 (“457 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. 2019/0325861 to Singh et al. (“ <i>Singh</i> ”)
Exhibit 1008	<i>Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning</i> to Zhao et al. (“ <i>Accentron</i> ”)
Exhibit 1009	U.S. Patent No. 11,164,592 to Wu et al. (“ <i>Wu</i> ”)
Exhibit 1010	U.S. Patent Application Publication No. 2008/0219459 to Bongiovi et al. (“ <i>Bongiovi</i> ”)
Exhibit 1011	G. Zhao et. al, “Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion” to Zhao et al, IEEE, October 2019 (“ <i>Osuna</i> ”).
Exhibit 1012	Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition to Ding et al. (“ <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	U.S. Patent No. 11,138,334 to Garrod et al. (“ <i>Garrod</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://caseyocallaghan.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	X. Li, Deep Segmental Phonetic Posterior-grams Based Discovery of non-Categories in L2 English Speech, arXiv, February 1, 2020. (“ <i>Li</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	J. Shen, Natural TTS Synthesis By Conditioning Wavenet on Mel Spectrogram Predications, arXiv, February 16, 2018 (“ <i>Shen</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	U.S. Patent Appl. No 2020/0134425 to Yoo et al. (“ <i>Yoo</i> ”)
Exhibit 1040	A. Tripathi et al, Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition, arXiv, May 21, 2018 (“ <i>Tripathi</i> ”)
Exhibit 1041	U.S. Patent No. 11,632,456 to Mevorah et al. (“ <i>Mevorah</i> ”)
Exhibit 1042	WO Publ. No. 2021/207131 to Jin et al. (“ <i>Jin</i> ”)
Exhibit 1043	U.S. Patent No. 11,132,997 to Pereira et al. (“ <i>Pereira</i> ”)
Exhibit 1044	U.S. Patent Appl. No. 2013/0282373 to Visser et al. (“ <i>Visser</i> ”)
Exhibit 1045	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 1046	Mina Ching Internet Archive Declaration

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 13590 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 *Inter Partes* Review and Exhibits were provided via Federal Express to the Patent Owner by serving the correspondence address of record for the '457 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