

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

DECLARATION OF CHRISTOPHER SCHMANDT

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS.....	11
A.	EDUCATIONAL BACKGROUND AND PROFESSIONAL EXPERIENCE	11
II.	METHODOLOGY AND MATERIALS CONSIDERED	15
III.	OVERVIEW AND LEGAL STANDARDS	18
A.	PERSON OF ORDINARY SKILL IN THE ART	19
B.	OBVIOUSNESS	20
C.	ANALOGOUS ART.....	25
D.	CLAIM CONSTRUCTION	26
IV.	LEVEL OF A PERSON OF ORDINARY SKILL.....	26
V.	OVERVIEW OF TECHNOLOGY	27
A.	AUDIO CHUNKS	27
B.	ACOUSTIC FEATURES	28
C.	STRESS OR ACCENT.....	29
1.	<i>Pitch or Fundamental Frequency (F0)</i>	30
2.	<i>Energy or Volume</i>	31
3.	<i>Voice Activity Detection (VAD)</i>	32
D.	LINGUISTIC FEATURES	33
1.	<i>Phonetic Posteriorgrams (PPGs)</i>	35
E.	PRIOR ART ACCENT CONVERSION SYSTEMS	38
1.	<i>Encoder and Decoder</i>	40
2.	<i>Mel Spectrograms</i>	41
3.	<i>Vocoder</i>	44
VI.	OVERVIEW OF THE '457 PATENT.....	45
A.	SUMMARY OF THE '457 PATENT.....	45
B.	FIELD OF ENDEAVOR OF THE '457 PATENT.....	46
C.	PROBLEM PURPORTEDLY SOLVED BY THE INVENTORS OF THE '457 PATENT	47
D.	FILE HISTORY OF '457 PATENT	48
VII.	OVERVIEW OF THE PRIOR ART.....	48
A.	OVERVIEW OF FEINAUER.....	49
B.	OVERVIEW OF FAN.....	51
C.	OVERVIEW OF PRABHAVALKAR	55

D.	OVERVIEW OF ACCENTRON.....	58
E.	OVERVIEW OF SINGH	60
F.	OVERVIEW OF BONGIOVI	61
G.	OVERVIEW OF WU	63
H.	OVERVIEW OF OSUNA	64
VIII.	SUMMARY OF UNPATENTABILITY	65
IX.	OPINIONS REGARDING GROUND I: <i>FEINAUER, PRABHAVALKAR, AND FAN</i>.....	66
A.	CLAIM 1	66
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>66</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>69</i>
a.	Feinauer’s Teachings.....	69
b.	Prabhavalkar’s Teachings.....	73
c.	Motivation to Combine Feinauer and Prabhavalkar.....	75
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>79</i>
a)	Feinauer’s Teachings.....	80
b)	Fan’s Teachings.....	81
c)	Motivation to Combine Feinauer-Prabhavalkar and Fan	82
i.	“a context associated with the input chunk”	86
d)	Prabhavalkar’s Teachings.....	87
e)	Motivation to Combine Prabhavalkar’s Context.....	91
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>	<i>94</i>
a)	Fan’s Teachings.....	95
b)	Motivation to Combine Feinauer-Prabhavalkar and Fan	98
i.	“a context associated with the input chunk”	101

5.	<i>Claim 1(d): “receiving a speaker embedding for a human speaker; and”</i>	102
a.	Feinauer’s Teachings.....	103
b.	Fan’s Teachings.....	105
c.	Motivation to Combine Fan’s speaker embedding	108
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>	115
a.	Feinauer’s Teachings.....	116
B.	CLAIM 2	125
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>	125
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>	125
a.	Prabhavalkar’s Teachings.....	125
b.	Feinauer’s Teachings.....	126
a)	Motivation to Combine Prabhavalkar	128
C.	CLAIM 3	131
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>	131
a.	Motivation to Combine Fan’s speaker embedding	134
2.	<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>	136
3.	<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>	137
D.	CLAIM 14	137

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>	<i>137</i>
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 input chunks including a pre-defined number of frames of the speech audio signal;”</i>	<i>137</i>
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>	<i>137</i>
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>	<i>138</i>
5.	<i>Claim 14(d): “receive a speaker embedding for a human speaker; and;”</i>	<i>138</i>
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>	<i>138</i>
E.	CLAIM 15	<i>138</i>
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>	<i>138</i>
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>	<i>138</i>
F.	CLAIM 16	<i>138</i>
1.	<i>Claim 16: “The computing apparatus of claim 14, wherein the speaker embedding is generated based on the speech audio signal.”</i>	<i>138</i>
G.	CLAIM 17	<i>139</i>
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>	<i>139</i>

H.	CLAIM 20	139
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:”	139
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;”	139
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;”	139
4.	Claim 20(c): “extract, by a linguistic features extraction module from the input chunk and the context, linguistic features with a reduced accent;”	139
5.	Claim 20(d): “receive a speaker embedding for a human speaker; and”	139
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.”	140
X.	GROUND 2: CLAIMS 9-11 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND ACCENTRON	140
A.	CLAIM 9	140
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.”	140
a.	Accentron’s Teachings	140
b.	Motivation to Combine the modified Feinauer with Accentron and Fan	144
B.	CLAIM 10	151
1.	Claim 10(a): “The method for real-time correction of claim 9, wherein: the synthesis module includes an	

	<i>encoder, a decoder, and a post-net module designed to improve an output of the decoder; and;</i>	151
	a. Accentron’s Teachings	151
	b. Motivation to Combine Accentron.....	153
2.	<i>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.”</i>	160
	a) Motivation to Combine Accentron.....	162
C.	CLAIM 11	167
	<i>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.”</i>	167
XI.	GROUND 3: CLAIMS 6 AND 19 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND SINGH.....	169
A.	CLAIM 6	169

1.	<i>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.”</i>	169
a.	Singh’s Teachings	170
b.	Motivation to Combine Singh	173
B.	CLAIM 19	175
1.	<i>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.”</i>	175
XII.	GROUND 4: CLAIMS 7 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND WU	175
A.	CLAIM 7	175
1.	<i>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.”</i>	175
a.	Wu’s Teachings	176
b.	Motivation to Combine Wu	178
XIII.	GROUND 5: CLAIMS 12-13 ARE OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND BONGIOVI	179
A.	CLAIM 12	179
1.	<i>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</i>	

	<i>extraction of the linguistic features and the acoustic features.”</i>	179
a.	Feinauer-Prabhavalkar-Fan’s Teachings	180
i.	“prior to dividing the speech audio signal”	180
ii.	“processing the speech audio signal...”	180
b.	Bongiovi’s Teachings	180
i.	“to improve extraction of the linguistic features and the acoustic features”	182
c.	Motivation to Combine Bongiovi	184
B.	CLAIM 13	190
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>	190
a)	Motivation to Combine Bongiovi	192
XIV.	GROUND 6: CLAIM 5 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND OSUNA	196
A.	CLAIM 5	196
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>	196
a.	Osuna’s Teachings	197
b.	Motivation to Combine Osuna	200
XV.	GROUND 7: CLAIM 18 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, WU, AND OSUNA	204
A.	CLAIM 18	204
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</i>	

	<i>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>	204
XVI. CONCLUSION		204

I, Christopher Schmandt, declare as follows:

I. INTRODUCTION AND QUALIFICATIONS

1. I am over the age of 21 and am competent to make this declaration.

A. Educational Background and Professional Experience

2. I retired in early 2019 after a 40-year career at the Massachusetts Institute of Technology (“MIT”); for most of that time I was employed as a Principal Research Scientist at the Media Laboratory. In that role I also served as faculty for the MIT Media Arts and Sciences academic program. I was a founder of the Media Laboratory, a research lab which now spans two buildings.

3. I received my B.S. degree in Electrical Engineering and Computer Science from MIT in 1978, and my M.S. in Visual Studies (Computer Graphics) also from MIT. I was employed at MIT since 1980, initially at the Architecture Machine Group which was an early computer graphics and interactive systems research lab. In 1985, I helped found the Media Laboratory and continued to work there until retirement. I was director of a research group titled “Living Mobile.” My research spanned distributed communication and collaborative systems, with an emphasis on multi-media and user interfaces, with a strong focus on speech-based systems. I have over 70 published conference and journal papers and one book in the field of speech technology and user interaction.

4. For the first fifteen years of my career, my research emphasized speech processing and speech user interfaces. I built the first conversational computer system utilizing speech recognition and synthesis (“Put That There”) starting in 1980. I continued to innovate speech processing user interfaces using recognition, text-to-speech synthesis, and recorded audio in a wide variety of projects. I built one of the first graphical user interfaces for audio editing, employing keyword recognition on voice memos in 1982 (Intelligent Ear). I built the first research-grade unified messaging system, which combined text and voice messages into a single inbox, with speech recognition over the phone for remote access, and a graphical user interface for desktop access in 1983 (Phone Slave). Along with my students, we built the first system for real time spoken driving directions, including speech-accessible maps of Cambridge, Massachusetts in 1987 (Back Seat Driver). We also built some of the earliest speech-based personal assistants for managing messages, calendar, contacts, and other information. (Conversational Desktop 1985, Chatter 1993, MailCall 1996). Quite a few of the systems we built employed speech processing in handheld mobile devices (ComMotion 1999, Nomadic Radio 2000, Impromptu 2001, and Symphony 2004, for example). We applied speech recognition to large bodies of everyday conversations captured with a wearable device and utilized as a memory aid (Memory Prosthesis 2004). We used speech recognition on radio newscasts to build a personalized version of audio newscasts (Synthetic News

Radio, 1999) and also investigated adding speech recognition to a mouse-based window system a few years earlier.

5. I was later awarded the prestigious Association for Computing Machinery (ACM) Computer Human Interface (CHI) Academy membership specifically for those years of work pioneering speech user interfaces.

6. In addition to the above mentioned recitation of systems which utilized speech recognition, synthesis, and digital audio, a number of my projects incorporated machine learning and speech processing more closely tied to the subject matter of the patent about which I have been asked to opine.

7. In 1983 in a project titled Zero Bandwidth Video I used real-time spectral analysis of speech to select one of a class of phonemes based on facial lip positions to animate a computer generated “talking head” of the speaker. Around 1986 my project Say It Like This analyzed recorded speech for pitch and energy attributes, and used these to resynthesize speech from another talker to mimic the prosody of the original talker. Starting in 1988 I explored how similar speech attributes, also including speech rate and syllable duration, could be incorporated into synthetic speech to mimic various affects or emotional states, as expressed by machine. Around 1987-1988 I used real-time analysis of speech energy, pitch, and voice activity to infer intent in context from spoken utterances without performing

recognition or other linguistic analysis. Real-time energy detection was also used in computer-based voice mail systems I built in 1984 and 1990.

8. Several projects incorporated other aspects of intonation and pitch tracking. Intonation was found to be useful to understand the semantics of user repairs of speech recognition errors in conversational systems in 1993, and to detect emphasis and semantic structure of recorded audio documents in various projects from 1994 to 1996, culminating in the Audio Notebook, a paper based audio document browser with topic detection. In 2003 and 2006 real-time detection of pitch characteristics and voice activity was used to determine when to mask speech in an environment which transmitted ambient audio from a home to a remote location.

9. In around 2000 I worked for several years using the machine learning technique of Latent Semantic Indexing, which includes singular value decomposition, to classify and group audio news stories for Synthetic News Radio. In 1995 the Newscom project included a back propagation neural network for the purpose of detecting different speakers during a conversation without *a priori* training.

10. Because of my early work with distributed speech systems, I served for several years in the mid-1990s with a working group on the impact of multimedia systems on the Internet reporting to the Internet Engineering Task Force (IETF) and

later the Internet Activities Board (IAB). This work impacted emerging standards such as Session Initiation Protocol (SIP).

11. In my faculty position I taught graduate level courses in speech technology and user interaction design, and I directly supervised student research and theses at the Bachelors, Masters, and PhD level. I oversaw the Masters and PhD thesis programs for the entire Media Arts and Sciences academic program during my more senior years. I also served on the Media Laboratory intellectual property committee for many years.

II. METHODOLOGY AND MATERIALS CONSIDERED

12. I have relied upon my education, knowledge and experience with audio speech signal processing systems more generally, as well as the other materials as discussed in this declaration in forming my opinions.

13. For this work, I have been asked to review U.S. Patent No. 11,715,457 (“the ‘457 Patent”) (Ex. 1001) including the specification and claims, and the ‘457 Patent’s prosecution history (“‘457 File History”) (Ex. 1002). In developing my opinions relating to the ‘457 Patent, I have considered the materials cited or discussed herein, including those itemized in the Exhibit Table below.

Exhibit	Description
Exhibit 1001	U.S. Patent No. 11,715,457 (“‘457 Patent”)
Exhibit 1002	Prosecution History for the 11,715,457 Patent (“‘457 File History”)
Exhibit 1004	U.S. Patent Application Publication No. 2020/0193971 to Feinauer et al. (“Feinauer”)

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, <i>Voice Communication with Computers</i> , Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)
Exhibit 1022	C. O’Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallaghan.com/research/papers/Pitch.pdf) September 30, 2015, (“ <i>O’Callaghan</i> ”)
Exhibit 1023	B. Sisman et al., <i>Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion</i> ,

	IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, April, 2019, 1085-1097. (“ <i>Sisman</i> ”)
Exhibit 1024	J. Holmes, <i>Speech Synthesis and Recognition</i> , Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003, (“ <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, <i>Chapter 13 English Sounds in Context: The Pronunciation of Phonemes and Morphemes</i> , in <i>Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners</i> , 303-324, 2019 (“ <i>Wiltshire</i> ”)
Exhibit 1028	N. Kazanina, et al., <i>Phonemes: Lexical access and beyond</i> , Psychon Bull Rev 25, 560–585, September 5, 2017 (“ <i>Kazanina</i> ”)
Exhibit 1029	D. Jurafsky et al., <i>Speech and Language Processing</i> , Chapter 27, October 16, 2019 (“ <i>Jurafsky</i> ”)
Exhibit 1030	L. Sun, <i>Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training</i> , IEEE on Multimedia and Expo, August 29, 2016 (“ <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, <i>Understanding the Mel Spectrogram</i> , (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, <i>Natural TTS Synthesis By Conditioning Wavenet on Mel Spectrogram Predications</i> , 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, <i>Tacotron: Towards End-to End Speech Synthesis</i> , arXiv, April 6, 2017 (“ <i>Wang</i> ”)
Exhibit 1039	U.S. Patent Appl. No 2020/0134425 to Chen et al. (“ <i>Yoo</i> ”)

Exhibit 1040	A. Tripathi et al, <i>Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition</i> , 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

14. I have considered these materials from the viewpoint of a PHOSITA as of the priority date of the '457 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the '457 Patent is January 10, 2022. I note that my opinions provided in this Declaration are made from the perspective of a PHOSITA as of this priority date of the '457 Patent, unless expressly stated otherwise. To the extent that I use any verb tense in this Declaration that is present tense (e.g., “a PHOSITA would understand” instead of “a PHOSITA would have understood”), such verb tense should be understood to be my opinion as of the '457 Patent’s priority date (again, unless expressly stated otherwise). I merely use the present verb tense for ease of reading.

III. OVERVIEW AND LEGAL STANDARDS

15. In formulating my opinions, I have been instructed to apply certain legal standards. I am not a lawyer. I do not expect to offer any testimony regarding what the law is. Instead, the following sections summarize the law as I have been

instructed to apply it in formulating and rendering my opinions found later in this declaration. I understand that, in an *inter partes* review (“IPR”) proceeding, patent claims may be deemed unpatentable if it is shown that they were anticipated or rendered obvious in view of the prior art. I understand that prior art in an IPR review is limited to patents or printed publications that predate the priority date of the patent at issue. I understand that questions of claim clarity (definiteness) and enablement cannot be considered as a ground for considering the patentability of a claim in these proceedings.

A. Person of Ordinary Skill in the Art

16. I understand that the ’457 Patent, the record of proceedings at the Patent Office (which I understand is called the “File History” or “Prosecution History”), and the teachings of the prior art are evaluated from the perspective of a person of ordinary skill in the art (“PHOSITA”). I understand that the factors considered in determining the ordinary level of skill in the art include: (i) the levels of education and experience of persons working in the field; (ii) the types of problems encountered in the field; and (iii) the sophistication of the technology. I may also consider, if available, the education level of the inventor, prior art solutions to the problems encountered in the art, and the rapidity with which innovations are made in the relevant art.

17. I understand that a person of ordinary skill in the art is not a specific real individual, but rather a hypothetical individual having the qualities reflected by the factors above. This hypothetical person has knowledge of all prior art in the relevant field as if it were arranged on a workshop wall and takes from each reference what it would teach to a person having the skills of a PHOSITA.

B. Obviousness

18. I understand that a claim may be invalid under § 103(a) if the subject matter described by the claim as a whole would have been “obvious” to a hypothetical PHOSITA in view of a single prior art reference or in view of a combination of references at the time the claimed invention was made. Therefore, I understand that obviousness is determined from the perspective of a hypothetical PHOSITA. I further understand that a hypothetical PHOSITA is assumed to know and to have all relevant prior art in the field of endeavor covered by the patent in suit and all analogous prior art. I understand that obviousness in an IPR review proceeding is evaluated using a preponderance of the evidence standard, which means that the claims must be more likely obvious than nonobvious.

19. I also understand that an analysis of whether a claimed invention would have been obvious should be considered in light of the scope and content of the prior art, the differences (if any) between the prior art and the claimed invention, and the level of ordinary skill in the pertinent art involved. I understand as well that a prior

art reference should be viewed as a whole. I understand that in considering whether an invention for a claimed combination would have been obvious, I may assess whether there are apparent reasons to combine known elements in the prior art in the manner claimed in view of interrelated teachings of multiple prior art references, the effects of demands known to the design community or present in the marketplace, and/or the background knowledge possessed by a PHOSITA. I also understand that other principles may be relied on in evaluating whether a claimed invention would have been obvious, and that these principles include the following:

- A combination of familiar elements according to known methods is likely to be obvious when it does no more than yield predictable results;
- When a device or technology is available in one field of endeavor, design incentives and other market forces can prompt variations of it, either in the same field or in a different one, so that if a PHOSITA can implement a predictable variation, the variation is likely obvious;
- If a technique has been used to improve one device, and a PHOSITA would have recognized that it would improve similar devices in the same way, using the technique is obvious unless its actual application is beyond his or her skill;
- An explicit or implicit teaching, suggestion, or motivation to combine two prior art references to form the claimed combination may

demonstrate obviousness, but proof of obviousness does not depend on or require showing a teaching, suggestion, or motivation to combine;

- Market demand, rather than scientific literature, can drive design trends and may show obviousness;
- In determining whether the subject matter of a patent claim would have been obvious, neither the particular motivation nor the avowed purpose of the named inventor controls;
- One of the ways in which a patent's subject can be proved obvious is by noting that there existed at the time of invention a known problem for which there was an obvious solution encompassed by the patent's claims;
- Any need or problem known in the field of endeavor at the time of invention and addressed by the patent can provide a reason for combining the elements in the manner claimed;
- "Common sense" teaches that familiar items may have obvious uses beyond their primary purposes, and in many cases a PHOSITA will be able to fit the teachings of multiple patents together like pieces of a puzzle;
- A PHOSITA is also a person of ordinary creativity, and is not an automaton;

- A patent claim can be proved obvious by showing that the claimed combination of elements was “obvious to try,” particularly when there is a design need or market pressure to solve a problem and there are a finite number of identified, predictable solutions such that a PHOSITA would have had good reason to pursue the known options within his or her technical grasp; and
- One should not use hindsight in evaluating whether a claimed invention would have been obvious.

20. I also understand that an obviousness determination includes the consideration of various factors such as: (1) the scope and content of the prior art, (2) the differences between the prior art and the asserted claims, (3) the level of ordinary skill in the pertinent art, and (4) the existence of secondary considerations such as commercial success, long-felt but unresolved needs, failure of others, and so forth.

21. I am informed that it is improper to combine references where the references teach away from their combination. I am informed that a reference may be said to teach away when a person of ordinary skill in the relevant art, upon reading the reference, would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path that was taken by the patent applicant. In general, a reference will teach away if it suggests that the line of

development flowing from the reference's disclosure is unlikely to be productive of the result sought by the patentee. I am informed that a reference teaches away, for example, if (1) the combination would produce a seemingly inoperative device, or (2) the references leave the impression that the product would not have the property sought by the patentee. I also am informed, however, that a reference does not teach away if it merely expresses a general preference for an alternative invention but does not criticize, discredit, or otherwise discourage investigation into the invention claimed.

22. I am informed that even if a *prima facie* case of obviousness is established, the final determination of obviousness must also consider "secondary considerations" if presented. In most instances, the patentee raises these secondary considerations of non-obviousness. In that context, the patentee argues an invention would not have been obvious in view of these considerations, which include: (a) commercial success of a product due to the merits of the claimed invention; (b) a long-felt, but unsatisfied need for the invention; (c) failure of others to find the solution provided by the claimed invention; (d) deliberate copying of the invention by others; (e) unexpected results achieved by the invention; (f) praise of the invention by others skilled in the art; (g) lack of independent simultaneous invention within a comparatively short space of time; (h) teaching away from the invention in the prior art.

23. I am further informed that secondary considerations evidence is only relevant if the offering party establishes a connection, or nexus, between the evidence and the claimed invention. The nexus cannot be based on prior art features. The establishment of a nexus is a question of fact. While I understand that the Patent Owner here has not offered any secondary considerations at this time, I will supplement my opinions in the event that the Patent Owner raises secondary considerations during the course of this proceeding.

24. Finally, I understand that obviousness in an IPR must be proven by a preponderance of the evidence.

C. Analogous Art

25. I have been informed that for a prior art reference to be proper for use in an obviousness analysis, the reference must be “analogous art” to the claimed invention. I have been informed that a reference is analogous art to the claimed invention if: (1) the reference is from the same field of endeavor as the claimed invention (even if it addresses a different problem); or (2) the reference is reasonably pertinent to the problem faced by the inventor (even if it is not in the same field of endeavor as the claimed invention). In order for a reference to be “reasonably pertinent” to the problem, it must logically have commended itself to an inventor’s attention in considering his problem. In determining whether a reference is reasonably pertinent, one should consider the problem faced by the inventor, as

reflected either explicitly or implicitly, in the specification. I believe that all of the references that my opinions in this IPR are based upon are well within the range of references a person having ordinary skill in the art would consult to address the type of problems described in the Challenged Claims.

D. Claim Construction

26. I understand that the claim terms in this IPR will be construed according to their plain and ordinary meaning as understood in light of the claim language, the patent's description, and the prosecution history viewed from the perspective of a PHOSITA. I further understand that where a patent defines claim language, the definition in the patent controls, even if there are other definitions that might be understood by those working in the art.

IV. LEVEL OF A PERSON OF ORDINARY SKILL

27. Based on my review and analysis of the '457 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a PHOSITA in the field of the '457 Patent at the time of the earliest possible priority date (January 10, 2022) would have been knowledgeable regarding the field of speech audio processing. In my experience working in this field, most workers of ordinary skill in the art as of the earliest possible priority date of January 10, 2022, 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. *See, e.g.*, the '457 Patent (Ex. 1001) at 1:23-3:3 (describing the “Background” and “Summary”). A person with less relevant practical experience but with additional education may also qualify as a PHOSITA in the field of the '457 Patent provided the additional education focused on speech processing technology. When I refer to the understanding of a PHOSITA, I am referring to the understanding of such a person as of January 10, 2022.

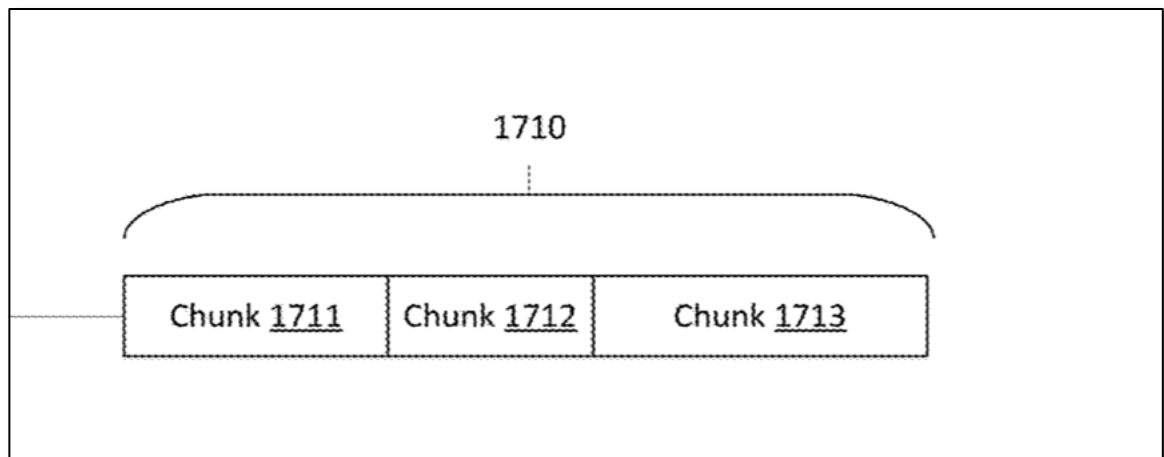
V. OVERVIEW OF TECHNOLOGY

28. I was asked to briefly summarize the background of the technology from the standpoint of a PHOSITA prior to January 10, 2022.

A. Audio Chunks

29. It was well known to process audio signals utterances in smaller portions, especially to process audio in real-time. For example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” in order to maintain “low real time delay” EX1018-Wasserblat, Abstract, [0025]. U.S. Patent Appl. No. 2021/0272571 to Balasubramaniam et al. (“*Balasubramaniam*”) teaches waveforms “of each speech utterance may be split into chunks of 200 ms with 10 ms overlapping frames.” EX1019-Balasubramaniam, [0310].

30. As another example, U.S. Patent Appl. No. 2020/0294492 to Sypniewski et al. (“*Sypniewski*”) teaches a front end module 201 that splits raw audio data into a series of frames. EX1020-Sypniewski, [0059], Fig. 2. *Sypniewski* provides an example in which a 7-10 second long portion of input audio “may be segmented on relative silence such that the input audio comprises an entire utterance.” EX1020-Sypniewski, [0057], [0110]. *Sypniewski* characterizes these 7-10 second portions of audio as “chunks.” EX1020-Sypniewski, [0162]. Fig. 17 below illustrates an example of audio 1710 split into chunks 1711-1713.



EX1020-Sypniewski, Fig. 17. (excerpt) *Sypniewski* notes “[a]ny frame duration, spacing, and overlap may be used as appropriate for any given implementation as determined by one skilled in the art.” EX1020-Sypniewski, [0059].

B. Acoustic Features

31. It was well known that acoustics are a measured property of sound such as, but not limited to, a characteristic of speech analyzed in physical terms. Further,

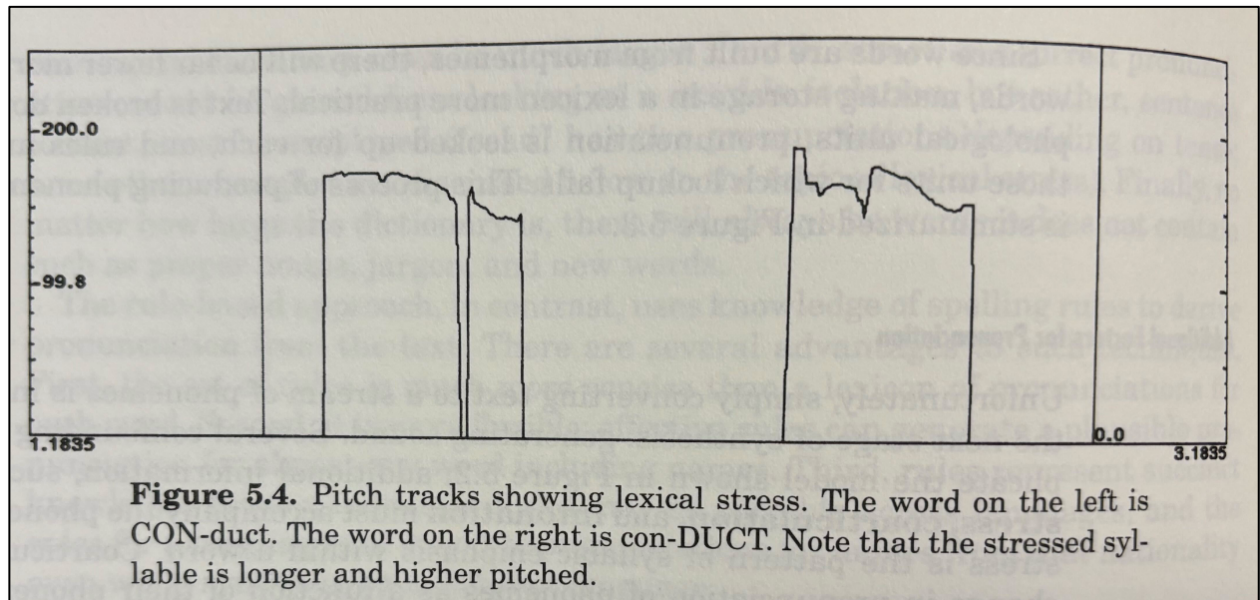
it was well known in the prior art that acoustic features may include several features, including but not limited to stress, accent, volume, energy, tone, pitch, main frequency F0, and voice activity detector as discussed below.

32. It was common in the prior art that several of these features were also referred to as “style” and/or “prosody”. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches that prosody “refers to the spoken style of discourse independent of lexical content[.]” EX1021-Schmandt, 301-302. *Schmandt* further teaches that stress, pitch, duration, rhythm, meter, and intonation are all examples of prosody speech style. EX1021-Schmandt, p. 17, 301-302.

C. Stress or Accent

33. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches lexical “stress is part of a complex interplay between the intonation and rhythm of a whole sentence...Stress covers many phenomena, all realized as variations in rhythm and meter; lexical stress is what differentiates syllabic ‘accent’ as opposed to emphasizing one word over another in a sentence. Lexical stress is important for understanding spoken language; incorrectly assigned stress may make a word unintelligible. For some English words the assignment of stress differentiates noun and verb forms, e.g., ‘convert’ and ‘conflict’...stress is not directly phonemic. In addition to producing the proper phonemes, the syllables containing those phonemes also must be marked for stress. This is done by assigning pitch and

duration for each voiced phoneme during the conversion to phonemes.” EX1021-Schmandt, 88.

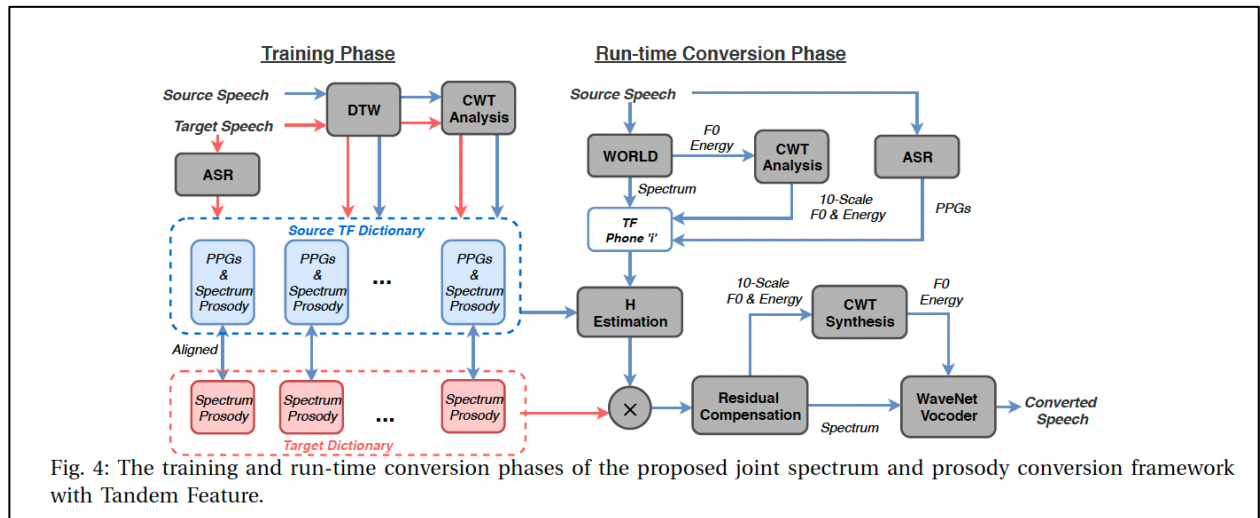


EX1021-Schmandt, Fig. 5.4

1. *Pitch or Fundamental Frequency (F0)*

34. *Pitch* to O’Callaghan (“O’Callaghan”) teaches pitch “of a sound is identical with its fundamental frequency; that is, the pitch of a periodic sound is the greatest whole-number frequency by which the frequency of each of its sinusoidal components is divisible without remainder.” EX1022-O’Callaghan, 4. In other words, “the fundamental frequency of a complex periodic tone determines its perceived pitch.” EX1022-O’Callaghan, 4. As another example, EX1023, *Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion* to Sisman et al. (“Sisman”) teaches “[a]t run-time conversion,

we use CWT to obtain the 10-scale representations of F0 and energy contour[.]” EX1023-Sisman, 6-7.

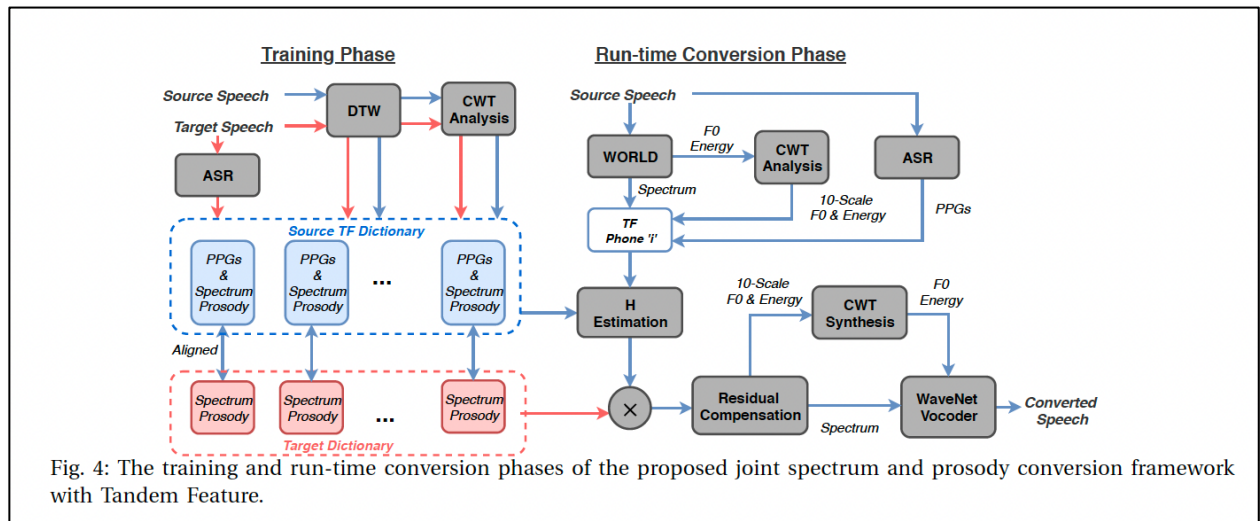


EX1023-Sisman, Fig. 4. Pitch or fundamental frequency is “a crucial prosodic feature in speech[.]” EX1023-Sisman 3rhc. *Speech Synthesis and Recognition* to Holmes et al. (“*Holmes*”) teaches that pitch may change by “the pattern of stressed syllables...that are influenced by the phonetic detail of an utterance.” EX1024-Homles, 104-105.

2. Energy or Volume

35. It was common to measure energy and volume in signals. To calculate loudness, the energy of a signal is measured over time. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“*Fletcher*”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to

this direction at the point.” EX1025-Fletcher, 83. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. *Sisman* teaches “[a]t run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour.” EX1023-Sisman, 6-7.



EX1023-Sisman, Fig. 4.

3. Voice Activity Detection (VAD)

36. *Personal VAD: Speaker-Conditioned Voice Activity Detection* to Chang et al. (“Chang”) teaches “a system to detect the voice activity of a target speaker at the frame level.” EX1026-Chang, Abstract. *Chang* teaches to find the “voice activity of a target speaker in real time.” EX1026-Chang, 11hc. The voice activity determines if audio contains “speech/non-speech.” EX1026-Chang, 11hc.

37. It was well known to perform VAD so that the “performance of downstream components by discarding non-speech signals, but also significantly

reduces the overall computational cost due to its relatively small size.” EX1026-Chang 1rhc. Detecting VAD “reduce[s] the computational cost and battery consumption[.]” EX1026-Chang, Abstract.

D. Linguistic Features

38. Linguistic features were commonly known in the prior art to include an expression of a phoneme. A phoneme is the smallest unit of sound that distinguishes meaning in a language. EX1027-Wiltshire, 304.

39. *Phonemes: Lexical access and beyond* to Kazanina et al. (“Kazanina”) teaches phonemes have two essential properties: “they are ‘segment-sized’ (the size of a consonant or a vowel) and abstract (a single phoneme may have different acoustic realisations).” EX1028-Kazanina, 560. *English Sounds in Context: The Pronunciation of Phonemes and Morphemes* to Wiltshire et al. (“Wiltshire”) teaches changing one phoneme changes the meaning of a word. EX1027-Wiltshire, 560. For example, “the words *pat* and *bat* are identical except for the initial sounds, which are therefore responsible for indicating the difference in meaning between the two words.” EX1027-Wiltshire, 304. Such pairs of words are called “minimal pairs,” because they only differ by a single sound but have completely different meanings. EX1027-Wiltshire, 304. Thus, phonemes are extremely helpful in differentiating minimal pairs because “the phoneme is the smallest potential unit of difference

between similar words recognizable as different to the native [speaker].” EX1027-Wiltshire, 304.

40. As such, “two words that are distinct in meaning differ in phonological form, with a minimal difference being exactly one phoneme within the same position in the word (e.g. /kæt/ ‘cat’ vs. /mæt/ ‘mat’).” EX1028-Kazanina, 561. Additionally, different words with distinct meanings can employ the same set of phonemes but in different orders (e.g. cat /kæt/ vs. act /ækt/ vs. tack /tæk/). EX1028-Kazanina, 561.

41. Different phonetic alphabet systems have been designed to represent phonemes. As discussed in *Speech and Language Processing*, Chapter 27 “Phonetics” to Jurafsky et al. (“*Jurafsky*”), one phonetic alphabet is the International Phonetic Alphabet (IPA), which is an evolving standard originally developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. EX1029-Jurafsky, 2. Another phonetic alphabet is the ARPAbet, which uses “ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters.” EX1021-Schmandt, 16. ARPAbet was developed “for the convenience of computer manipulation and representation of speech using ASCII-printable characters.” EX1021-Schmandt, 16.

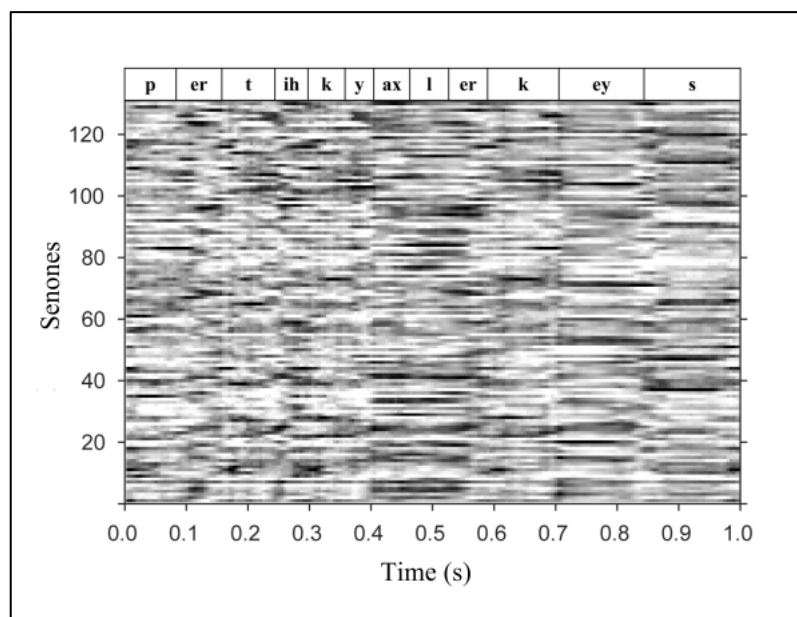
ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɜ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

Figure 27.2 ARPAbet symbols for American English vowels, with IPA equivalents.

EX1029-Jurafsky, 2.

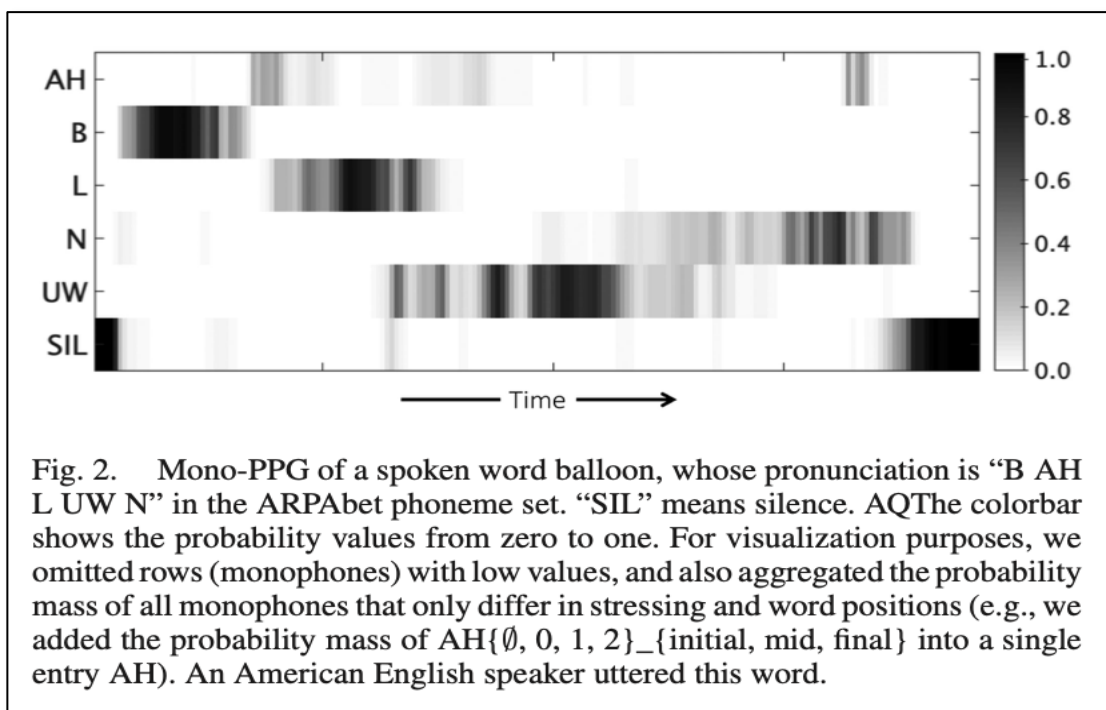
1. *Phonetic Posteriorgrams (PPGs)*

42. One common way to represent phonemes in accent conversion systems is in the form of a phonetic posteriorgram. *Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training* to Sun et al. (“Sun”) teaches a Phonetic Posteriorgram (PPG) “is a time-versus-class matrix representing the posterior probabilities of each phonetic class” per utterance. EX1030-Sun, 2–3. “A phonetic class may refer to a word, a phone or a senone.” EX1030-Sun, 3. An example of a PPG (using senones) is as follows:



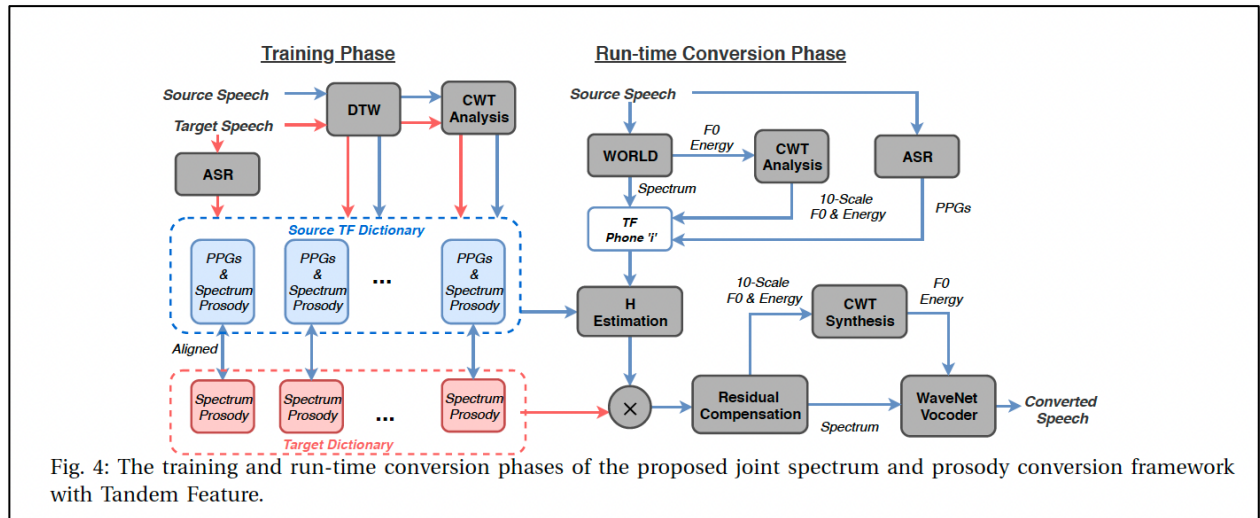
EX1030-Sun, Fig 4. In this example, the “horizontal axis represents time in seconds and the vertical one contain indices of phonetic classes.” EX1030-Sun, 3. A “[d]arker shade implies a higher posterior probability.” EX1030-Sun, 3.

43. Another example of a phonetic posteriorgrams used to represent phonemes is taught in EX1045, *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) using the ARPAbet phonetic alphabet and depicted below:



EX1045-Zhao, Fig. 2. *Zhao* teaches this “phonetic posteriorgram obtained by collapsing the senones into monophone symbols (346 monophones with word positions, e.g., word-initials, word-finals). For each monophone symbol, we aggregate the probability mass of all the senones that share the same root monophone.” EX1045-Zhao, 2370rhc.

44. As another example, EX1023, *Sisman* teaches “to convert all speaker’s voice into PPGs, and construct dictionaries that consist of source spectrum and prosody features as well as PPGs.” EX1023-Sisman, 7. *Sisman* teaches “a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary[.]” EX1023-Sisman, 7.



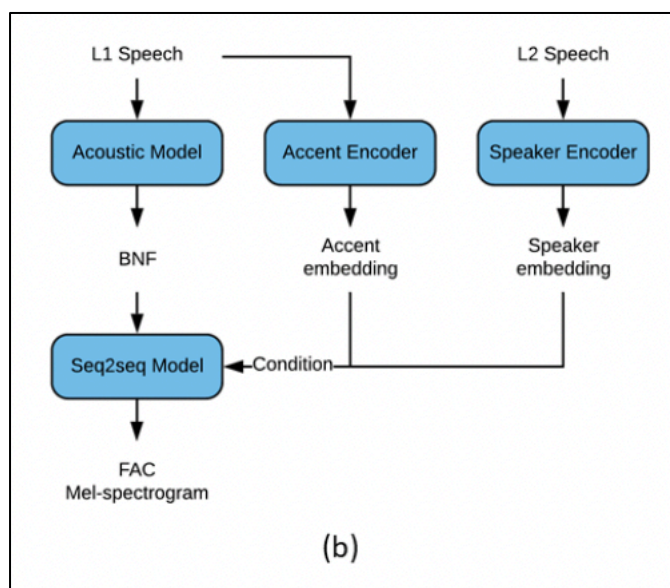
EX1023-Sisman, Fig. 4. *Sisman*'s PPG dictionary has "multiple coupled dictionaries $[A_i; B_i]$, one for each phone i , where $i = 1, \dots, N$, A_i is the source phonetic dictionary, and B_i the target phonetic dictionary. The source phonetic dictionary here consists of both PPGs, source spectrum features as well as CWT-based prosody features, called source TF dictionary while the target phonetic dictionary B_i only includes spectrum features and prosody features, called target dictionary." EX1023-Sisman, 8.

E. Prior Art Accent Conversion Systems

45. After extracting the linguistic and acoustic features, along with any other needed information, it was common practice to perform accent conversion. For example, U.S. Patent Appl. No. 2022/0122579 to Biadsy et al. ("*Biadsy*") teaches the speech-to-speech conversion server within the computing device "convert[s] audio data of an utterance spoken by a user who may speak with an accent into audio

data of a synthesized utterance that may have a different accent.” EX1032-Biadsy, [0023]; *see also* [0021].

46. As another example, *Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning* to Zhao et al. (“*Accentron*”) teaches an accent conversion system that “can generate accent conversions to arbitrary L2 speakers given a few seconds of audio (i.e., enough speech to compute a speaker embedding)[.]” EX1008-Accentron, 3. *Accentron* teaches “*Accentron*, an approach that can generate accent-converted speech for arbitrary L2 speakers unseen during training. In the proposed approach, we first train a speaker-independent acoustic model on L1 corpora to extract bottleneck features that represent the linguistic content of utterances. Then, we develop a speaker encoder and an accent encoder to generate embedding vectors for the desired voice identity (L2 speaker’s) and accent (L1 accent), respectively.” EX1008-Accentron, 1.



EX1008-Accentron, Fig. 1 (excerpt). *Accentron* “use[s] a sequence-to-sequence model to transform bottleneck-features to Mel-spectrograms, conditioned on the L2 speaker embedding and the L1 accent embedding. We conducted experiments on the L2-ARCTIC corpus under two testing conditions: the standard FAC setting where test L2 speakers were seen during training, and a zero-shot FAC setting where test L2 speakers were unseen during training.” EX1008-Accentron, 1.

47. Further, U.S. Patent Appl. No. 2018/0174595 to Dirac et al. (“*Dirac*”) teaches an accent conversion system that uses “collected audio samples to generate one or more accent translation models for translating speech from one accent to another.” EX1033-*Dirac*, [0015]. That is, audio samples “associated with at least a first accent” are compared to audio samples of “at least one other accent” to determine “various audio characteristics associated with...the accent.” EX1033-*Dirac*, [0048]. Resultingly, *Dirac* teaches outputting an “audio portion substantially associated with the second accent in the...same spoken language as the input audio” via the accent translation model, including adjusting the audio characteristics of the input audio ...to more closely resemble the audio characteristics of the second accent. EX1033-*Dirac*, [0053] and [0049].

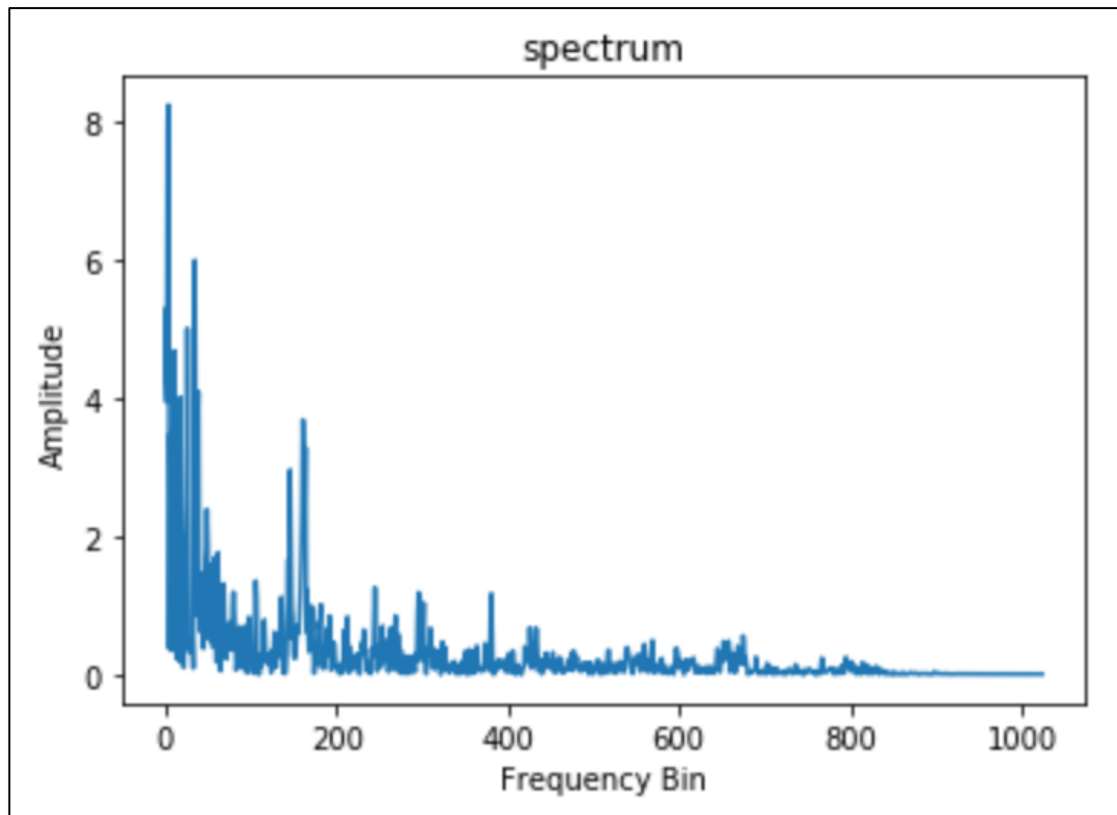
1. Encoder and Decoder

48. *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“*Zhao*”) teaches one example of “an encoder-decoder architecture[.]” EX1045-

Zhao, 2371. This encoder-decoder architecture uses an encoder to “consume” input sequences and generate a high-level hidden representation sequence. EX1045-Zhao, 2371. Then, *Zhao* passes “the hidden linguistic representation sequence to the decoder, which consists of a location-sensitive attention mechanism [47] and a decoder LSTM, to predict the raw Mel- spectrogram.” EX1045-Zhao, 2371.

2. *Mel Spectrograms*

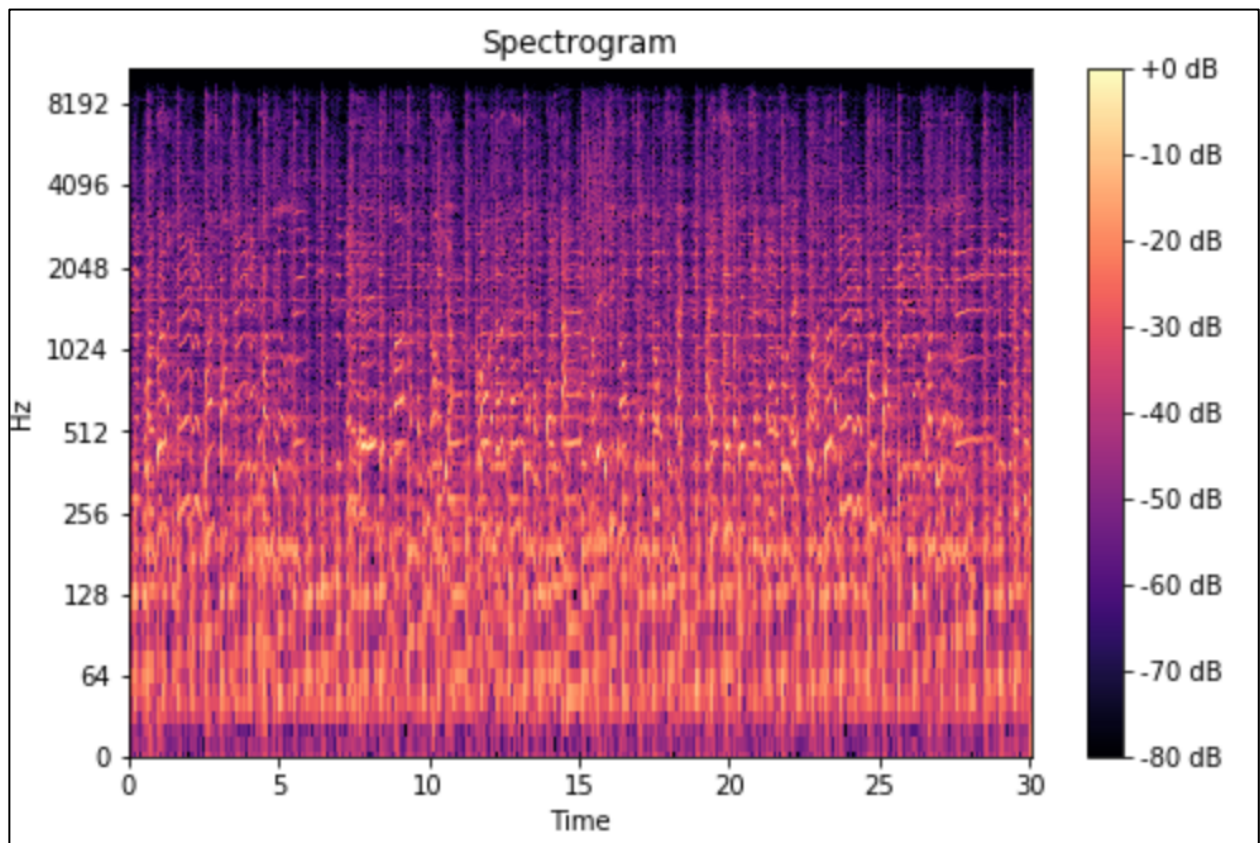
49. An audio signal is “comprised of several single-frequency sound waves...The Fourier transform is a mathematical formula that allows us to decompose a signal into it’s (sic) individual frequencies and the frequency’s amplitude. In other words, it converts the signal from the time domain into the frequency domain. The result is called a spectrum.” EX1034-Roberts, 4-5.



EX1034-Roberts, 4.

50. Either a “short-time Fourier transform” or a “fast Fourier transform” (FFT) is performed on the frequency domain signal in order to “analyze the frequency content of a signal”, and if the “FFT is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram” in order to visually represent the signals “loudness, or amplitude, as it varies over time at different frequencies.” EX1034-Roberts, 7-8. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches a “spectrogram is a transformation of the waveform into the frequency domain.” EX1021-Schmandt, 13. In a spectrogram, the “horizontal axis represents time, the vertical axis represents frequency, and the

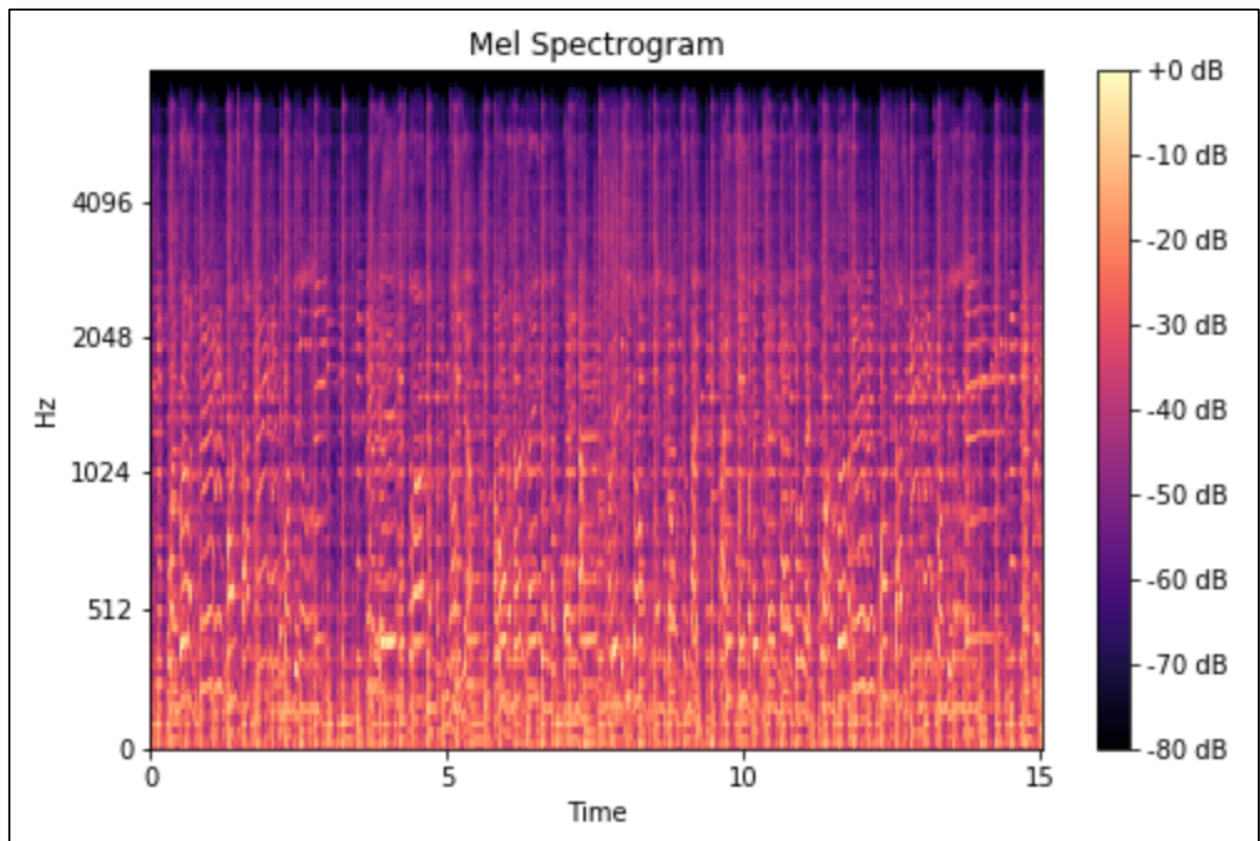
intensity or blackness at a point indicates the acoustic energy at that frequency and time.” EX1021-Schmandt, 13. A spectrogram “requires much less storage than the original waveform and reveals acoustic features specific to speech.” EX1021-Schmandt, 13.



EX1034-Roberts, 6.

51. Studies “have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies.” EX1034-Roberts, 11. In order to combat this, in “1937, Stevens, Volkman, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the mel scale.” EX1034-

Roberts, 12. From the spectrogram in the prior paragraph, you “perform a mathematical operation on frequencies to convert them to the mel scale.” EX1034-Roberts, 12. A Mel-spectrogram “is a spectrogram where the frequencies are converted to the mel scale.” EX1034-Roberts, 13.



EX1034-Roberts, 8.

3. *Vocoder*

52. A vocoder converts a computed representation of speech, such as a spectrogram or LPC coefficients, into an audible speech waveform. *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches that to obtain hearable synthesized audio, mel spectrograms are transformed into

waveforms using a real-time vocoder. EX1045-Zhao, 6rhc (“The predicted Mel-spectrograms are converted back to audio waveforms using a WaveGlow neural vocoder trained on the L2 utterances[.]”); EX1045-Zhao, 2rhc. (“we use a high-quality WaveGlow [15] real-time neural vocoder to convert Mel-spectrograms to speech waveform.”).

53. Another type of vocoder is a “WaveNet vocoder [21] to generate the speech waveforms.” EX1023-Sisman, 9 lhc.

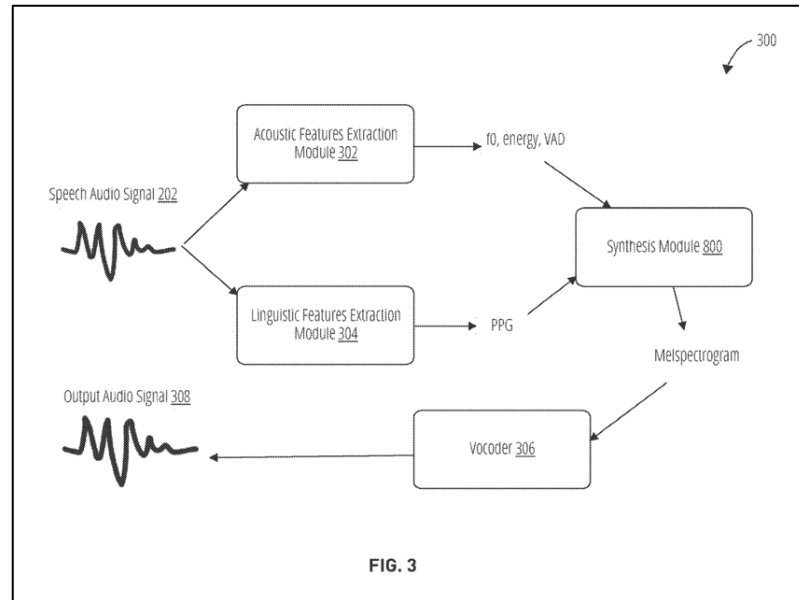
VI. OVERVIEW OF THE '457 PATENT

54. I have reviewed the '457 Patent (Ex. 1001). Descriptions of the '457 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the '457 Patent are intended only as a summary of the '457 Patent and/or intended as being pertinent to the Challenged Claims of the '457 Patent. My opinions provided herein are not intended as a discussion of my entire understanding of the '457 Patent or any other claims in any related patent to the '457 Patent.

A. Summary of the '457 Patent

55. The '457 Patent describes “real-time correction of an accent in a speech audio signal are provided.” 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. Field of Endeavor of the '457 Patent

56. I have been informed that the field of endeavor of the claimed invention can be determined by reference to explanations of the invention’s subject matter in the patent application, including the embodiments, function, and structure of the claimed invention.

57. The '457 Patent defines the “Technical Field” as relating to “audio processing” and, more specifically, to “systems and methods for real-time correction of accent in speech audio signals[.]” EX1001, 1:15-19. After reviewing the '457 Patent, it is my opinion that a PHOSITA would understand that the field of endeavor

of the '457 Patent includes speech audio processing. *See* EX1001, Abstract, 1:16-19, Fig. 3.

C. Problem Purportedly Solved by the Inventors of the '457 Patent

58. I have been informed that a prior art reference is “reasonably pertinent” if a PHOSITA would have consulted it and applied its teachings when faced with the problem that the inventor was trying to solve. As such, I have been asked to analyze the '457 Patent and determine the problem that the inventors were trying to solve.

59. The '457 Patent notes that a concern exists regarding the understandability between speakers possessing different accents during live video or audio conversations and therefore provides “real-time correction of accent in speech audio signals.” EX1001, 1:16-19, *See also* Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3. The '457 Patent also describes alleged improvements to audio processing quality for accent conversion, implying a need for advancements to systems and methods in audio processing quality for such real-time corrections. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3. That is, the '457 Patent details that a method for the “improved extraction” of acoustic and linguistic features for the generation of a modified, accented audio is desirable. EX1001, 14:15-19; *see also* 4:17-18, 14:15-19, Fig. 3. To solve these problems, the '457 Patent describes “processing the speech audio signal by a digital signal processing module to adjust

one or more characteristics of the speech audio signal” before “dividing the speech audio signal.” EX1001, 14:20-27; *see also* 11:64-67. Thus, a PHOSITA would understand that the ’457 Patent is directed to solving the problems of real-time audio modification, accent conversion, feature extraction from audio data, and audio processing quality improvements.

D. File History of ’457 Patent

60. I have reviewed the prosecution history for the ’457 Patent. The ’457 Patent claims priority to Provisional Application No. 63/297,901 filed January 10, 2022. The ’457 Patent did not face any office actions. EX1002.

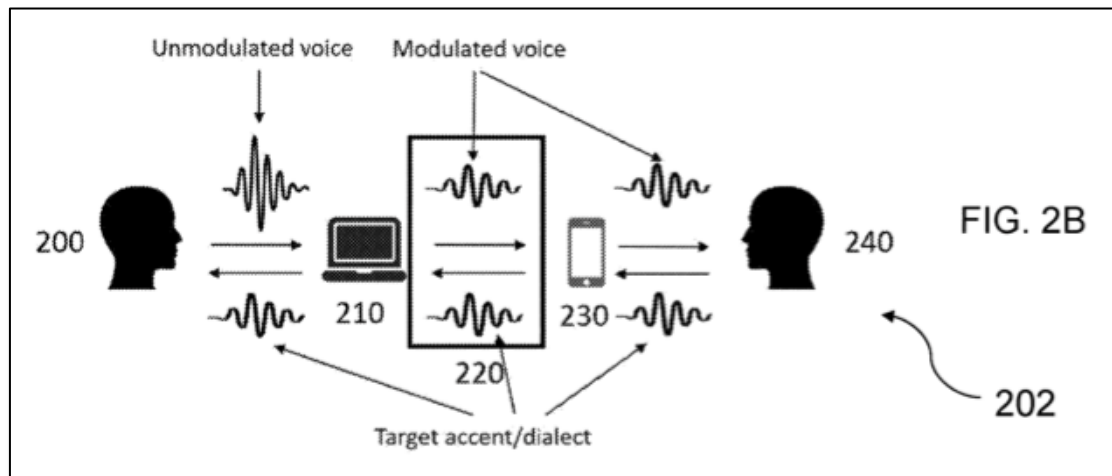
VII. OVERVIEW OF THE PRIOR ART

61. I have been informed that for a prior art reference to be proper for use in an obviousness analysis, the reference must be “analogous art” to the claimed invention. I have been informed that a prior art reference is analogous to the claimed invention if the reference is from the same field of endeavor as the claimed invention or if it is reasonably pertinent to the particular problem that the inventor was trying to solve.

62. The prior art is analogous to the claimed invention of the ’457 Patent, as discussed with greater detail for each prior art reference included. In general, the prior art is from the same field of endeavor of the ’457 Patent, speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3.

A. Overview of Feinauer

63. I have been informed that U.S. Patent Appl. No. 2020/0193971 to Feinauer et al. (“*Feinauer*”) was filed on December 13, 2018, and published on June 18, 2020. *Feinauer* teaches “methods for accent and dialect modification[.]” EX1004-Feinauer, Abstract.



EX1004-Feinauer, Fig. 2B.

64. A PHOSITA would have classified *Feinauer* within the same field of endeavor as the '457 Patent because both *Feinauer* and the '457 Patent relate to speech audio processing. Compare EX1004-Feinauer, Title, Abstract, [0001], [0091] with EX1001, Title, Abstract, 1:16-19.

65. Additionally, a PHOSITA would have found *Feinauer* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Feinauer* and the '457 Patent describe the difficulties posed by accent conversion. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3. For example, *Feinauer*

teaches “[c]ommunication between people with different dialects”, including “miscommunication, misunderstanding, and negative opinions based on suboptimal pairings of language, dialect and/or accent between conversation partners” as potential outcomes particularly without the context of visual cues. EX1004-Feinauer, [0002]-[0004]. *Feinauer* then identifies a system and method for addressing such problems via accent conversion, particularly the identification of an “accent and dialect modifier” utilized to modify “a spoken word to within a distance vector of the target threshold” of the desired accent. EX1004-Feinauer, [0086].

66. Further, the pertinence of *Feinauer* to the problem faced by the inventors of the ’457 Patent would have been identifiable by a PHOSITA as real-time audio modification. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1004-Feinauer, Abstract, [0002]-[0004], [0086].

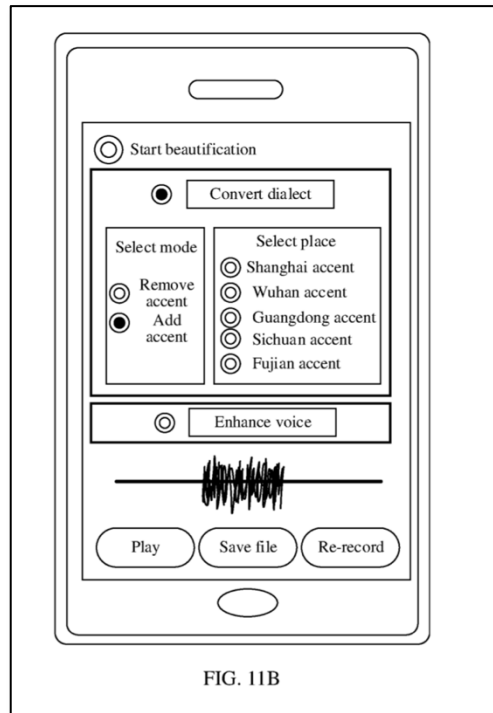
67. Further, the pertinence of *Feinauer* to the problem faced by the inventors of the ’457 Patent would have been identifiable by a PHOSITA as feature extraction from audio data. EX1001, 4:17-18, 14:15-19, Fig. 3. *Feinauer*’s real-time audio modification extracts “prosodic vocal features” from a detected dialect for the generation of the target audio data such that effective communication “across multiple dialects and/or accents” is achievable. EX1004-Feinauer, [0002]-[0004], [0086] [0095], [0099].

B. Overview of Fan

68. I have been informed that U.S. Patent Appl. No. 2023/0223006 to Fan et al. (“*Fan*”)¹ was filed on 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*. The US application incorporates both the PCT application (EX1015) and the Chinese application (EX1013) by reference. The PCT application (EX1015) incorporates the Chinese application (EX1013) by reference.

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

¹ I have been informed that EX1006 is the U.S. Patent Appl. to *Fan*. I have been informed that also included is the Chinese priority application (EX1013), the English translation of the Chinese priority application (EX1013, 1), and a certification that the translation is true and correct (EX1013, 63). I have been informed that also included is 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). I have been informed that *Fan* made a foreign priority claim within 12 months of the foreign priority application (EX1014, 3-4) and the Office acknowledged same (EX1014, 10).



EX1006-Fan, Fig. 11B.

70. A PHOSITA would have identified *Fan* as being within the same field of endeavor as the '457 Patent because both *Fan* and the '457 Patent relate, speech audio processing. EX1006-Fan, Abstract, [0134], [0201] with '457 *Patent*, Title, Abstract, 1:16-19, Fig. 3. *Fan* discloses a system having the capacity for “speaking style conversion on a to-be-converted first voice...[and] a dialect conversion mode.” EX1006-Fan, Abstract.

71. A PHOSITA would also have found *Fan* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Fan* and the '457 Patent describe accent conversion. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27; EX1006-Fan, Abstract, [0120], [0135], [0201].

72. Further, a PHOSITA would have found both *Fan* and the '457 Patent directed to feature extraction from audio data. EX1001, 4:17-18, 14:15-19, Fig. 3. Comparatively, *Fan*'s dialect conversion mode utilizes a voice feature extraction model to extract a phoneme posteriorgram feature from a first voice (the input voice) having a first accent and a style separation model to extract a style feature from the input voice. EX1006-Fan, Abstract, [0113], [0120], [0127], [0135], [0201].

73. The subject matter relied upon from *Fan* is supported in *Fan*'s priority application and PCT, as shown in the table below. I rely on disclosures in the patent and priority application, as set forth below:

<i>Fan</i> publication (EX1006) (figure or paragraph)	<i>Fan</i> PCT (EX1015) (page, figure or paragraph)	<i>Fan</i> Priority (EX1013) (page ² , figure or 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

² EX1013 pages 1-8 contain a listing of the claims. The specification follows, and the page numbers start over at this point. All page citations to EX1013 in this chart and in this Declaration refer to the page numbers assigned to the specification.

[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

C. Overview of Prabhavalkar

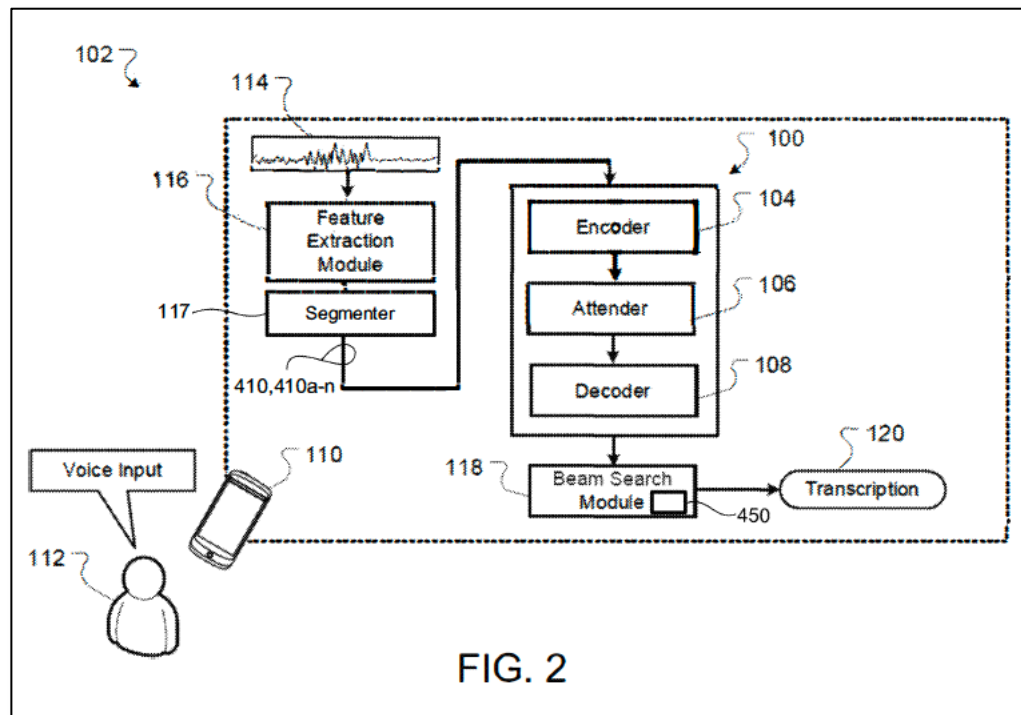
74. I have been informed that U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“*Prabhavalkar*”) was filed on September 20, 2021, and published October 1, 2024. *Prabhavalkar* determines, “from the audio data for the utterance, speech frames[,],” and “successively processing chunks that each include 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.

75. A PHOSITA would have identified *Prabhavalkar* as being within the same field of endeavor as the '457 Patent because both *Prabhavalkar* and the '457 Patent are both directed to speech audio processing. EX1001, Abstract, 1:15-25, Fig. 3. For example, *Prabhavalkar* teaches a method for “receiving audio data for an utterance, providing features indicative of acoustic characteristics of the utterance as input to an encoder neural network, and processing an output of the encoder neural network using an attender neural network to generate a context vector” as well as “generating speech recognition scores using the context vector and a decoder neural network.” EX1005-Prabhavalkar, 2:41-3:56. *See also* EX1005-Prabhavalkar, 6:56-7:4, 15:16-22.

76. Further, a PHOSITA would also have found *Prabhavalkar* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Prabhavalkar* and the '457 Patent describe real-time audio modification. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1005-Prabhavalkar, 17:46-51, 18:33-37, 20:16-21; 22:16-27.

77. Further, a PHOSITA would also have found *Prabhavalkar* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Prabhavalkar* and the '457 Patent describe feature extraction from audio data. EX1001, 4:17-18, 14:15-19, Fig. 3. Specifically, *Prabhavalkar* discloses that “[a] feature extraction module 116 processes the audio data 114 to extract (e.g., generate)

a set of feature values that are indicative of acoustic characteristics of the utterance. For example, the feature values may be mel-frequency cepstral coefficients.” EX1005-Prabhavalkar, 11:36-41.



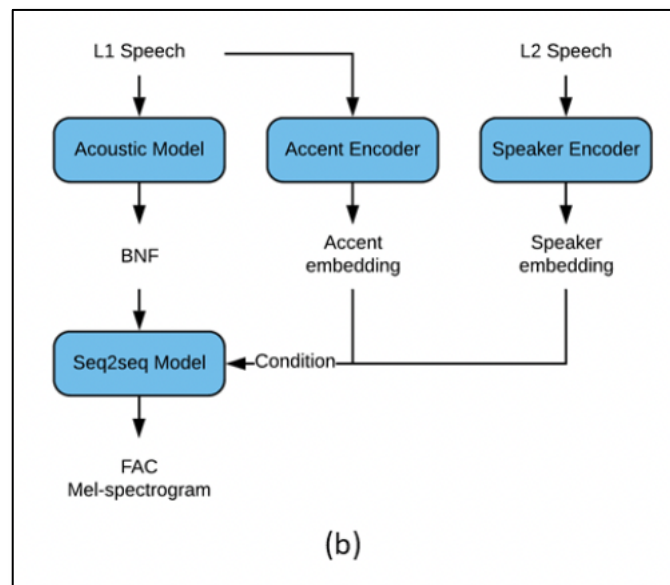
EX1005-Prabhavalkar, Fig. 2. See also EX1005-Prabhavalkar, 2:44-50.

78. Moreover, pertinence of *Prabhavalkar* to the purported problem described in the '457 Patent of audio processing quality improvement. EX1001, 11:64-67, 14:20-27. For instance, *Prabhavalkar* discloses “[v]arious techniques for enhancing speech recognition accuracy” including “implementations involv[ing] enhanced listen, attend, and spell (LAS) models, and others involv[ing] neural transducer models.” EX1005-Prabhavalkar, 1:29-35.

D. Overview of Accentron

79. I have been informed that “*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. EX1016-Munford, 7-35 (declaration of June Munford establishing public availability of *Accentron*), EX1046-Ching (affidavit of Mina Ching at the Internet Archive authenticating the support for *Accentron*’s public availability.)

80. *Accentron* teaches an approach that generates “accent-converted speech for arbitrary L2 speakers[.]” EX1008-Accentron, 1.



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

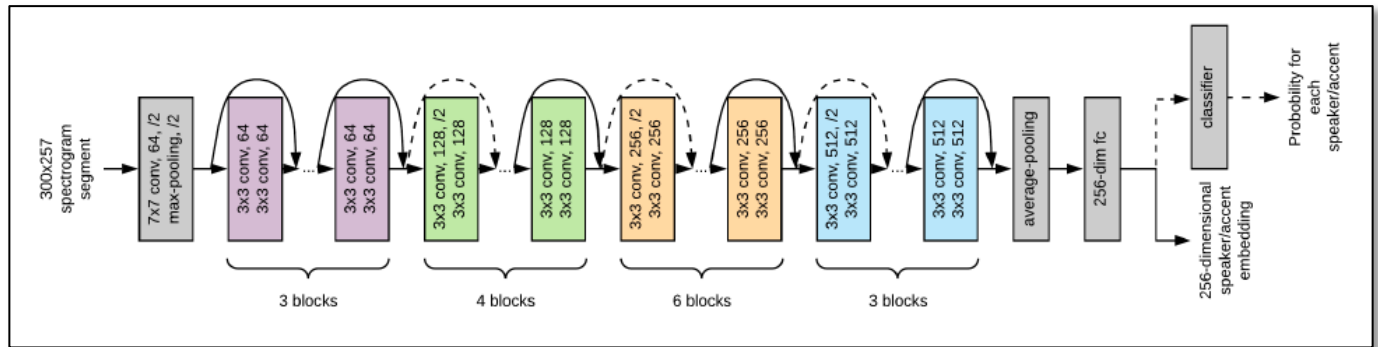
81. A PHOSITA would have identified *Accentron* as being within the same field of endeavor as the ’457 Patent because both *Accentron* and the ’457 Patent are directed to speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3. *Accentron*

teaches foreign accent conversion comprising the creation of a new voice having the “voice identity of a given second-language (L2) speaker but with a native (L1) accent.” EX1008-Accentron, 1. *Accentron* further discloses the generation of such voice (i.e., foreign accent conversion synthesis) via a model inference pipeline that may be accomplished in real-time. EX1008-Accentron, 33.

82. A PHOSITA would also have found *Accentron* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Accentron* and the '457 Patent address real-time audio modification. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1008-Accentron, 1, 33.

83. A PHOSITA would also have found *Accentron* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Accentron* and the '457 Patent address accent conversion. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1008-Accentron, 1.

84. A PHOSITA would also have found *Accentron* reasonably pertinent to the problem faced by the inventors of the '457 Patent because both *Accentron* and the '457 Patent address feature extraction from audio data. EX1008-Accentron, 1, 8, Fig. 1.



EX1008-Accentron, Fig. 2.

E. Overview of Singh

85. I have been informed that U.S. Patent Appl. No. 2019/0325861 to Singh et al. (“Singh”) was filed on April 18, 2019, and published on August 24, 2019. 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].

86. A PHOSITA would have identified *Singh* as being within the same field of endeavor as the ’457 Patent because both *Singh* and the ’457 Patent are both directed to speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3; EX1007-Singh, [0003]-[0004], [0007].

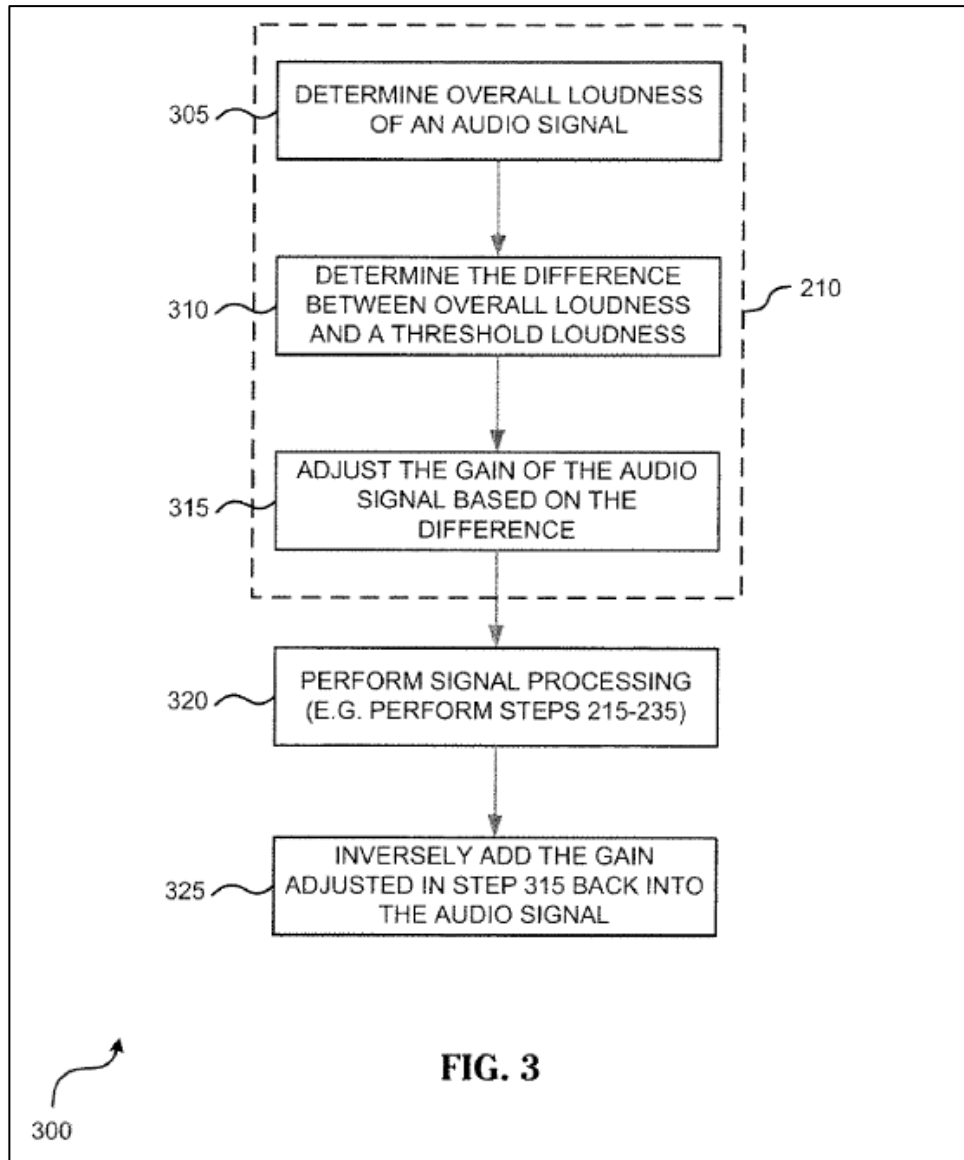
87. A PHOSITA would have found *Singh* reasonably pertinent to the particular problem faced by the inventors of the ’457 Patent including feature

extraction from audio data. EX1001, 4:17-18, 14:15-19, Fig. 3; EX1007-Singh, Abstract, [0003]-[0004], [0007], [0022].

88. A PHOSITA would have found *Singh* reasonably pertinent to the particular problem faced by the inventors of the '457 Patent including audio processing quality improvement. EX1001, 11:64-67, 14:20-27. *Singh* teaches the use of unsupervised domain adaptation to reduce the acoustic variability resulting from factors including, but not limited to, the gender and accent of a speaker. EX1007-Singh, [0009]. Correspondingly, *Singh* describes systems and methods capable of learning domain invariant features “directly from raw speech with significant improvement over the baseline acoustic models trained without domain adaptation. EX1007-Singh, [0009], [0022].

F. Overview of Bongiovi

89. I have been informed that U.S. Patent Appl. No. 2008/0219459 to Bongiovi et al. (“*Bongiovi*”) was filed on March 14, 2008, and published on September 11, 2008. *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]. *Bongiovi* teaches altering the loudness of the audio signal. EX1010-Bongiovi, [0046].



EX1010-Bongiovi, Fig. 3.

90. A PHOSITA would have understood *Bongiovi* as within the same field of endeavor as the '457 Patent because both Bongiovi and the '457 Patent are directed to speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3.

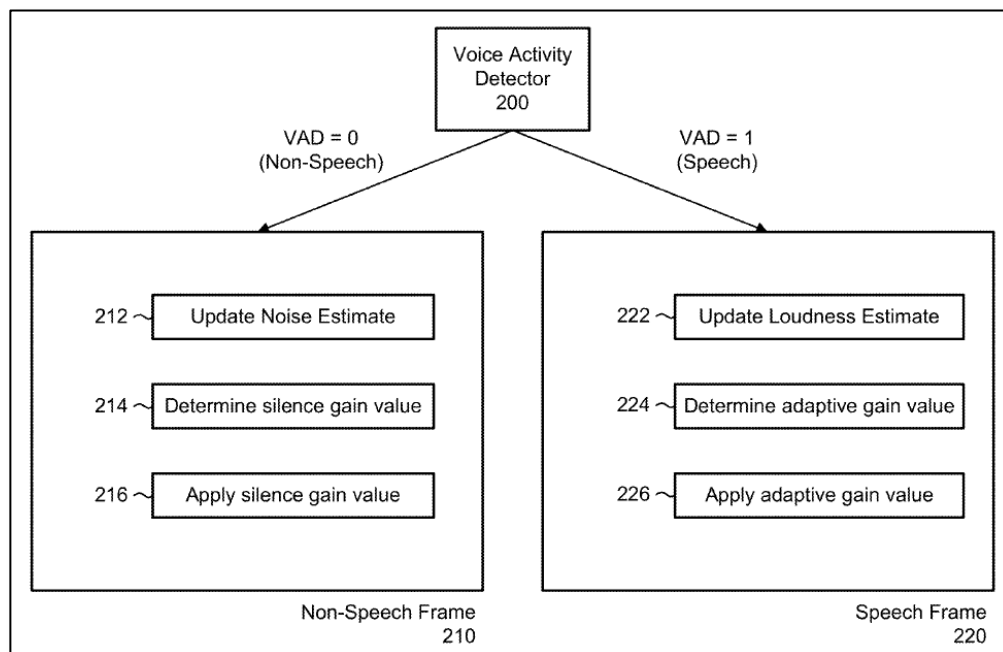
91. A PHOSITA would have found *Bongiovi* reasonably pertinent to the particular problem faced by the inventors of the '457 Patent including audio

processing quality improvement. EX1001, 11:64-67, 14:20-27; EX1010-Bongiovi, [0002]-[0003], [0034]-[0035].

G. Overview of Wu

92. I have been informed that U.S. Patent No. 11,164,592 to Wu et al. (“*Wu*”) was filed on May 9, 2019, and granted on November 2, 2021. *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.

93. A PHOSITA would have understood *Wu* as within the same field of endeavor as the ’457 Patent as both *Wu* and the ’457 Patent address speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3. *Wu* is directed toward “[e]lectronic devices...commonly used to receive audio data and generate output audio data based on the received audio data.” EX1009-Wu, 1:6-8.



EX1009-Wu, Fig. 2. *See also* EX1009-Wu, 5:46-47, 5:55-56.

94. Moreover, a PHOSITA would have identified *Wu* as reasonably pertinent to the problem faced by the inventors of the '457 Patent, audio processing quality improvement. EX1001, 11:64-67, 14:20-27. For example, *Wu* teaches “improving audio quality and maintaining dynamic range for an individual voice[.]” EX1009-Wu, 3:43-44, Abstract, 2:43-57. *See also* EX1009-Wu, Fig. 2.

H. Overview of *Osuna*

95. I have been informed that “Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion” to Zhao et al. (“*Osuna*”) was publicly available October 2019 in IEEE. EX1016-Munford, 44-47 (declaration of June Munford establishing public availability of *Osuna*).

96. *Osuna* teaches “accent conversion” using phonetic posteriorgrams (PPGs) to “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, 1651.

97. A PHOSITA would have understood *Osuna* as within the same field of endeavor as the '457 Patent as both *Osuna* and the '457 Patent address speech audio processing. EX1001, Abstract, 1:16-19, Fig. 3; EX1011-*Osuna*, Abstract.

98. A PHOSITA would have identified *Osuna* as reasonably pertinent to the problem faced by the inventors of the '457 Patent, accent conversion. EX1001, Title, Abstract, 1:16-19, 1:24-29, 4:23-27, Fig. 3; EX1011-Osuna, Abstract.

99. A PHOSITA would have identified *Osuna* as reasonably pertinent to the problem faced by the inventors of the '457 Patent, feature extraction from audio data. EX1001, 4:17-18, 14:15-19, Fig. 3; EX1011-Osuna, Abstract.

VIII. SUMMARY OF UNPATENTABILITY

100. I have reproduced the Proposed Grounds of Unpatentability from the Petition for ease of reference:

Proposed Grounds of Unpatentability	Exhibits
Ground 1: Claims 1-4, 8, 14-17, and 20 are obvious over Feinauer (EX1004), Prabhavalkar (EX1005), and Fan (EX1006)	EX1004, EX1005, EX1006
Ground 2: Claims 9-11 are obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006), and Accentron (EX1008)	EX1004, EX1005, EX1006, EX1008
Ground 3: Claims 6 and 19 are obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006), and Singh (EX1007).	EX1004, EX1005, EX1006, EX1007
Ground 4: Claim 7 is obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006), and Wu (EX1009).	EX1004, EX1005, EX1006, EX1009
Ground 5: Claims 12-13 are obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006), and Bongiovi (EX1010).	EX1004, EX1005, EX1006, EX1010
Ground 6: Claim 5 is obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006) and Osuna (EX1011).	EX1004, EX1005, EX1006, EX1011
Ground 7: Claim 18 is obvious over Feinauer (EX1004), Prabhavalkar (EX1005), Fan (EX1006) Osuna (EX1011) and Wu (EX1009).	EX1004, EX1005, EX1006, EX1009

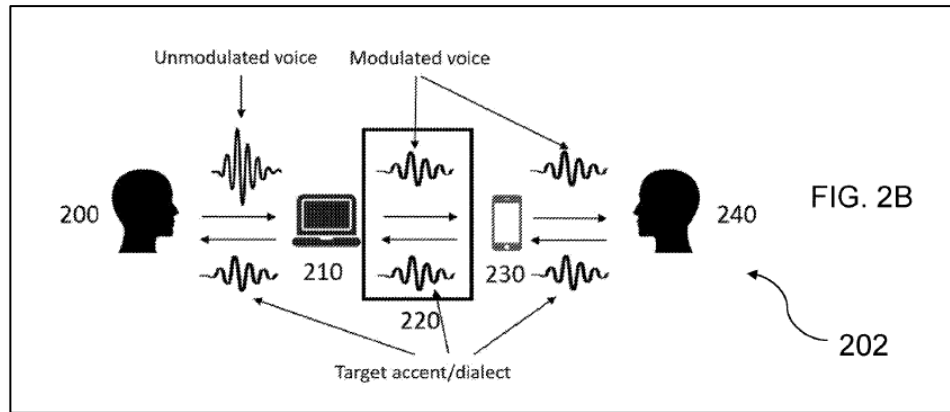
IX. OPINIONS REGARDING GROUND I: *FEINAUER, PRABHAVALKAR, AND FAN*

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:”*

101. 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 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.”

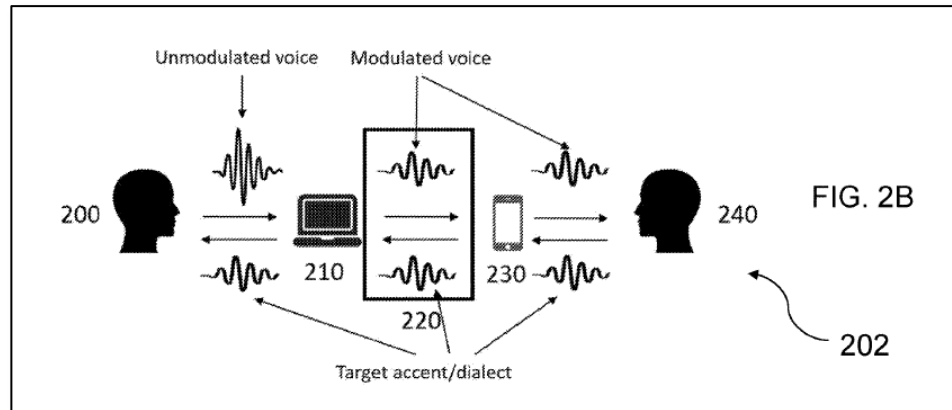
³ Claim language is *italicized*. All drawing annotations added. All emphases added unless otherwise noted.



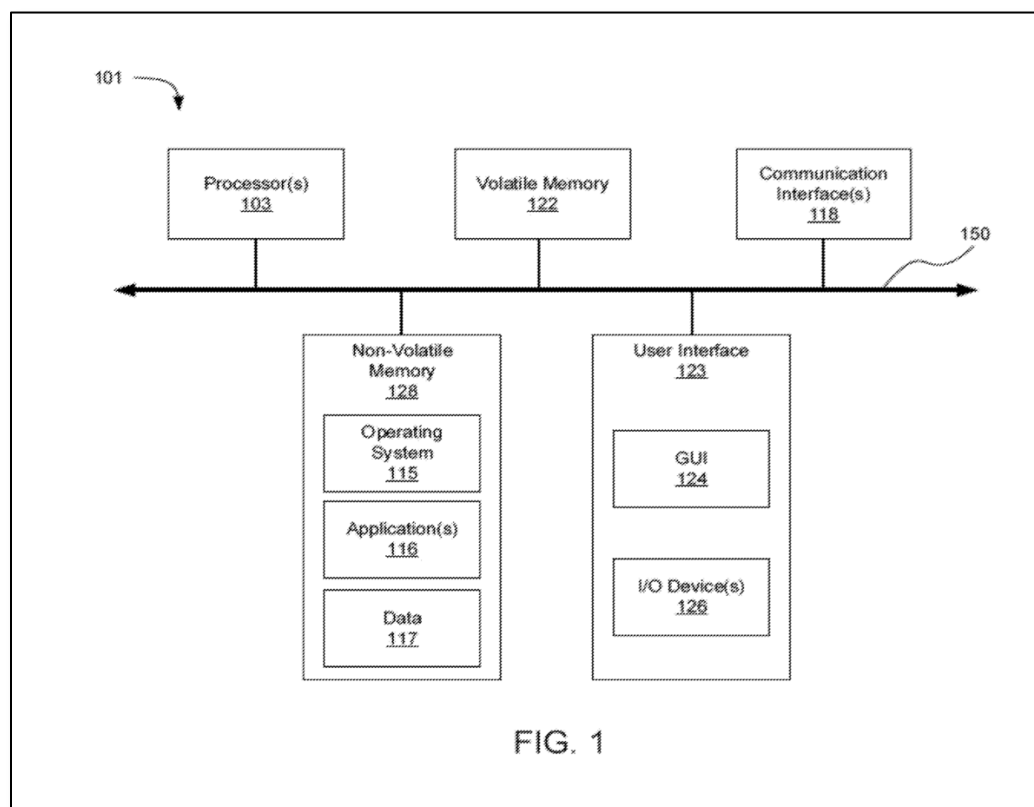
EX1004-Feinauer, [0035]-[0036], [0041], Fig. 2B.

102. *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]. It was well known that accent conversion may be performed in real time as demonstrated by *Feinauer*. As another example, *Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning* to Zhao et. al (“*Accentron*”) teaches, “accent conversion” where the “the inference can be accomplished in real time[.]” EX1008-Accentron, 33. A PHOSITA thus would have recognized *Feinauer* teaches a method for real-time correction of accent in a speech audio signal.

103. *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*. EX1004-Feinauer, [0121]. For example, *Feinauer* teaches that “communication device 230 can be a computing device 210.” EX1004-Feinauer, [0041].



EX1004-Feinauer, Fig. 2B. *Feinauer* teaches that communication device 230 may be “any embodiments of device 101, described above in connection with FIG. 1[.]” EX1004-Feinauer, [0043].



EX1004-Feinauer, Fig. 1.

104. *Feinauer's* computing device includes “one or more processors 103, volatile memory 122 (e.g., random access memory (RAM)), non-volatile memory 128” and processor “may be implemented by one or more programmable processors to execute one or more executable instructions, such as a computer program, to perform the functions of the system.” *See generally*, EX1004-Feinauer, [0031]-[0032]. *Feinauer's* function was to modify “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]. Therefore, a PHOSITA would have recognized that *Feinauer* teaches a real-time correction of an accent implemented by a computing device.

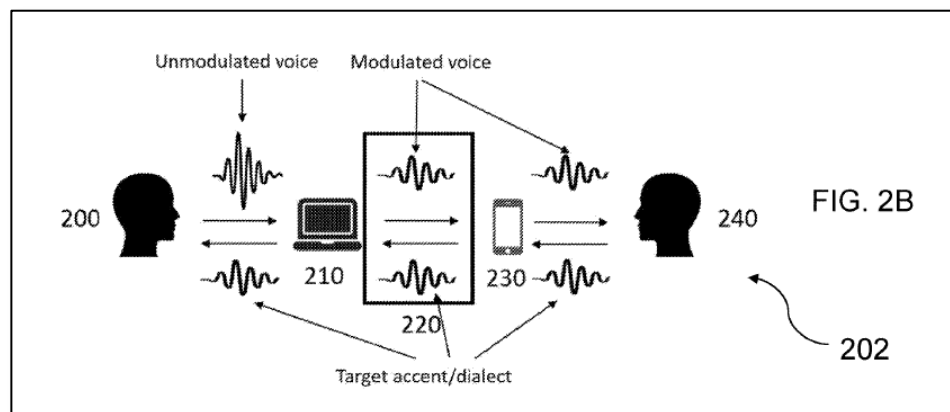
2. ***Claim 1(a): “dividing the speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”***

105. *Feinauer* as modified by *Prabhavalkar* renders obvious 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.

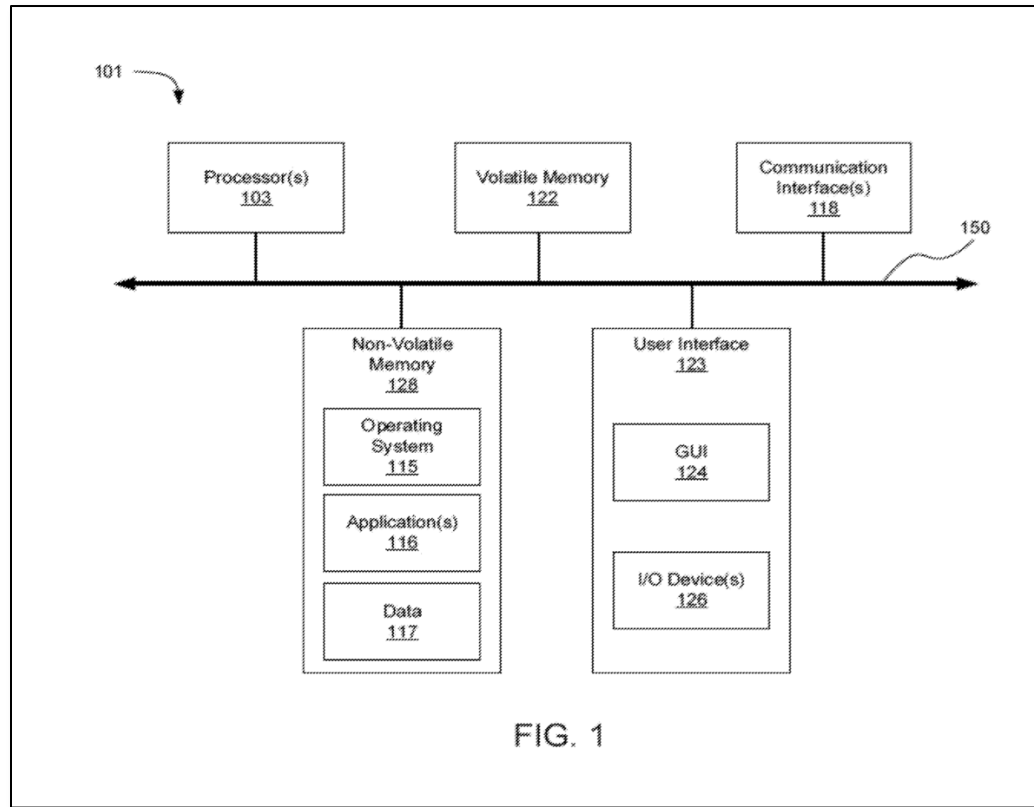
a. ***Feinauer's Teachings***

106. *Feinauer* teaches receiving an unmodulated voice signal (*speech audio signal*) via a microphone. *Feinauer* teaches that communication device 230 “may include any type and form of hardware and software for receiving, detecting or **being provide[d] voice or audio as input, such as via a microphone**” and which includes

“any type and form of hardware and software for playing, outputting or providing voice or audio as output, such as via a speaker.” EX1004-Feinauer, [0043] (emphasis added). *See also* EX1004-Feinauer, [0031]-[0032].



EX1004-Feinauer, Fig. 2B. Device 230 may be “any embodiments of device 101, described above in connection with FIG. 1” which includes “one or more input/output (I/O) devices 126 (e.g., ...a microphone...” EX1004-Feinauer, [0043].



EX1004-Feinauer, Fig. 1. It was well known to correct an accent spoken into a microphone. For example, *Accentron* teaches foreign accent conversion “has focused on acoustic methods, since they only require recording speech with a microphone.” EX1008-Accentron, 5.

107. *Feinauer* further teaches data may “represent...or contain...audio of conversation(s), such as audio fragments.” EX1004-Feinauer, [0053]. This audio fragment is represented by an “unmodulated voice or signal waveform. The unmodulated voice or signal waveform can be received by communication device 210 and transmitted via the communication network 220 without modification.” EX1004-Feinauer, [0044]. *Feinauer’s* device receives in real-time

data in the form of “data packets” that include unmodulated voice signal audio fragments from “I/O device(s)” which as discussed above includes a microphone. EX1004-Feinauer, [0044], [0053]. Therefore, a PHOSITA would have understood that *Feinauer* teaches receiving an unmodulated voice signal (*speech audio signal*) via a microphone. *See also* EX1004-Feinauer, [0042]-[0046], [0056].

108. *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. *Feinauer* discloses “the accent and dialect modifier 318 can determine that certain data portions of the audio content of data packet including a conversation, a conversation fragment, a sentence or a single word should be modified.” EX1004-Feinauer, [0054]. It was well known in the prior art that in order to perform accent conversion in real time, a conversion should be processed in “portions.” As an example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” to maintain “low real time delay” EX1018-

Wasserblat, Abstract, [0025]. In sum, 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*

109. *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. A PHOSITA would have appreciated that *Prabhavalkar* teaches a quick, non-textual, real-time under 300ms latency, speech processing system. EX1005-*Prabhavalkar*, 22:1-38. A PHOSITA would have appreciated that it was common in the prior art for real-time speech processing systems to not translate received speech audio to text as demonstrated by *Prabhavalkar*, allowing the system to provide a quick, low-latency, real-time speech modification experience to the user. *See, e.g.*, EX1008-Accentron, 14; 33.

110. *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. It is well known in speech processing to divide input audio into speech frames that include speech content from a speech utterance. Often these frames of speech are on the order of 10-40ms in length each, such as in *Prabhavalkar* where the speech frames are envisioned as being 30ms long. EX1005-*Prabhavalkar*, 15:18-20. *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[.]” EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Here, *Prabhavalkar* is describing a process by which incoming speech is divided into a string of 30ms-long speech frames. Every five (5) of these speech frames are apportioned into a 150ms chunk (much like *Feinauer’s* “portions”) for subsequent processing. EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Because every chunk has 5 frames, *Prabhavalkar’s* chunks include a *predetermined number* of frames of the speech audio signal. The chunks are processed “successively,” that is, one-at-a-time, though successive processing such does not preclude content from one chunk being used and/or considered relative to another chunk, for example to provide context between chunks. Thus, *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.

111. *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, 2:14-40. *Prabhavalkar* teaches that “streaming can be considered to mean that the system has a maximum allowable delay of 300ms.” EX1005-*Prabhavalkar*, 18:28-41. A PHOSITA would have appreciated that one example of “streaming” occurs with an “audio and/or video conversation” as taught by *Feinauer*. EX1004-*Feinauer*, [0041]. Therefore, as also discussed in paragraph 159, a PHOSITA would have thus recognized *Prabhavalkar*’s system as performing a *real-time* speech processing method, like *Feinauer*. *Prabhavalkar* envisions implementations may specifically generate an output based on input data representing no more than 300 ms of an utterance. EX1005-*Prabhavalkar*, 6:24-28.

c. *Motivation to Combine Feinauer and Prabhavalkar*

112. 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. *Feinauer* already teaches that “[t]he accent and dialect modifier 318 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.” EX1004-Feinauer, [0057]. (emphasis added) A PHOSITA would have understood that all this combination would entail is making sure there are frames within *Feinauer*’s portions, which was well within a PHOSITA’s skillset. A PHOSITA would have made any modifications necessary to make this modification operable since this combination was well within a PHOSITA’s skillset.

113. *Prabhavalkar* further gives details that a PHOSITA would have been motivated to look towards for this idea of “portions” that *Feinauer* teaches such that “speech frames each representing audio characteristics over different periods having a same duration; successively processing chunks that each include different sets of speech frames with the speech recognition model, wherein the speech recognition model is configured to predict a variable number of word elements for each chunk processed, wherein each chunk corresponds to a time step of processing of the speech recognition model, each chunk comprising a same predetermined number of speech frames...” EX1005-Prabhavalkar, 6:56-7:4,. *See also* EX1005-Prabhavalkar, 15:16-22, 18:18-22, 18:28-41, 22:1-38.

114. 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.

EX1005-Prabhavalkar, 17:47. Specifically, *Prabhavalkar* teaches that in a traditional LAS system, there is the “limitation that the entire utterance must be seen by the encoder, before any labels can be decoded...” EX1005-Prabhavalkar, 17:33-36. One way to “address this limitation” is to “operate on segments of an utterance less than the full utterance.” EX1005-Prabhavalkar, 17:37-43. This is referred to in *Prabhavalkar* as “a ‘chunk’-based approach” to “process attention on small segments of the utterance, such as a certain number of frames or a certain duration of the utterance at a time.” EX1005-Prabhavalkar, 17:43-46. To do this, *Prabhavalkar*’s model shifts “attention from one chunk to the next to **limit the amount of latency incurred**. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance while a second chunk or segment of the utterance is still being spoken.” EX1005-Prabhavalkar, 17:46-51 (emphasis added). Specifically, the time span necessary to produce an output would have been shorter in the combined *Fan-Prabhavalkar* system than in the unmodified *Fan* due to the improvements provided by *Prabhavalkar*’s chunking techniques.

115. 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. It was well known to process audio signals utterances in smaller portions, especially to process audio in real-time as taught by *Feinauer*. EX1004-Feinauer, [0057], [0086]. For example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” to maintain “low real time delay” EX1018-Wasserblat, Abstract, [0025]. 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. EX1005-Prabhavalkar, 17:15-16 (“an attention-based model can be designed for sequence-to-sequence speech recognition”). *Prabhavalkar* provides extensive support for processor architecture capable of implementing its method. For example, *Prabhavalkar* teaches the “processor 552 can process instructions for execution within the computing device 550, including instructions stored in the memory 564. The processor may also include separate analog and digital processors. The processor may provide, for example, for coordination of the other components of the device 550, such as control of user interfaces, applications run by device 550, and wireless communication by device 550. Processor 552 may communicate with a user through control interface 558 and display interface 556

coupled to a display 554.” EX1005-Prabhavalkar, 25:55-65. *See also* EX1005-Prabhavalkar, 6:56-7:4, 15:16-22.

116. To the extent necessary to implement *Prabhavalkar*’s chunking approach, a PHOSITA would have incorporated the enabling hardware from *Prabhavalkar* into *Feinauer*’s system to result in an operable real-time accent modification system performing chunking using ordinary skill without undue experimentation.

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;”*

117. *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. *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”

118. 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. As I described in the background section, it was common to measure all of these acoustic features in speech audio signals. See Section IV.B.

a) Feinauer’s Teachings

119. As discussed above regarding Claim 1(a), a PHOSITA would have modified *Feinauer’s* speech “portions” to be input chunks per *Prabhavalkar*. See Section IX.I.2.

120. *Feinauer* as modified by *Prabhavalkar* teaches extracting “prosodic vocal features and characteristics” from chunks of input speech audio. EX1004-*Feinauer*, [0099], Fig. 6 (step 620). *Feinauer* further envisions prosodic features including “any voice characteristic such as *pitch, tone, volume, etc.*” EX1004-*Feinauer*, [0057]. A PHOSITA would have understood pitch, tone, and volume to each satisfy “acoustic features” of the input chunk of user’s speech. A PHOSITA would have understood that one example of acoustic features are prosodic features. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches that prosody “refers to the spoken style of discourse independent of lexical content.” EX1021-Schmandt, 301-302. *Schmandt* further teaches stress, pitch, duration, and intonation are all examples of prosody speech style. EX1021-Schmandt, 301-302.

As another example, to calculate loudness, the energy of a signal is measured over time. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“*Fletcher*”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point.” EX1025-Fletcher, 83. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. In order to maintain or modify the voice characteristics identified in [0057], *Feinauer* must extract them, but does not provide further detail.

b) Fan’s Teachings

121. *Fan* teaches a voice conversion method for “adding an accent to or removing an accent from” a user’s voice. EX1006-Fan, Abstract. Like *Feinauer*, *Fan* teaches modifying acoustic features such as the user’s tone, pitch, rhythm, and speaking speed which are examples of prosody speech style. EX1006-Fan, [0050], [0106], EX1021-Schmandt, 301-302. As previously discussed at paragraph 30-31, *Schmandt* teaches that prosody (including features alternatively referred to as acoustic features) “refers to the spoken style of discourse independent of lexical content” such as stress, pitch, duration, rhythm, meter, and intonation. EX1021-Schmandt, 17, 301-302. Accordingly, a PHOSITA would have understood that the rhythm and speaking speed, defined broadly as prosody, are acoustic features.

122. *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. I note the description of the style enhancement model is discussed in *Fan* with regard to an embodiment employing a “third voice,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable to all of *Fan*’s envisioned embodiments.

c) Motivation to Combine Feinauer-Prabhavalkar and Fan

123. 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. 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] (“the voice processing neural network...includes...a dialect conversion network”); [0047]

(“Sequence to sequence (sequence to sequence, seq2seq) neural network: seq2seq belongs to an encoding (encoder)-decoding (decoder) structure.”); [0050] (“Style feature: The style feature includes but is not limited to a timbre. The timbre refers to quality of a voice, which is referred to as tone quality. The timbre can reflect unique quality of a sound made by a speaker. The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.”); [0090] (“Optionally, the style conversion network further includes a plurality of sub-feature extraction models and a multihead attention structure. The terminal inputs the third voice into the sub-feature extraction model, and extracts a vector of the sub-feature by using the sub-feature extraction model. For example, the plurality of sub-feature extraction models include a prosody extraction model, an accent extraction model, and a speaking speed extraction model. For example, the prosody extraction model is for extracting a prosody feature of the third voice, to obtain a prosody vector. The accent extraction model is for extracting an accent feature of the third voice, to obtain an accent vector. The speaking speed extraction model is for extracting a speaking speed feature of the third voice, to obtain a speaking speed vector.”); [0094] (“For example, a structure output by the multihead attention mechanism may be a 256-dimensional style embedding vector (a style feature).”); [0127] (“Optionally, the terminal inputs a third voice into a style separation model, and separates a style feature of the first voice by using the style separation model, to obtain a timbre feature vector of the

first voice. Then, 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. Content of the second voice is the same as content of the input voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”) [0152]-[0155] (“The obtaining module 1330 is configured to obtain a third voice for extracting a style feature. The processing module 1310 is further configured to input the third voice into the style separation model, and separate the style feature of the third voice by using the style separation model. The processing module 1310 is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion, to obtain the second voice. In an optional implementation, the style feature includes a first feature, and the first feature includes a plurality of sub-features.”); *see also* [0166]-[0167], [0170]-[0171], [0090]-[0093], [0046], [0048]-[0049], Fig. 5, Fig. 7.

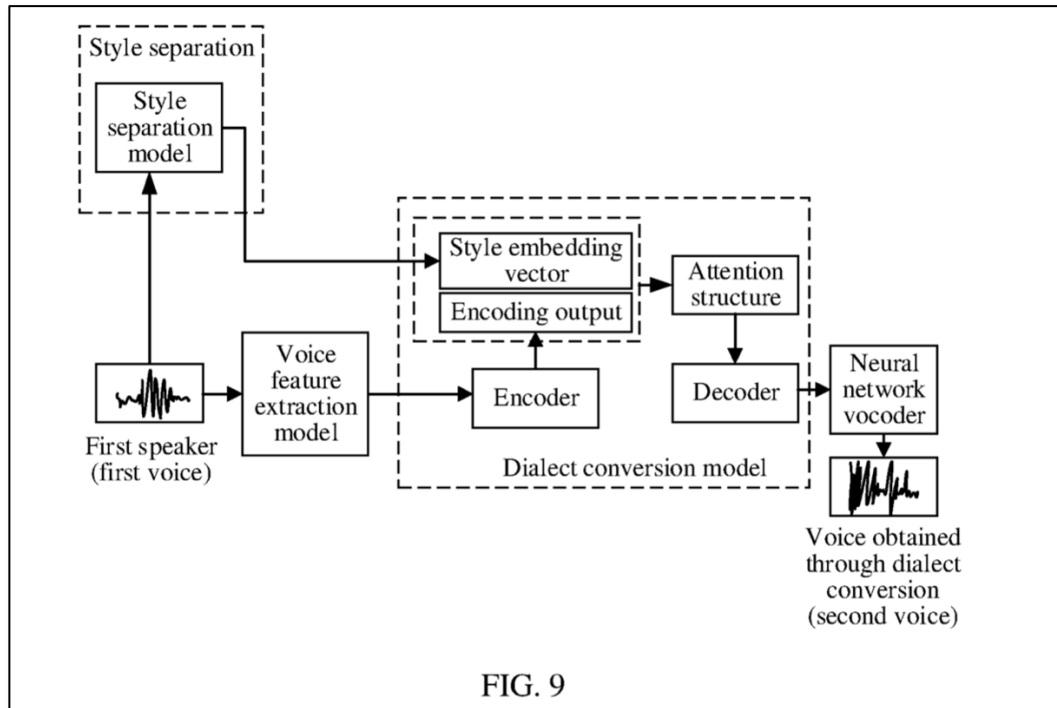


FIG. 9

EX1006-Fan, Fig. 9. *Feinauer* already teaches machine learning and neural network capabilities. EX1004-Feinauer, [0036]-[0037] (“a neural network machine learning model can detect a correlation between accent features and their success in achieving the optimal desired outcome in the specified context”), [0039] (“a neural network machine learning model can detect a correlation between dialect features and their success in achieving the optimal desired outcome in the specified context”), [0098] (“the features of the vocal characteristics in each category can be analyzed by a supervised neural network and the identified vocal characteristics correlated with their respective accent and dialect strength category labels.”). 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. *Feinauer* already teaches the “machine learning model 338 can extract vocal features of pronounced utterances” and that vocal features are “prosodic” EX1004-*Feinauer*, [0067]; [0095] (“prosodic vocal features and characteristics of the dialects can be extracted from the audio samples”); [0099]; [0100].

i. “a context associated with the input chunk”

124. 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).

125. 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 IX.I.2 citing EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Further, in the *Feinauer-Prabhavalkar-Fan* method, the style separation model (part of the mapped *acoustic feature module*) uses “attention” to “focus more on different key parts of an input instead of the entire input.” EX1006-Fan, [0048], [0090], [0094]. This was a very common way to focus on different parts of audio input in the prior art. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”). As another example, Zhao teaches forcing its decoder to compute the attention alignments from both the forward and backward directions during training, making

the decoder incorporate useful contextual information from both the past and future when producing the alignment. EX1045-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1008-Accentron, 11. (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”) *Prabhavalkar* explains that “attention” may be computed using chunks immediately preceding each input chunk. For example, *Prabhavalkar* teaches “the attender of the speech recognition model is configured to compute attention based speech frames in the current chunk and one or more previous chunks of the utterance.” 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

126. 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. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”); [0097] (“Because a length of the input sequence may be long, it is difficult for a vector to express abundant information during decoding. Therefore, the attention

mechanism is introduced. Then, a vector of the style feature and a PPG vector output by the encoder are spliced in width, to obtain an attention input matrix. Then, the attention input matrix is sent to the attention structure frame by frame, and the attention structure cooperates with the encoder to output a mel (or may also be referred to as a mel spectrum) feature of the second voice. The mel spectrum feature is a feature representation form of the second voice. After encoding ends, different weighting is performed on an output vector of an encoding module to obtain several vectors, and each vector corresponds to one output. This ensures that an output sequence is focused on an input sequence instead of depending on all input sequences.”). In another example, Zhao teaches forcing its decoder to compute the attention alignments from both the forward and backward directions during training, making the decoder incorporate useful contextual information from both the past and future when producing the alignment. EX1045-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1008-Accentron, 11. (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”).

127. *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* “compute[s]

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 (“each chunk having a first predetermined number of speech frames representing speech occurring before speech content being predicted at the current time step and a second predetermined number of speech frames representing speech occurring after the speech content being predicted in the current time step.”), 22:1-38. *Prabhavalkar teaches* “the attender is configured to compute attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:9-11. 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 (“As discussed above with reference to FIG. 1, attention-based models consist of three modules. First, an encoder module, represented by a multi-layer recurrent neural net-work (RNN), models the acoustics. Second, a decoder module, which consists of multiple RNN layers, predicts the output sub-word unit sequence. Third, an attention layer module selects frames in the encoder representation that the decoder should attend to when predicting each sub-word unit. Attention-based models, such as Listen, Attend and Spell (LAS) can be explored in

‘full-sequence’ mode, meaning attention is computed by seeing the entire input sequence. Thus, during inference, the model can produce the first output token only after all input speech frames have been consumed.”), 22:1-38; EX1001, 1:64-66 (“The context may include a pre-determined number of the frames belonging to chunks preceding the input chunk in the 65 stream of input chunks.”), 6:55-63 (“An input to module 302 may include a chunk 408 of frames concatenated with a context 402. The context 402 may include a pre-determined number of frames of the speech audio signal 202 preceding the chunk 408. Context 402 can be stored in a cache of module 302 and continuously updated. Thus, at each state of real time, input of the modules can include the chunk 408 ended at the previous state of real time and the context 402 corresponding to the chunk.”), 17:1-4 (Claim 8). It was well known that attention is a type of context. EX1008-Accentron, 12 (“the target speech using the *attention context*”); EX1008-Accentron, Table 1.

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).

e) Motivation to Combine Prabhavalkar’s Context

128. 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*. EX1005-*Prabhavalkar*, 22:1-38 (“a chunk size of 5 (300 ms latency) is the requirement for allowed streaming delay”). *Prabhavalkar teaches* “the attender is configured to compute attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:9-11. *Prabhavalkar teaches* that “streaming can be considered to mean that the system has a maximum allowable delay of 300 ms.” EX1005-*Prabhavalkar*, 18:1-41. A PHOSITA would have appreciated that one

example of “streaming” occurs with a “real-time” “audio and/or video conversation” as taught by *Feinauer*. EX1004-Feinauer, [0041]; [0086]. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-Prabhavalkar, 17:33-51.

129. 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*. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”); [0097] (“Because a length of the input sequence may be long, it is difficult for a vector to express abundant information during decoding. Therefore, the attention mechanism is introduced. Then, a vector of the style feature and a PPG vector output by the encoder are spliced in width, to obtain an attention input matrix. Then, the attention input matrix is sent to the attention structure frame by frame, and the attention structure cooperates with the encoder to output a mel (or may also be referred to as a mel spectrum) feature of the second voice. The mel spectrum feature is a feature representation form of the second voice. After encoding ends, different

weighting is performed on an output vector of an encoding module to obtain several vectors, and each vector corresponds to one output. This ensures that an output sequence is focused on an input sequence instead of depending on all input sequences.”). As another example, Zhao teaches “by forcing the decoder to compute the attention alignments from both the forward and backward directions during training, we can make the decoder incorporate useful contextual information from both the past and future when producing the alignment.” EX1045-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1008-Accentron, 11. (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”).

130. 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. As one example, Zhao teaches “by forcing the decoder to compute the attention alignments from both the forward and backward directions during training, we can make the decoder

incorporate useful contextual information from both the past and future when producing the alignment.” EX1045-Zhao, 2368. 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.

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

131. *Feinauer* in view of *Fan* teaches *extracting, by an acoustic features extraction module (Fan’s “voice feature extraction model”), linguistic features including words (per Feinauer) which include phonemes represented in the form of phonetic posteriorgrams (PPGs) (per Fan) with a reduced accent. 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. *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.

132. As discussed above regarding Claim 1(b), *Feinauer-Prabhavalkar-Fan* renders obvious extracting acoustic features from an input chunk and context. See Section IX.I.3. Further, *Feinauer* teaches identifying “individual words” in the user’s speech, and may determine “which words” to be accent-modified. EX1004-

Feinauer, [0083] (“an algorithm will measure each individual word in the conversation to determine which words to modify”), [0094] (“an algorithm will measure each individual word in the conversation to determine which words to modify.”), [0099] (“At step **620**, the accent and dialect modifier can define feature values of words. For example, prosodic vocal features and characteristics of individual **word dialects can be extracted from the individual word audio samples** using preprocessing tools to convert the audio sample waveforms into a spectrogram of MFCC. In some embodiments, the accent and dialect modifier can query the database to obtain instructions, data or information to define feature values of words. In some embodiments, the accent and dialect modifier can define the feature values of all the words in a conversation, conversation packet or voice data packet.”). A PHOSITA would have recognized that *Feinauer* thus extracts linguistic features, but does not provide further detail regarding how linguistic features are extracted, or by which module.

133. 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

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

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] (“The terminal inputs the feature information of the first voice into a voice feature extraction model, and extracts a phoneme posteriorgram PPG feature of the first voice by using the voice feature extraction model, where the PPG feature is for retaining the content information of the first voice. The PPG feature 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. In this example, the voice does not need to be converted into a text to retain the content information of the voice, instead, voice input is directly performed, and the content information of the first voice is retained by using the PPG feature, so that robustness can be improved.”), 85-86 (“The terminal inputs the first voice into the voice feature extraction model, and extracts the PPG feature of the first voice by using the voice feature extraction model.”), 100, Fig. 5, Fig. 9, 126-127, 150 (“In an optional implementation, the processing module **1310** is further configured to input the feature information of the first voice into a voice feature extraction model, and extract a phoneme posteriorgram PPG feature of the first voice by using the voice

feature extraction model, where the PPG feature is for retaining content information of the first voice.”). Fan’s PPGs are linguistic features because they represent phoneme probabilities and phonemes are a linguistic feature of any language. EX1006-Fan, [0085] (“The PPG feature describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and is equivalent to recognizing the phoneme.”). *See* EX1045-Zhao, 2368lhc.

136. 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). EX1006-Fan, [0023], [0085], Fig. 5. I note the description of a voice feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model in regard to a voice feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable to all of *Fan*’s envisioned embodiments.

137. 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

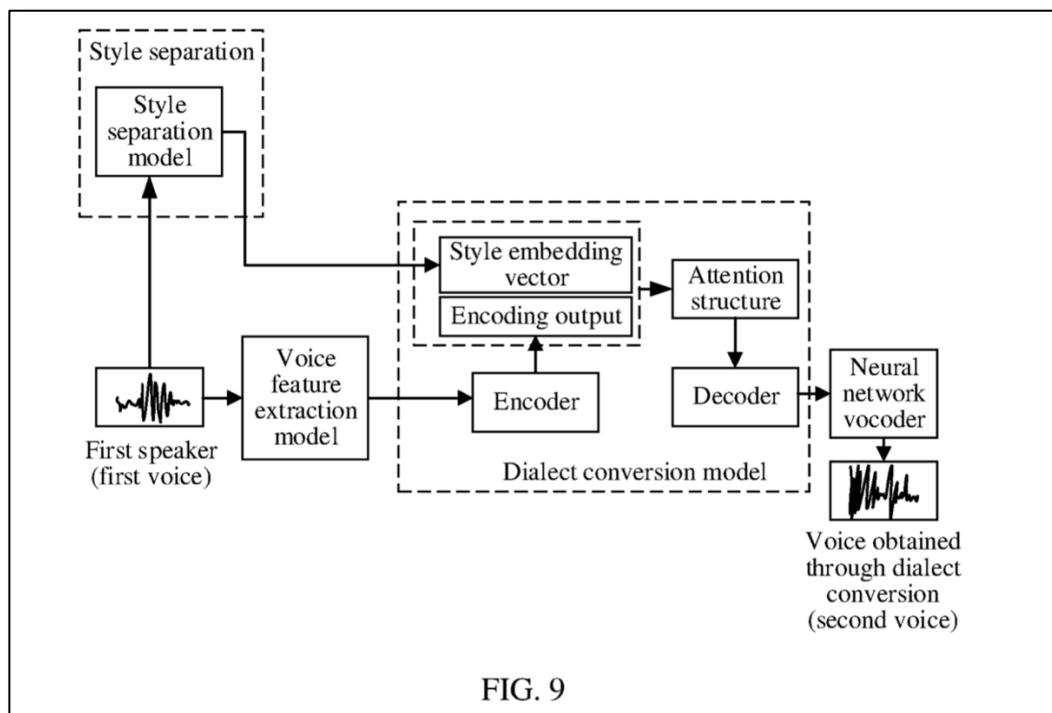
PPG feature is for retaining the content information of the first voice.” EX1006-Fan, [0007], [0085]. As supported by the prior art and the ’457 specification, a PHOSITA would have recognized that PPGs extracted from a user’s speech carry a reduced accent. *See* EX1001, 6:24-35, 13:22-32; EX1012-Ding, Abstract. Specifically, as supported by *Ding*, PPGs are mostly speaker-independent, but do carry some accent that may leak through to subsequently-generated speech. *Id.* For example, *Ding* teaches most methods “implicitly assume that PPGs are speaker-independent and contain only linguistic information in an utterance. In practice, however, PPGs carry speaker individuality cues, such as accent, intonation, and speaking rate. As a result, these cues can leak into the voice conversion, making it sound similar to the source speaker.” EX1012-Ding, Abstract. Therefore, it was well known that PPGs will still retain some accent within the PPG, though at a reduced amount. Therefore, PPGs satisfy *linguistic features with a reduced accent*. Thus, *Prabhavalkar-Fan* renders obvious *extracting linguistic features with a reduced accent from an input chunk*.

b) Motivation to Combine Feinauer-Prabhavalkar and Fan

138. 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. 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] (“the voice processing neural network...includes...a dialect conversion network”); [0047] (“Sequence to sequence (sequence to sequence, seq2seq) neural network: seq2seq belongs to an encoding (encoder)-decoding (decoder) structure.”); [0085] (“the voice feature extraction model may be a deep neural network.”); [0127] (“Optionally, the terminal inputs a third voice into a style separation model, and separates a style feature of the first voice by using the style separation model, to obtain a timbre feature vector of the first voice. Then, 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. Content of the second voice is the same as content of the input voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”); [0152]-[0155] (“The obtaining module 1330 is configured to obtain a third voice for extracting a style feature. The processing module 1310 is further configured to input the third voice into the style separation model, and separate the style feature of the third voice by using the style separation model. The processing module 1310 is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion, to obtain the

second voice. In an optional implementation, the style feature includes a first feature, and the first feature includes a plurality of sub-features.”); see also [0166]-[0167]; [0170]-[0171]; [0090]-[0093]; [0046], [0048]-[0049]; Figs. 5, 7.



EX1006-Fan, Fig. 9. *Feinauer* already teaches machine learning and neural network capabilities. EX1004-Feinauer, [0036]-[0037] (“a neural network machine learning model can detect a correlation between accent features and their success in achieving the optimal desired outcome in the specified context”), [0039] (“a neural network machine learning model can detect a correlation between dialect features and their success in achieving the optimal desired outcome in the specified context”), [0098] (“the features of the vocal characteristics in each category can be analyzed by a supervised neural network and the identified vocal characteristics correlated with

their respective accent and dialect strength category labels.”). 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. 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. A PHOSITA would have recognized PPGs were a well-known and widely-used tool for extracting linguistic features, and would have expected a speech modification system to be capable of extracting PPGs, further motivating the combination.

i. “a context associated with the input chunk”

139. As discussed regarding Claim 1(b), in the *Feinauer-Prabhavalkar-Fan*’s 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 IX.I.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. A PHOSITA would have appreciated that this combination would have saved time and improved

performance. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”); [0097]. As another example, Zhao teaches “by forcing the decoder to compute the attention alignments from both the forward and backward directions during training, we can make the decoder incorporate useful contextual information from both the past and future when producing the alignment.” EX1045-Zhao, 2368. The improved speed and efficiency would have reduced the lag experienced by a user, desirably improving the system and motivating the combination. 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”*

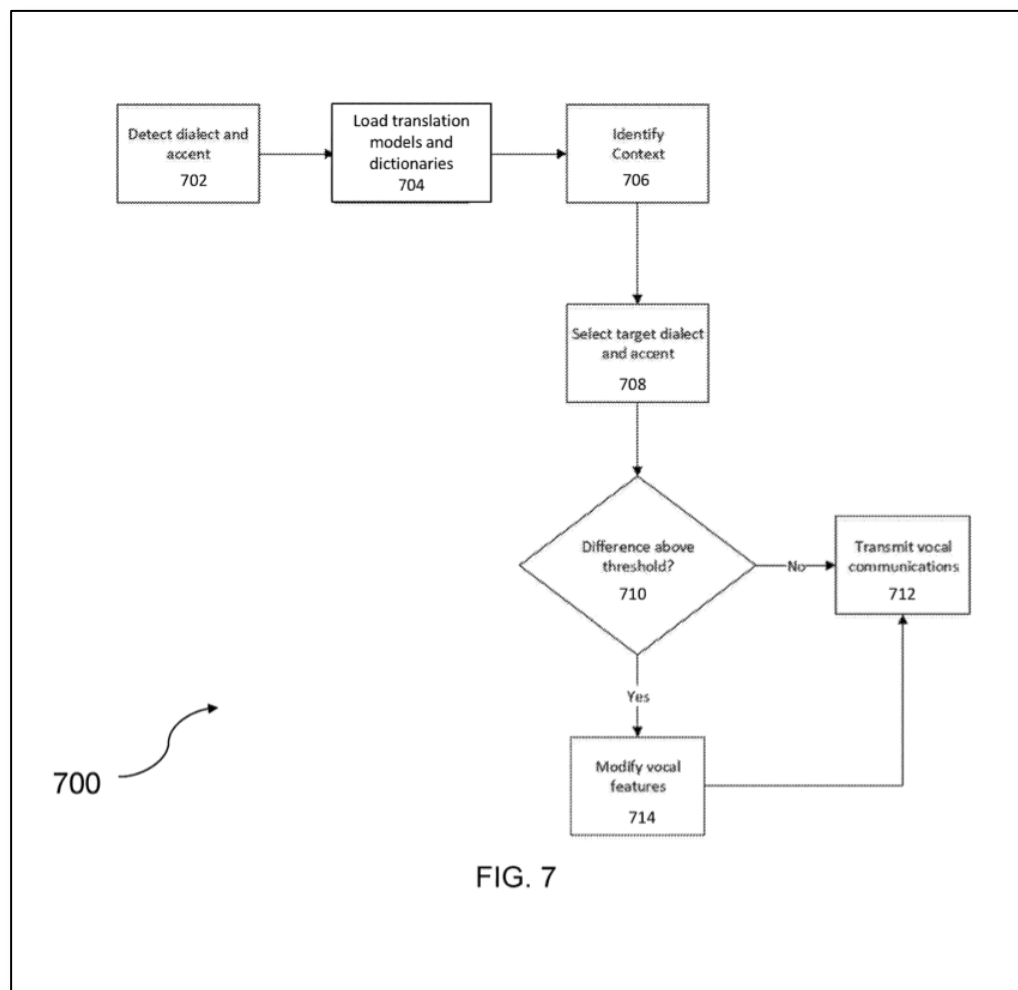
140. The ’457 Patent describes a speaker embedding as a “multidimensional vector” that is “pretrained based on audio data including a recorded speech of a target speaker having a further accent.” EX1001, 13:33-38. The ’457 Patent further describes the speaker embedding may include other “information 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. EX1001, 9:58-67.

141. *Feinauer* in view of *Fan* teaches receiving a speaker embedding for a human speaker (*Fan*'s "style embedding vector"). *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]. Although *Feinauer* does not have much detail as to how *Feinauer*'s voice profile is expressed, a PHOSITA would have been motivated to look towards *Fan* to find information on *Fan*'s "style embedding vector."

a. *Feinauer's Teachings*

142. *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], ("step 504, the accent and dialect modifier preloads relevant translation models and translation dictionaries. For example, the relevant translation models and translation dictionaries can be installed on a local machine, such as computing device 210, 230"), [0105] ("step 704, the accent and dialect modifier loads translation models and dictionaries"), [0110], Fig. 7. *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]. *Feinauer’s* translation model includes “English received pronunciation with the American way of emphasizing certain words.” EX1004-Feinauer, [0065]. *Feinauer* teaches that “English received pronunciation may be recognized as being more persuasive or trustworthy, but American clients may understand familiar American words, such as brand names, more easily in their native American accent compared to with an English accent.” EX1004-Feinauer, [0065].



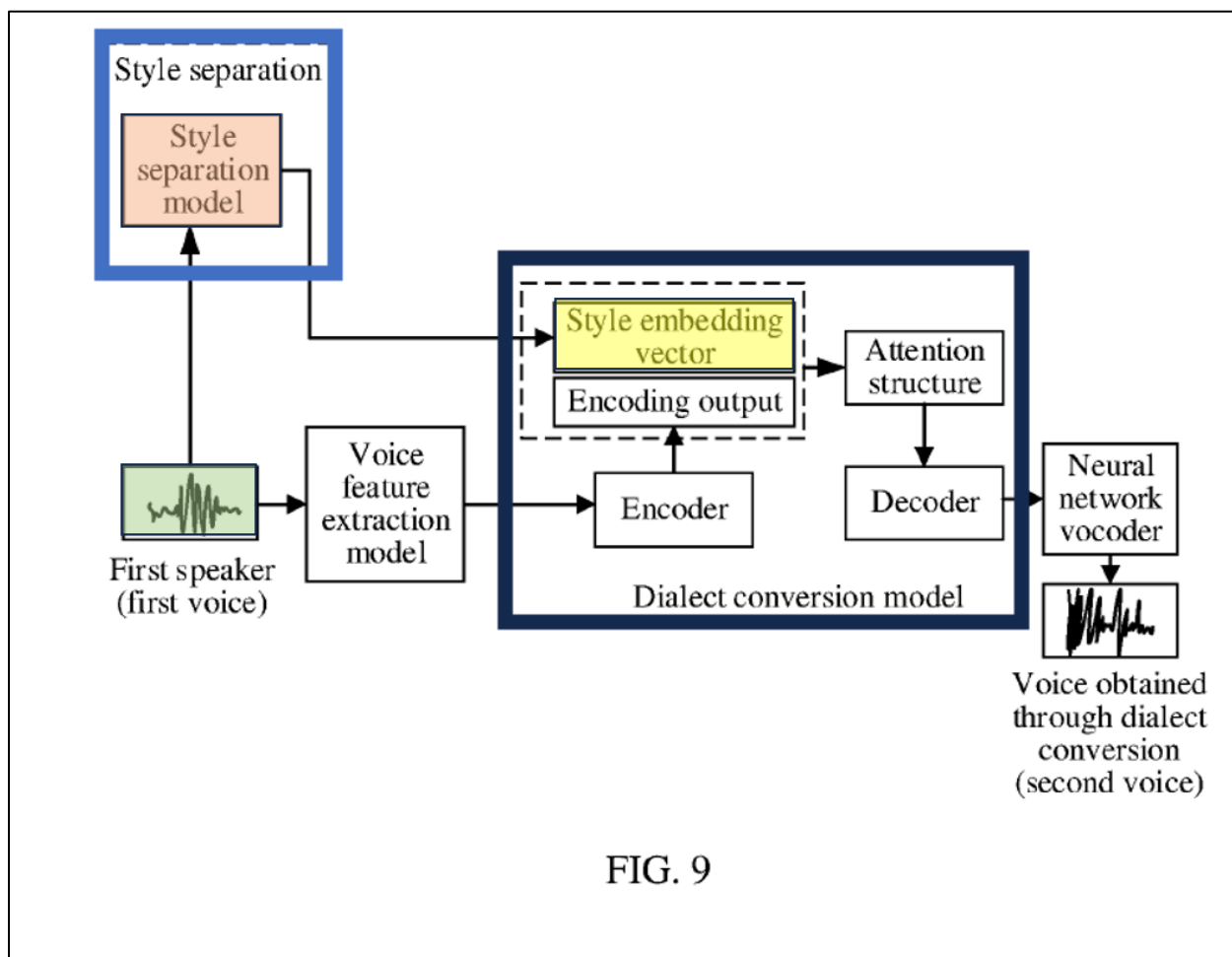
EX1004-Feinauer, Fig. 7. *Feinauer* does not provide further detail regarding the contents of its translation model.

b. *Fan's Teachings*

143. 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]. *Fan* envisions a “style embedding vector” that a PHOSITA would have recognized includes a representation of a target accent. EX1006-Fan, [0050], [0094], [0127], Figs. 7, 9, 12. It was well known in the prior art that a “speaker embedding vector” at least includes “timbre” as taught by Fan. EX1008-Accentron, 21 (“speaker embedding mainly encodes information such as voice quality/timbre and pitch”).

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

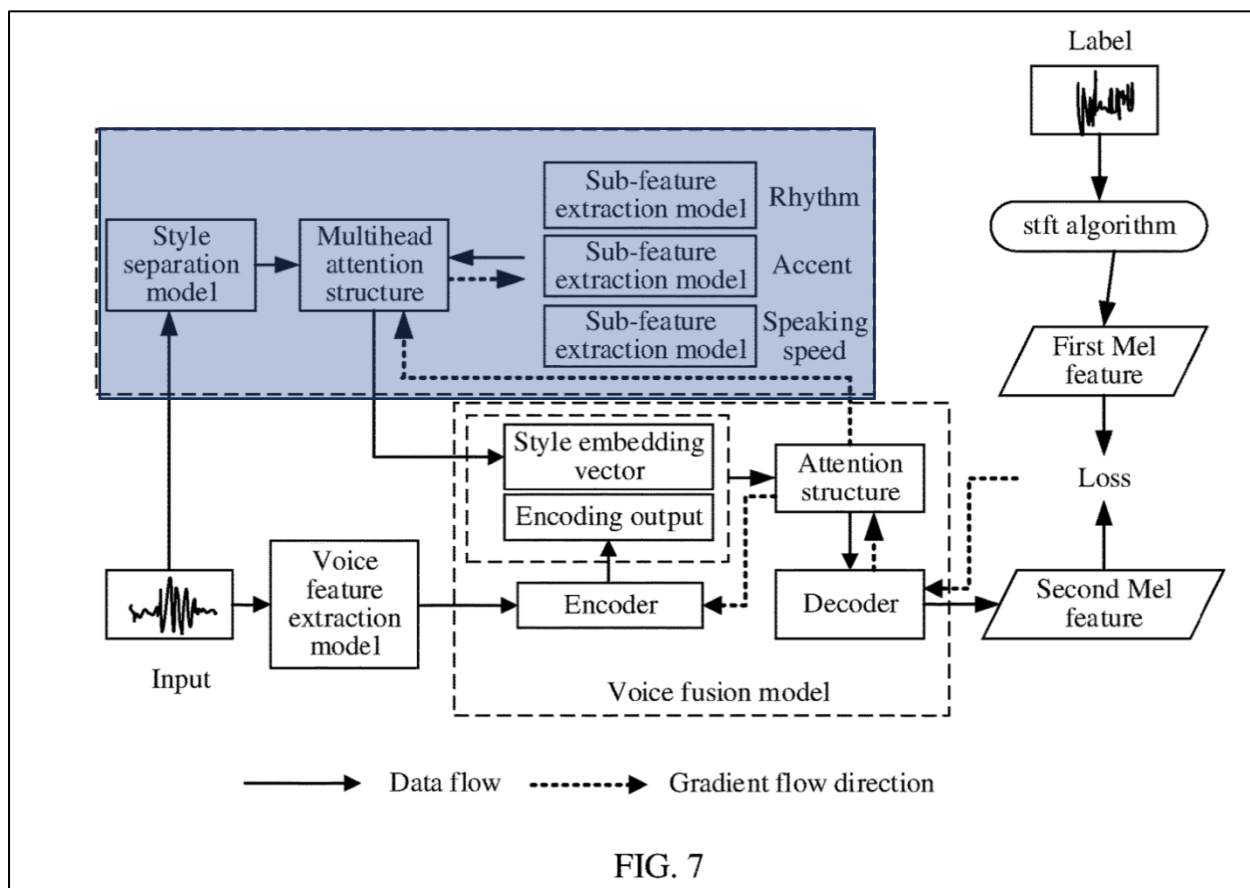
[0127].



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.

145. *Fan* teaches that the features from the target voice are extracted in the style separation model and put into a “multihead attention structure” to enable “a model to extract a feature expression from different subspaces.” EX1006-Fan,

[0094], Figs. 7, 9, 12 (annotated in blue). An output from the “the multihead attention mechanism includes some timbres of the third voice” in the form of a “256-dimensional style embedding vector (a style feature).” EX1006-Fan, [0094].



EX1006-Fan, Fig. 7. *Fan*’s timbre reflects the “unique quality of a sound made by a speaker.” EX1006-Fan, [0050].

146. *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. EX1006-Fan, [0050],

[0094], [0127], Figs. 7, 9, 12. *Fan* teaches this style feature includes “but is not limited to a timbre. The timbre refers to quality of a voice, which is referred to as tone quality. The timbre can reflect unique quality of a sound made by a speaker. The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.” EX1006-Fan, [0050]. It was well known in the prior art that a “speaker embedding vector” at least includes “timbre” as taught by Fan. EX1008-Accentron, 21 (“speaker embedding mainly encodes information such as voice quality/timbre and pitch”). For these reasons, *Fan*’s style embedding vector teaches a *speaker embedding*.

147. 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 style features (*e.g.* acoustic content) of the first (input) voice are input into the dialect conversion model (annotated with a black box) corresponding to the selection operation, and the second (modified) voice is output by using the dialect (accent) conversion model corresponding to a style embedding vector (speaker embedding). EX1006-Fan, [0127], Fig. 9.

c. *Motivation to Combine Fan’s speaker embedding*

148. 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” within *Feinauer*’s voice profile 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] (“At step 504, the accent and dialect modifier pre-loads relevant translation models and translation dictionaries...the accent and dialect modifier identifies relevant translation models and dictionaries based on the content and/or language, accent and/or dialect the accent and dialect modifier is to use or operate[,]...[and] the accent and dialect modifier loads relevant translation models and dictionaries prior to, at the beginning of a conversation or during the conversation”), [0105] (“...[T]he accent and dialect modifier loads translation models and dictionaries. Loading translation models and dictionaries 704 can include preparing relative translation models and dictionaries that are to be accessed and used for the appropriate target dialect and/or accent, such as the dialect and/or accent to be selected for the context.”), 81, 105, Fig. 7; EX1006-Fan, [0010] (“a similarity between styles of the to-be-converted voice and the third voice...is adjusted by using the weight that corresponds to each sub-feature and that is input by the user”); [0081]-[0082] (“the terminal device selects, based on the sub-template selection operation, a target voice corresponding to a sub-template...”); [0094] (“[t]he terminal multiplies the weight that is input by the user and that corresponds to each sub-feature and the vector of each sub-feature...[a] result obtained through

the multiplication and the vector (for example, a timbre vector) of the first feature are simultaneously input into a multihead attention structure for attention alignment, so that the multihead attention structure outputs a style vector, where the style vector is the style feature of the third voice”); [0126]-[0129](“ Refer to FIG. 9 . When the mode selection operation is for selecting the dialect switching mode, the terminal inputs a PPG feature into a dialect switching model (a seq2seq neural network) based on the mode selection operation, and outputs a melt feature form of the second voice by using the dialect switching model. Then, a melt feature is converted into a playable voice form by using a neural network vocoder. The first voice is a voice of a first dialect, and the second voice is a voice of a second dialect...Then, 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. Content of the second voice is the same as content of the input voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”);[0160]-[0162] (“In an optional implementation, the obtaining module 1330 is further configured to: receive a template selection operation input by the user,... obtain a voice corresponding to the target template, and use the voice corresponding to the target template as the third voice. [T]he obtaining module 1330 is further configured to receive the third voice input by a second speaker, where the

first voice is a voice of a first speaker, and the second speaker is any person different from the first speaker. [T]he target conversion mode is the dialect conversion mode, and the target voice conversion network is a dialect conversion network. The processing module 1310 is further configured to input the feature information of the first voice into the dialect conversion network, and output the second voice over the dialect conversion network, where the first voice is a voice of a first dialect, and the second voice is a voice of a second dialect.”), Figs. 9, 12-13.

149. 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. I note the description of a style feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model in regard to a style feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable to all of *Fan*’s envisioned embodiments.

150. As discussed, *Feinauer* itself teaches extracting linguistic and acoustic features from a user's input speech. *See*, Sections IX.3-IX.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]. *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 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*.

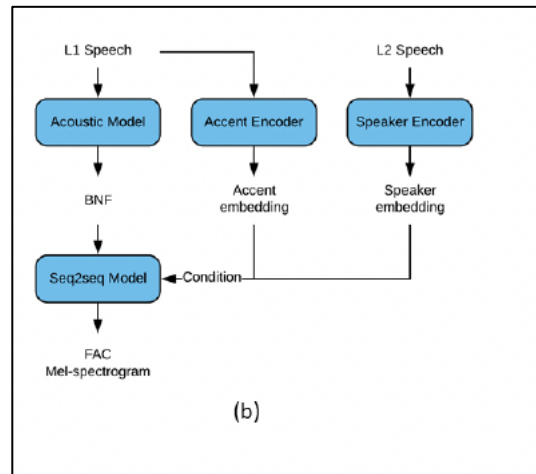
151. A PHOSITA would have combined the teachings such that the method of *Feinauer*, in which accent modification is performed as detailed by *Fan* regarding Fig. 9, in which linguistic content (*e.g.*, comprising PPGs) is combined with acoustic features (*e.g.*, pitch, tone, and/or volume per Claim 1(b)) and a speaker embedding (style embedding vector) to modify the accent (dialect) of the output audio signal. 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. *Fan* provides several examples of how to store and retrieve the style embedding vector with *Fan's* accent conversion system. See EX1006-Fan, [0024]-[0026] (“According to a third aspect, an embodiment of this application provides a terminal, including a processor, where the processor is coupled to at least one memory, and the processor is configured to read a computer program stored in the at least one memory, so that the terminal performs the method in any implementation of the first aspect. According to a fourth aspect, an embodiment of this application provides a computer-readable medium. The computer-readable storage medium stores a computer program, and when the computer program is run on a computer, the computer is enabled to perform the method in the first aspect. According to a fifth aspect, this application provides a chip system. The chip system includes a processor, configured to support a terminal device in implementing functions in the foregoing aspects. In a possible design, the chip system further includes a memory. The memory is configured to store program instructions and data that are necessary for the terminal device. The chip system may include a chip, or may include a chip and another discrete component.”); [0094] (“For example, a structure output by the multihead attention mechanism may be a 256-dimensional style embedding vector (a style feature).”); [0156]

(“The processing module **1310** is further configured to: input the third voice into the style separation model, and extract a vector of the first feature of the third voice by using the style separation model; and input the third voice into a sub-feature extraction model, and extract a vector of the sub-feature by using the sub-feature extraction model.”); [0158]-[0159] (“The processing module 1310 is further configured to determine the style feature of the third voice based on the vector of the first feature, the vector of the sub-feature received by the input module 1320, and the weight of the sub-feature received by the input module 1320. In an optional implementation, the processing module 1310 is further configured to input the vector of the first feature into a multihead attention structure, input the vector of the sub-feature and a product of the vector of the sub-feature and the weight corresponding to the sub-feature into the multihead attention structure, and output the style feature of the third voice by using the multihead attention structure.”). 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.

152. This would have been well within a PHOSITA’s skillset, who would have been familiar with incorporating speaker embeddings to a sample of input speech, as seen in prior art such as *Accentron*. A PHOSITA would have done any modifications necessary to make this modification operate since this combination

was well within a PHOSITA's skillset. A PHOSITA would have understood receiving a speaker embedding from a chunk of audio in accent conversion was common practice in the prior art. For example, *Accentron* teaches extracting a “speaker embedding” from a segment of audio. EX1008-Accentron, 11 (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65])”).



EX1008-Accentron, Fig. 1 (excerpt). *Accentron* teaches “foreign accent conversion.” EX1008-Accentron, 1.

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.”***

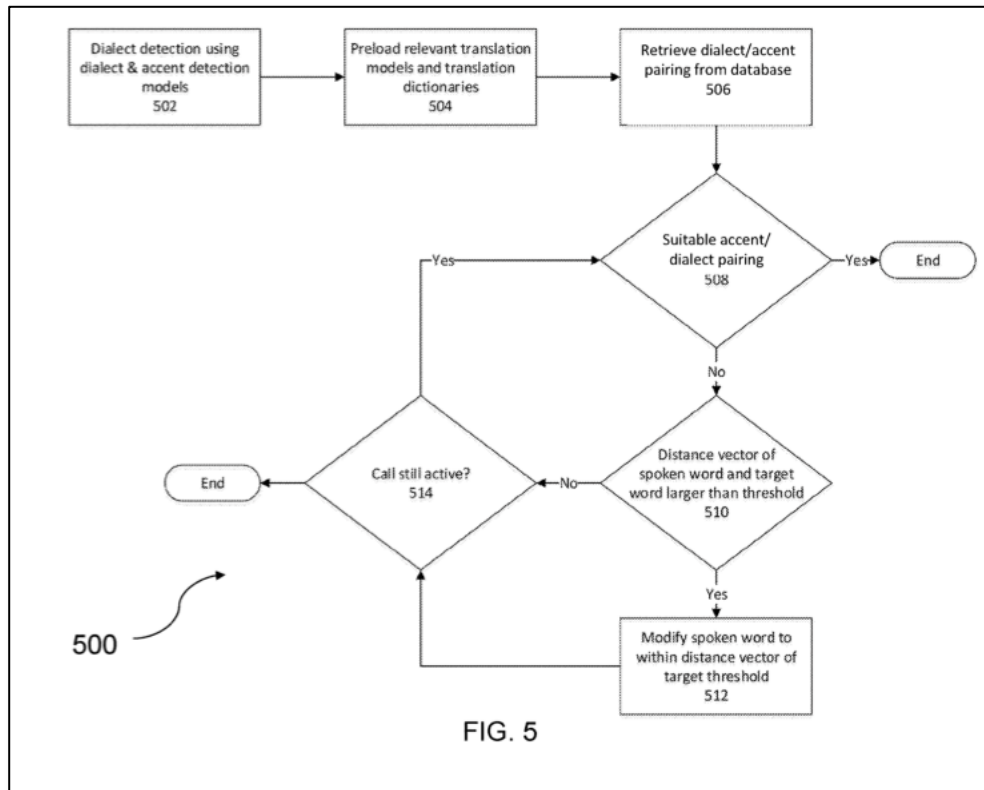
153. The '457 Patent describes output chunks, like input chunks, comprising a plurality of frames. EX1001, 11:49-53, 12:13-17. Specifically, the '457 Patent

describes output chunks as corresponding in size to that of the input chunk from which they were generated. EX1001, 11:49-53, 12:13-17.

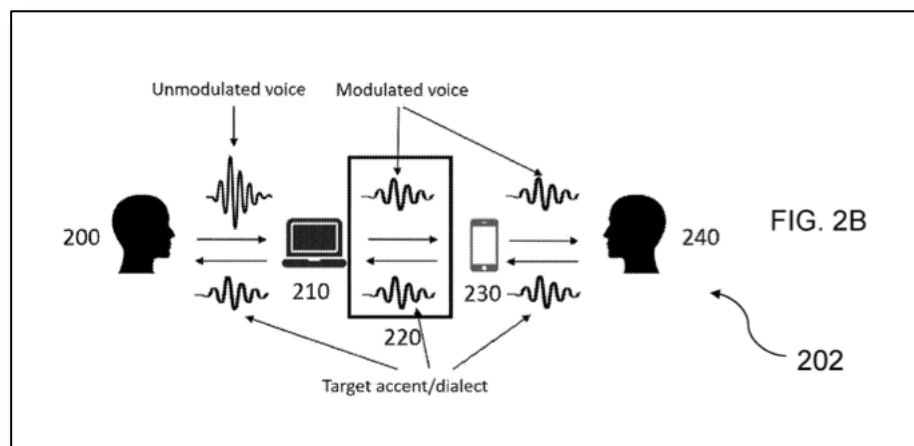
154. As discussed above, the modified *Feinauer* system combines a target accent *speaker embedding*, *acoustic features* such as pitch, tone, and/or volume per Claim 1(b), and *linguistic features* including PPGs of spoken words into an output audio signal having a modified accent. EX1004-Feinauer, [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*

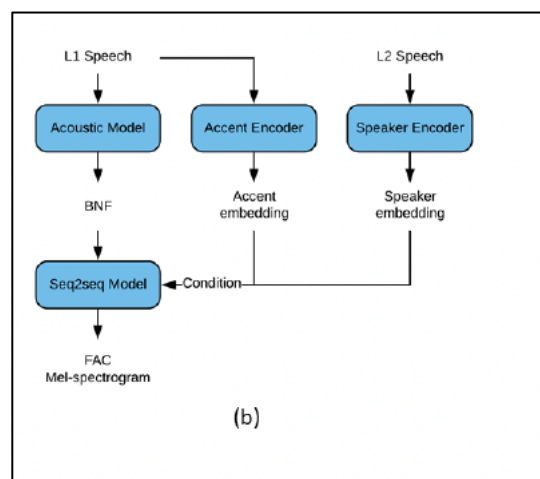
155. *Feinauer* employs an accent modifier 318 to modify the accent in “portions” of a user’s speech, modified per Claim 1(a)-1(b) to be *Prabhavalkar's* input chunks. EX1004-Feinauer, [0057], Fig. 3; See Sections IX.I.2-IX.I.3 discussing “chunk-based” processing. Specifically, *Feinauer's* method generates a conversation output audio signal with a modified accent. *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. For example, the lag duration can be a duration that sounds like real-time to a user (e.g. 100ms).” EX1004-Feinauer, [0086].



EX1004-Feinauer, Fig. 5. A PHOSITA would have understood that *Feinauer's* “broadcast” conversation teaches an output audio signal as is consistent with Fig. 2B.



EX1004-Feinauer, Fig. 2B. 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. A PHOSITA would have understood this was common practice in the prior art. For example, *Accentron* teaches “foreign accent conversion” EX1008-Accentron, 1.



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. 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. After

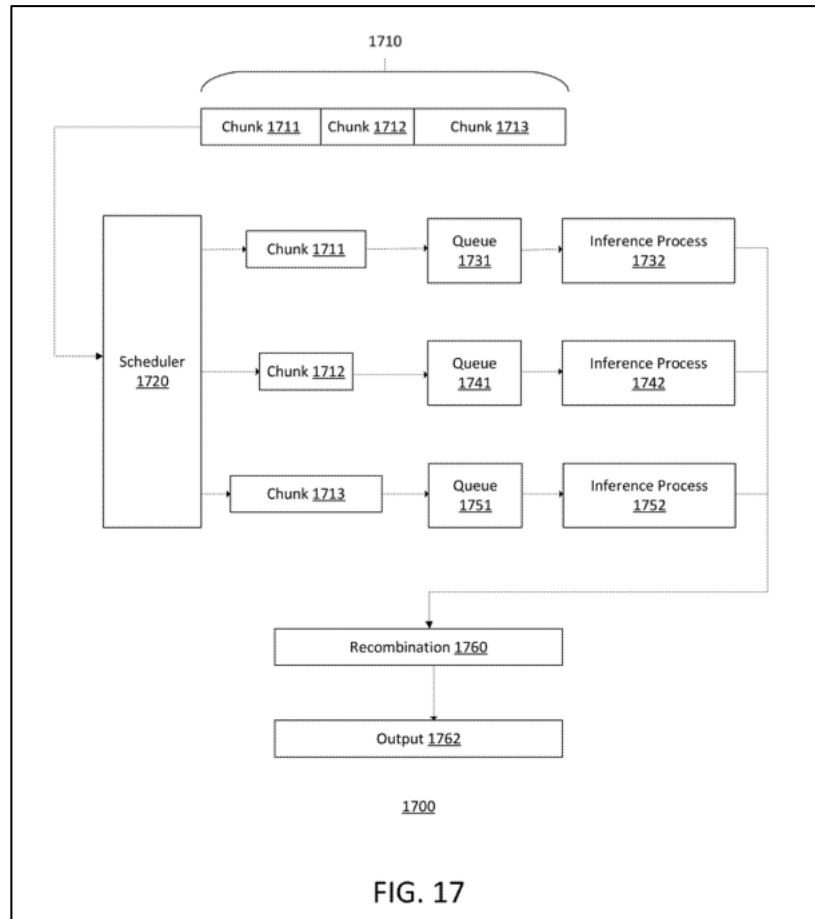
this, “the output Mel-spectrogram is converted back into a waveform through...a separately trained vocoder (e.g., WaveNet [37], WaveRNN [38]).” EX1008-Accentron, 3. *Accentron’s* “Mel-spectrogram is converted back to a waveform” using a WaveRNN vocoder. EX1008-Accentron, 9, 18.

156. *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]. *Feinauer* teaches to “modify the unmodulated voice of one or more of the participants to provide a modulated voice with a target accent and/or dialect.” EX1004-Feinauer, [0045]. *Feinauer’s* modified voice signal with “a different accent and/or dialect” is “transmitted via the communication network 220” and “received by communication device 230 which can then be communicated to, provided to or played to the participant 240.” EX1004-Feinauer, [0044]-[0046]. *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”

157. *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. *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. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-*Prabhavalkar*, 17:33-51. (emphasis added) Thus, *Prabhavalkar*’s generates output from a given chunk (an “output chunk”) before moving on to processing succeeding chunks. Based on this disclosure, a PHOSITA would have recognized or at least found obvious that *Prabhavalkar*’s “chunk-based approach,” which consumes input as chunks, correspondingly generates output as chunks. *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.” This was well known to do.

158. As an example, U.S. Patent Appl. No. 2020/0294492 to Sypniewski et al. (“*Sypniewski*”) teaches a front end module 201 splits raw audio data into a series of frames. EX1020-Sypniewski, [0059], Fig. 2. *Sypniewski* provides an example in which a 7-10 second long portion of input audio “may be segmented on relative silence such that the input audio comprises an entire utterance.” EX1020-Sypniewski, [0057], [0110]. *Sypniewski* characterizes these 7-10 second portions of audio as “chunks.” EX1020-Sypniewski, [0162]. Fig. 17 below illustrates an example of audio 1710 split into chunks 1711-1713 and then that each chunk is later recombined after the inference process. EX1020-Sypniewski, [0019] (“After inference, the resulting outputs may be recombined.”)



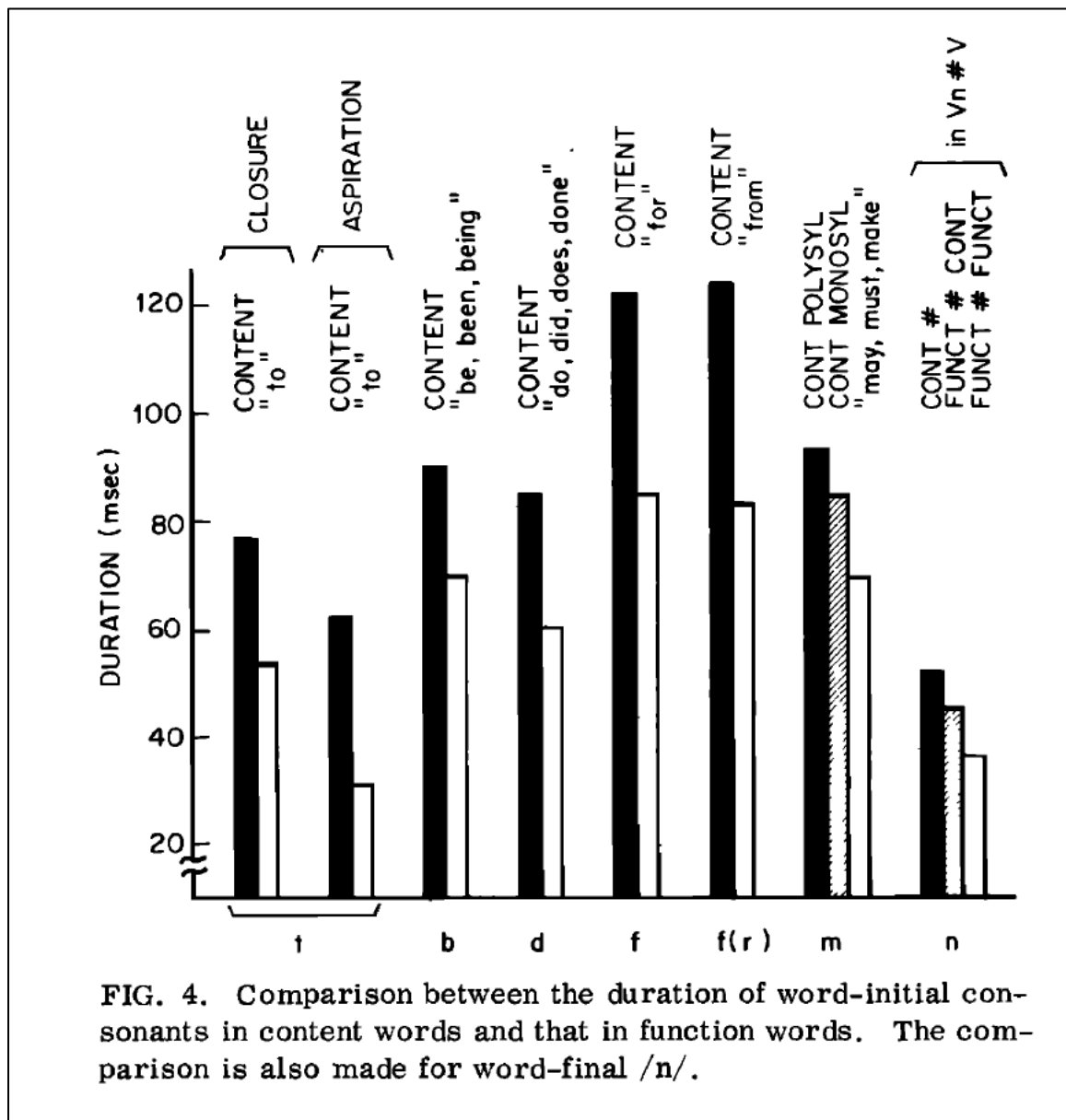
EX1020-Sypniewski, Fig. 17.

159. 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). *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. *Prabhavalkar* teaches that “streaming can

be considered to mean that the system has a maximum allowable delay of 300 ms.” EX1005-*Prabhavalkar*, 18:33-35. A PHOSITA would have appreciated that one example of “streaming” occurs with an “real-time “audio and/or video conversation” as taught by *Feinauer*. EX1004-*Feinauer*, [0041]; [0086]. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-*Prabhavalkar*, 17:33-51. (emphasis added) Therefore, a PHOSITA would have thus recognized *Prabhavalkar*’s system as performing a *real-time* speech processing method, like *Feinauer*. It was well known in the prior art that in order to perform accent conversion in real time, a conversion should be processed in “portions.” As an example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” to maintain “low real time delay” EX1018-*Wasserblat*, Abstract, [0025].

160. Further, it was well known that audio frames are short (~30ms) as taught by *Prabhavalkar* 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 EX1005-*Prabhavalkar*, 15:16-22. For

example, Umeda teaches many phonemes span multiple 30ms frames as shown in Figure 4.



EX1035-Umeda, Fig. 4. 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.

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:”***

161. *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 provide 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. EX1008-Accentron, 5 (“later work on FAC has focused on **acoustic methods**, since they only require recording speech with a **microphone**.”).

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.”***

162. *Feinauer-Prabhavalkar-Fan* 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*

163. 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 30 ms frame rate." EX1005-Prabhavalkar, 15:16-22; *see also* 20:14-20, Section IX.A.2. Therefore, there are 150ms of recorded audio data in 5 frames (as, utilizing basic mathematics, $(30 \text{ ms/frame}) * (5 \text{ frames}) = 150 \text{ ms}$). 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.

b. *Feinauer's Teachings*

164. As discussed in Claim 1(Pre), *Feinauer* teaches that the accent conversion process happens in 100ms. EX1004-Feinauer, [0086]. *See* Section IX.I.1. Specifically, *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. That is, while some range of lag duration occurs between

the time at which a user speaks (i.e., *the first timestamp*) and the time at which the audio segment has been converted (i.e., *the second timestamp*) in order to account for time required for a system to process such speech data, a PHOSITA would have understood that minimizing, or, rather, restricting it to a duration consistent with a user's perception of real-time, the lag duration rate at which the audio data is processed and, therefore, support the real-time nature of an accent conversion system.

165. A PHOSITA would have understood that the specific lag duration within this range is merely a design choice that does not require significant modification to the existing accent conversion system. As such, a 250ms lag duration, for example, falls within the aforementioned range and represents an example of a real-time processing speed. Moreover, a PHOSITA would have understood that minimizing the lag duration such that the duration corresponds with real-time yields an improved production rate of modified speech and would have been motivated to reduce the lag duration below the upper limitation of this range. As such, a PHOSITA would have understood a 250ms lag duration specifically to be effective for systems and methods for minimizing the processing time corresponding with real-time production of modified audio data. Thus, *Feinauer-Prabhavalkar-Fan's* 250 millisecond lag duration is *between 40 milliseconds and 300 milliseconds*, teaching Claim 2(b).

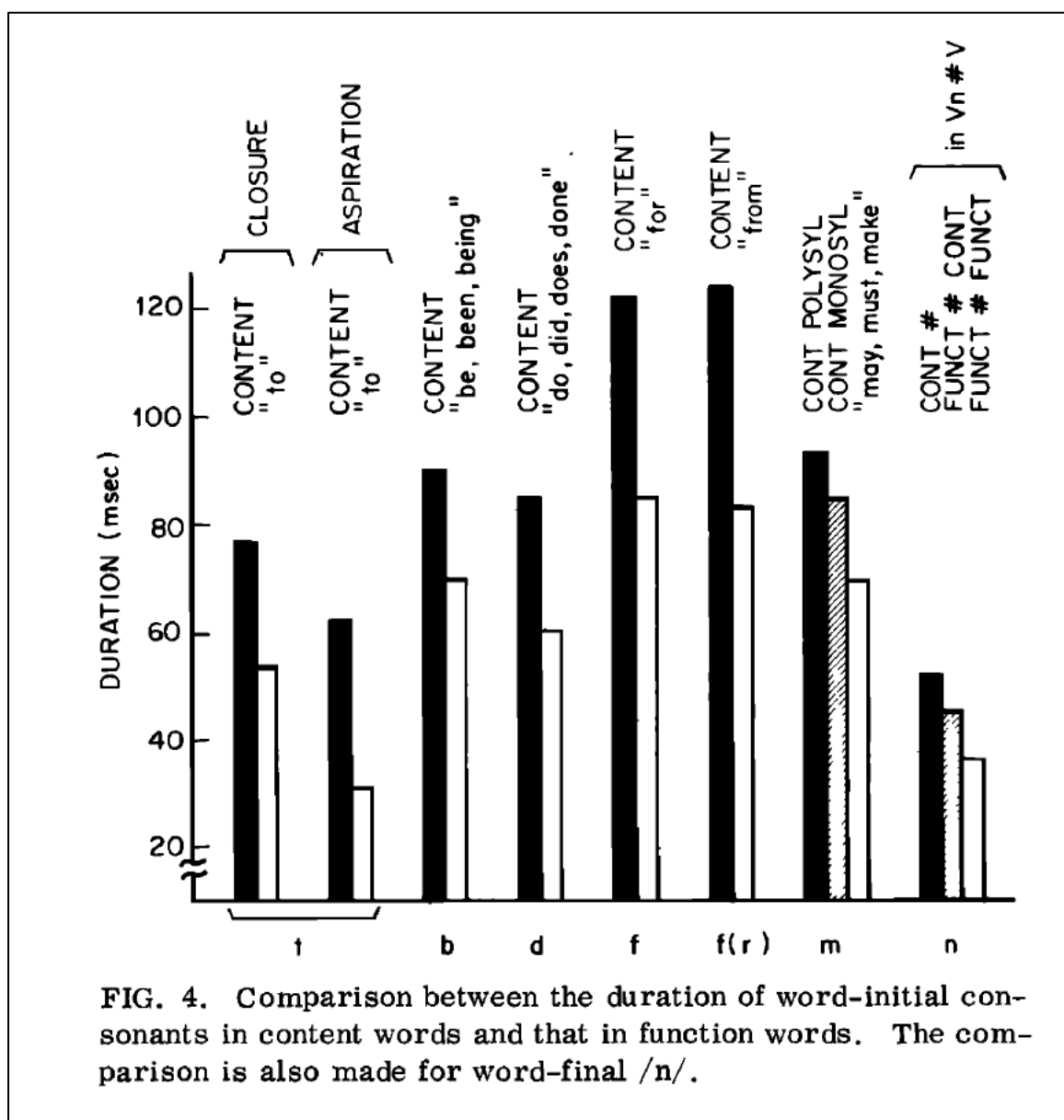
a) Motivation to Combine Prabhavalkar

166. As discussed in Claim 1(a), a PHOSITA would have been motivated to modify Feinauer to teach chunks of an audio signal. *See* Section IX.I.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. EX1005-Prabhavalkar, 15:16-22; *see also* 20:14-20. This was well known in the art. For example, *Schmandt* teaches “extract[ing] frames of Linear Predictive Coding (LPC) coefficients every 20 milliseconds” for speech recognition processes. EX1021-Schmandt, 134. The LPC coefficients within such frame size are a “concise representation of the articulatory characteristics of speech as they capture both the resonances of the vocal tract as well as its voicing characteristics. The 20 millisecond sampling interval results in 50 LPC frames per second, providing significant data reduction...[.]” EX1021-Schmandt, 134.

167. 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. For instance, *Ding* utilizes longer frame lengths, teaching a down-sampling of the waveforms of each utterance “from 48kHz to 16kHz to match the sampling rate of other modules.” EX1012-*Ding*, 778rhc. *Ding* then discloses the

extraction of “an 80-dim Mel-spectrogram with a 50ms window and 12.5ms shift” and, following the frame shift, extraction of the “PPG (collapsed into a 40-dim mono-phone PPG from the 5,816-dim senone PPG) and the d-vector (256-dim) for each utterance using the acoustic model and speaker recognition model, respectively.” EX1012-Ding, 778rhc.

168. 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. EX1005-Prabhavalkar, 15:16-22; *see also* 20:14-20. It was well known that audio frames are short (~30ms) as taught by *Prabhavalkar* 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. For example, *Umeda* teaches many phonemes span multiple 30ms frames as shown in Figure 4.



EX1035-Umeda, Fig. 4. 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.

169. A PHOSITA would have been motivated to look towards *Prabhavalkar's* teachings since it was well known that frames (alternatively referred

to as windows or segments) of an audio signal correspond to a time. Particularly for real time applications, a PHOSITA would have understood the corresponding time to be adequately expressed in a magnitude of units consistent with milliseconds reflective of the required rate modification inherent to such process. For example, *Ding* teaches resampling “audio waveforms hav[ing] a 44.1 kHz sampling rate...to 16 kHz” and “extract[ing] 80-dimensional Mel-spectrogram with a 25ms window and 10ms shift”. EX1012-Ding, 18. 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. 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.

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.”

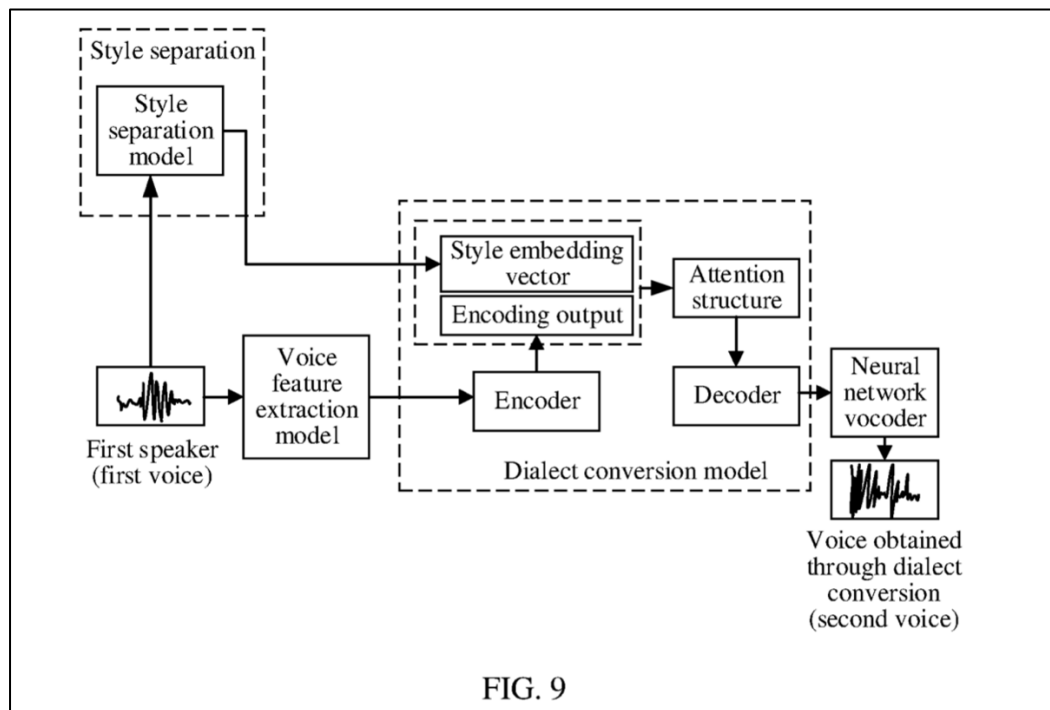
170. *Feinauer-Prabhavalkar-Fan* teaches the speaker embedding (*Fan’s* style embedding vector) is generated based on the speech audio signal.

171. *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]. 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, [0127]; Fig. 9. Specifically, *Fan* teaches that a style separation feature separates a style of the first voice to obtain a timbre feature vector of the first voice. EX1006-Fan, [0127] (“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. Content of the second voice is the same as content of the input voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”). In this case, *Fan*’s system and method are being used for both accent conversion and style enhancement. *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. See ¶149.

172. *Fan* teaches this style feature includes “but is not limited to a timbre. The timbre refers to quality of a voice, which is referred to as tone quality. The timbre can reflect unique quality of a sound made by a speaker. The timbre includes

but is not limited to a rhythm, an accent, and a speaking speed.” EX1006-Fan, [0050].



EX1006-Fan, Fig. 9.

173. *Fan* envisions that a user’s voice’s timbre may be represented as a “style embedding **vector**.” EX1006-Fan, [0094], Figs. 9, 12. A PHOSITA would have recognized this “style embedding vector” teaches a speaker embedding vector consistent with the ’457 specification. It was well known in the prior art that a “speaker embedding vector” at least includes “timbre” as taught by Fan. EX1008-Accentron, 21 (“speaker embedding mainly encodes information such as voice quality/timbre and pitch”). Thus, in *Fan*, a *speaker embedding* vector at least

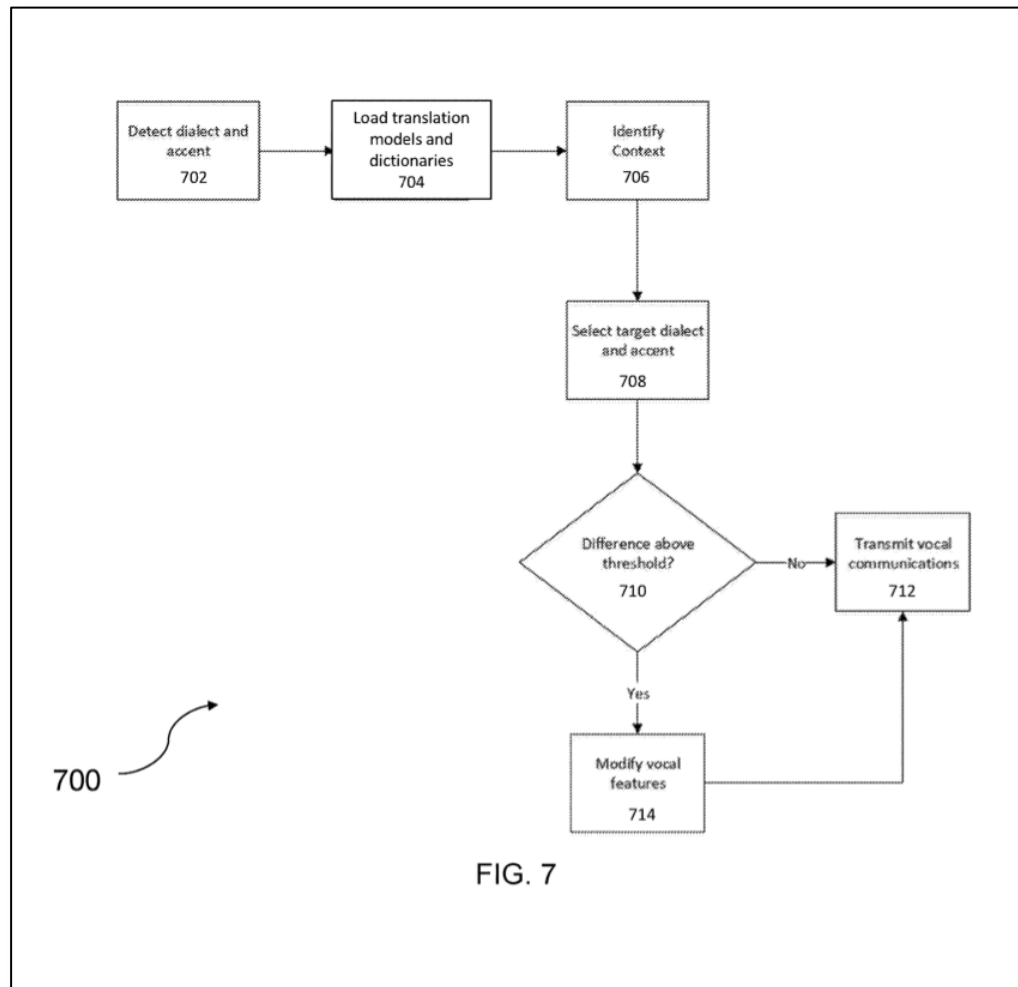
reflecting the timbre of a human speaker is received by the processor of the device for modifying the accent of the input speech.

a. *Motivation to Combine Fan's speaker embedding*

174. 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” voice profile 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. A PHOSITA would have been motivated to use the same software and hardware needed to make the combination in Claim 1 in Claim 3. *See* Section IX.I; EX1004-Feinauer, [0038]; [0081], [0105], Fig. 7; EX1006-Fan, [0010], [0081]-[0082], [0094], [0160]-[0162], Figs. 12-13.

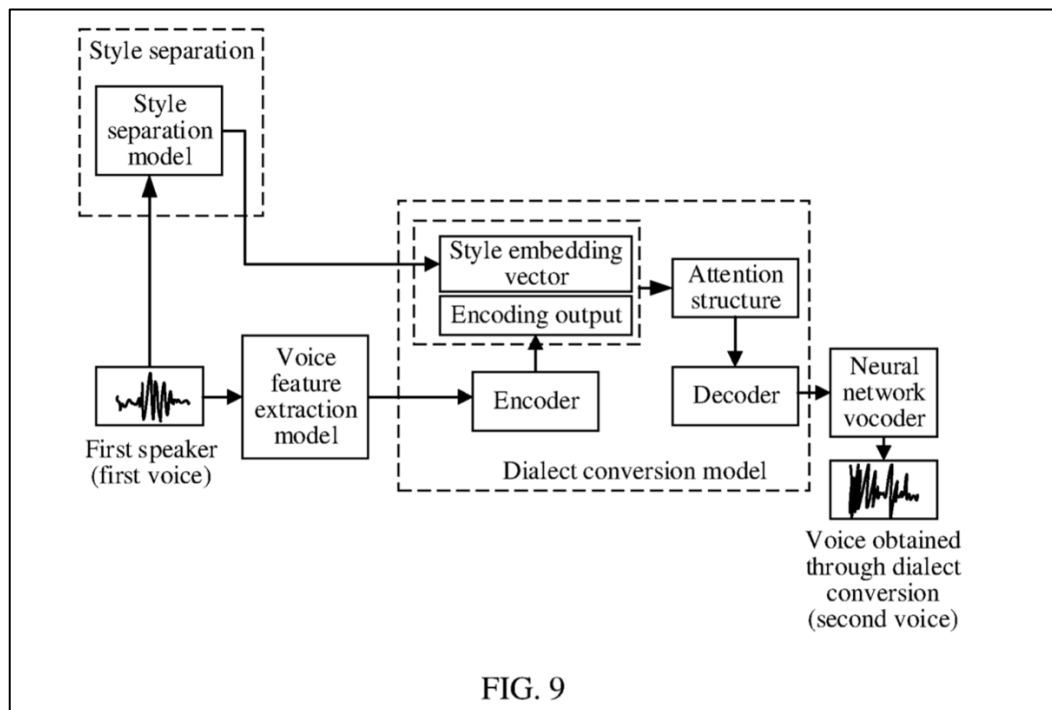
175. *Feinauer* teaches adjusting the accent of the user to match a target accent. EX1004-Feinauer, [0045]-[0046], [0096], [0103]-[0110], Figs. 2B, 7. *Feinauer's* system loads a relevant “translation model” corresponding to the particular target accent. EX1004-Feinauer, [0081] (“step 504, the accent and dialect modifier preloads relevant translation models and translation dictionaries. For example, the relevant translation models and translation dictionaries can be installed on a local machine, such as computing device 210, 230”), [0105] (“step 704, the

accent and dialect modifier loads translation models and dictionaries. Loading translation models and dictionaries”), Fig. 7.



EX1004-Feinauer, Fig. 7. As discussed above, *Fan* teaches that a style separation feature separates a style of the first voice to obtain a timbre feature vector of the first voice. EX1006-Fan, [0050], [0127]. A PHOSITA 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. A PHOSITA would have appreciated

that *Fan* already teaches generating an output signal with a source user speaker embedding as shown in Figure 9.



EX1006-Fan, Fig. 9. 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. 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.

2. ***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."***

176. As discussed regarding Claim 1(d), Feinauer's system loads a relevant "translation model" speaker embedding corresponding to the particular target accent. See Section IX.I.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].

3. ***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."***

177. See Claim 1(b).

D. 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:"***

178. 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 IX.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;"***

179. 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;"***

180. 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;”*

181. See Claim 1(c).

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

182. 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.”*

183. See Claim 1(e).

E. 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”*

184. 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.”*

185. See Claim 2(b).

F. Claim 16

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

186. See Claim 3.

G. 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.”***

187. See Claim 4.

H. 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:”***

188. 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 IX.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;”***

189. 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;”***

190. 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;”***

191. See Claim 1(c).

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

192. 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.”*

193. See Claim 1(e).

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

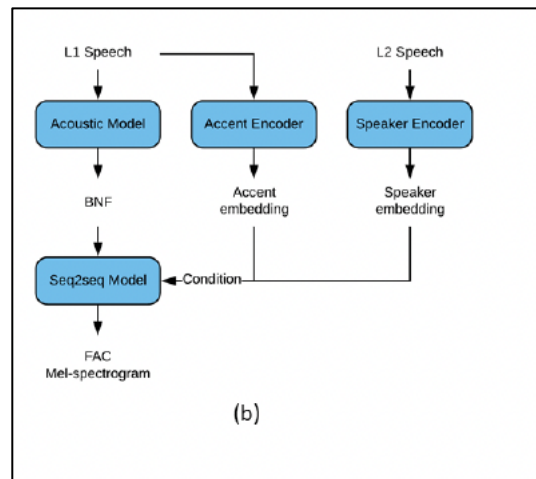
A. Claim 9

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.”*

194. 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 IX.I.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* and *Fan* teaches such.

a. Accentron’s Teachings

195. *Accentron* teaches a seq2seq *synthesis module*. EX1008-Accentron, 7 (“a seq2seq model that consumes the previous three representations to synthesize Mel-spectrogram for an arbitrary L2 speaker.”).



EX1008-Accentron, Fig. 1. (excerpt). *Accentron*’s seq2seq model consumes bottleneck features, accent embedding, and speaker embedding “to synthesize Mel-spectrograms 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, therefore in the combination with Accentron a PHOSITA would have appreciated that the input PPGs (in place of BNFs) would have a *reduced accent*. As an example, *Ding* teaches most methods “implicitly assume that PPGs are speaker-independent and contain only linguistic information in an utterance. In practice, however, PPGs carry speaker individuality cues, such

as accent, intonation, and speaking rate. As a result, these cues can leak into the voice conversion, making it sound similar to the source speaker.” EX1012-Ding, Abstract. That is, extracting linguistic features using PPGs reduces the amount of accent present, but some accent remains embedded in the PPGs. Therefore, it was well known that PPGs will still retain some accent within the PPG, though at a reduced amount. *See* Section IX.I.4.

196. *Accentron’s* seq2seq model is a *synthesis module* which includes a decoder. EX1008-Accentron, 7 (“a seq2seq model that consumes the previous three representations to synthesize Mel-spectrogram for an arbitrary L2 speaker.”); Table 1.

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). The decoder produces the predicted FAC (Foreign Accent Conversion) Mel-spectrogram based on the hidden representations from the bottleneck features, accent embedding, and speaker embedding (the output from the previous decoding step) and “a context vector produced by the attention mechanism.” EX1008-Accentron, 3.

197. During accent conversion the “predicted Mel-spectrogram is passed through a post-net consisting of five convolution layers to predict the residual. Each of these layers has 512 kernels with 5x1 shape and 1x1 stride, followed by tanh

activation and batch normalization. The residual is added back to the original prediction to form the final prediction:

$$\hat{y}^t = \hat{y}_{\text{pre}}^t + \text{PostNet}(\hat{y}_{\text{pre}}^t)$$

EX1008-Accentron, 15.

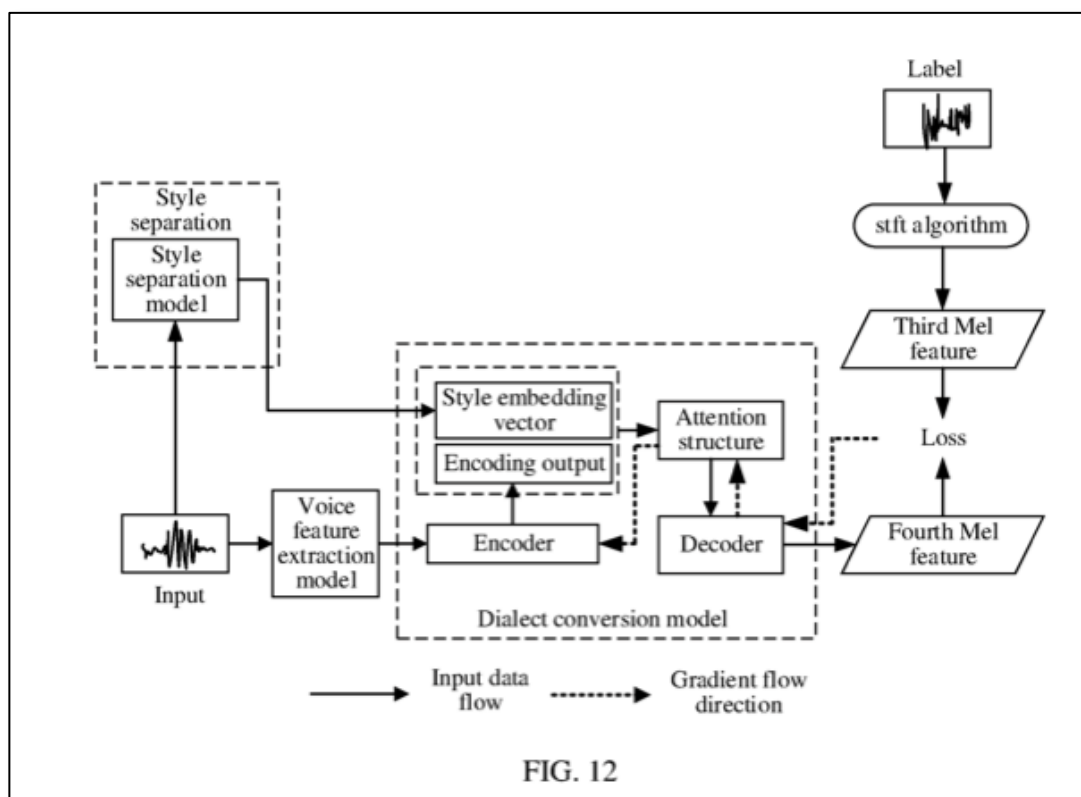
198. The decoder “produces an 80-dimensional Mel-spectrogram as the prediction of the L2 speech” with a “25ms window and 10ms shift.” EX1008-Accentron, 14, 18. After this, “the output Mel-spectrogram is converted back into a waveform through...a separately trained vocoder (e.g., WaveNet [37], WaveRNN [38]).” EX1008-Accentron, 3. *Accentron*’s “Mel-spectrogram is converted back to a waveform” using a WaveRNN vocoder. EX1008-Accentron, 9; 18. Therefore, *Accentron* teaches *providing the melspectrogram to a vocoder*.

b. *Motivation to Combine the modified Feinauer with Accentron and Fan*

199. 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. *Accentron* teaches the hidden representations from the bottle neck features that include a reduced accent, accent embedding, and speaker embedding is fed “to the

decoder to produce the predicted FAC Mel-spectrogram.” EX1008-Accentron, 12-13; see also EX1008-Accentron, 10. 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).

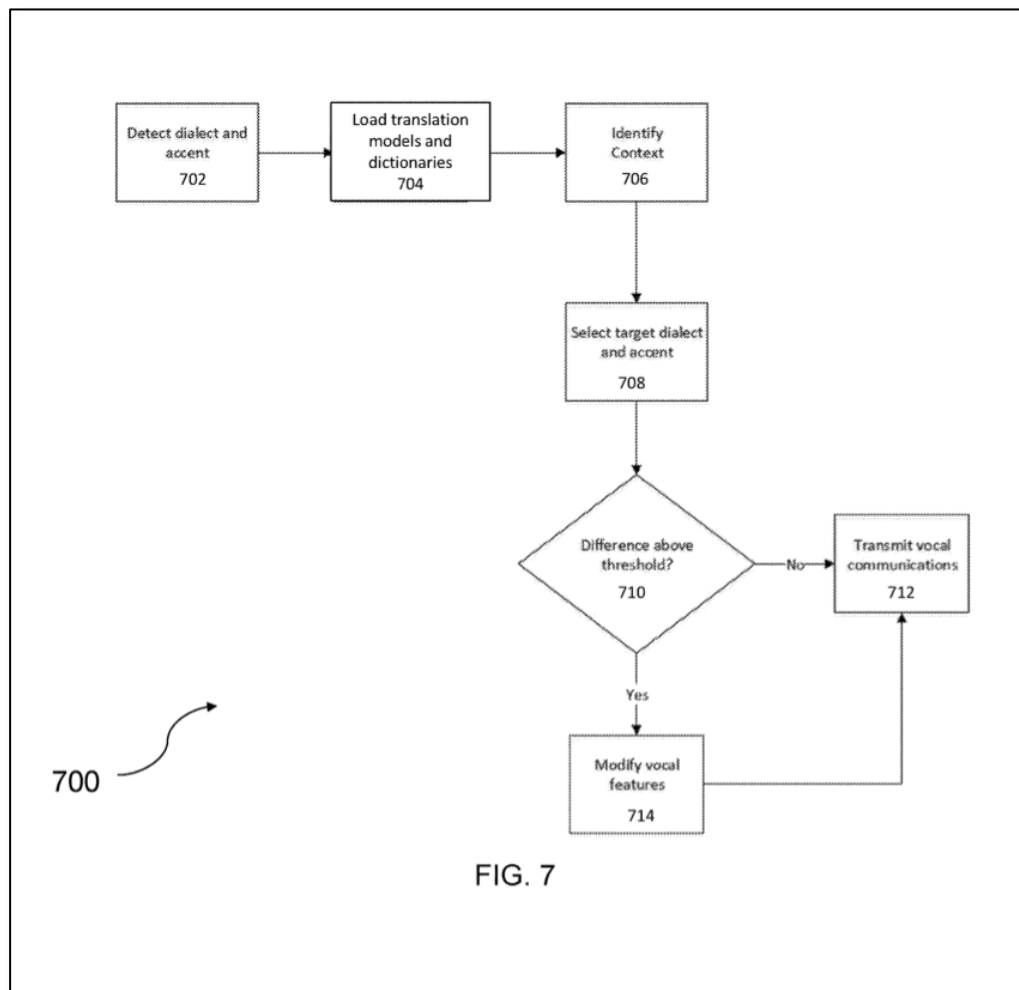
200. A PHOSITA would have appreciated that *Fan*, already in the combination teaches using mel features when generating an output. EX1006-Fan, [0099] (“The vocoder may be a neural network vocoder, and the network is responsible for converting the mel feature into a voice signal with a high natural degree.”).



EX1006-Fan, Fig. 12. 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] (“the dialect detection model 332 can output an accent or dialect probability, for example, by matching the biometric characteristics of different accents and/or dialect on a spectrogram.”); [0080] (“the dialect detection model 332 can output an accent or dialect probability, for example, by matching the biometric characteristics of different accents and/or dialect on a spectrogram.”); [0095] (“using preprocessing tools to convert the audio sample waveforms into a spectrogram”); and [0099] (“using preprocessing tools to convert the audio sample waveforms into a spectrogram of MFCC.”) 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. *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].

201. 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

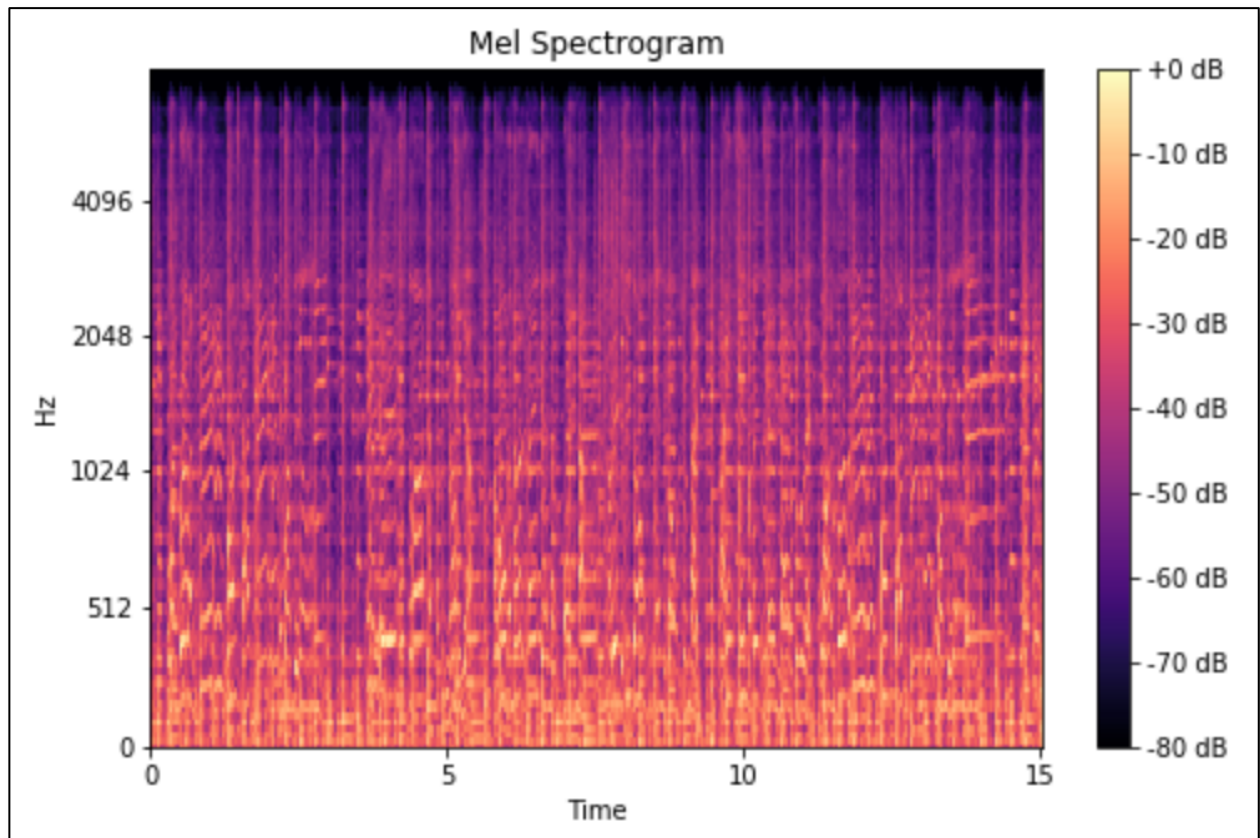
modification. EX1004-Feinauer, [0001] (“accent and dialect modification”), Fig. 7 (step 714); [0110] (“modifying vocal features 714 can include modifying the dialect and/or accent of at least a portion of the voice communication.”).



EX1004-Feinauer, Fig. 7.

202. A PHOSITA would have expressly implemented the 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 IX.I.3-IX.I.4. A PHOSITA would have been motivated to make the combination since it

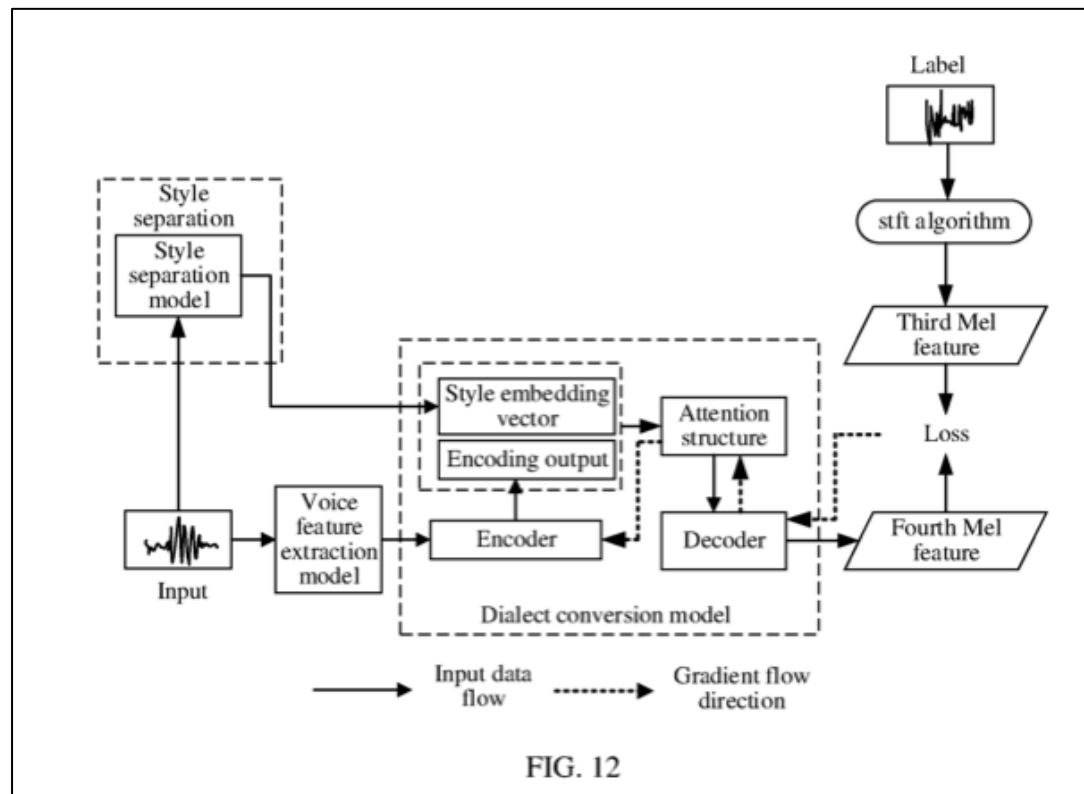
was well known mel-spectrograms provide a convenient representation of audio for the vocoder to reconstruct and that neural networks are commonly used in accent conversion systems and implementing a neural network in Feinauer's accent modifier 318 which would have increased efficiency. Studies "have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies." EX1034-Roberts, 11. In order to combat this, in "1937, Stevens, Volkmann, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the mel scale." EX1034-Roberts, 12. From a spectrogram, one may "perform a mathematical operation on frequencies to convert them to the mel scale." EX1034-Roberts, 12. A Mel-spectrogram is "is a spectrogram where the frequencies are converted to the mel scale." EX1034-Roberts, 13.



EX1034-Roberts, 8. A PHOSITA would have appreciated that a spectrogram “requires much less storage than the original waveform and reveals acoustic features specific to speech,” motivating the combination to consume less storage in speech processing. EX1021-Schmandt, 13.

203. 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. As one example, *Fan* teaches “the attention structure cooperates with

the encoder to output a mel (or may also be referred to as a mel spectrum) feature of the second voice.” EX1006-Fan, [0097].



EX1006-Fan, Fig. 12.

204. *Fan's* “mel spectrum feature is an intermediate representation. Although a mel spectrum contains information about an output voice, the mel spectrum cannot be directly played, and an inverse process of converting a voice signal into the mel spectrum needs to be performed. In this application, a playable audio file may be generated by using a vocoder by using an acoustic feature of the mel spectrum. The vocoder may be a neural network vocoder, and the network is responsible for converting the mel feature into a voice signal with a high natural

degree.” EX1006-Fan, [0099]. *See also* EX1006-Fan, [0047]-[0049], [0053]. 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. 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.

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;”***

205. *Feinauer-Prabhavalkar-Fan* in view of *Accentron* teaches per *Accentron*, 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*

206. *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 (optimize) 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).

207. During accent conversion and “to incorporate the spectral residual and improve synthesis quality,” *Accentron*’s “predicted Mel-spectrogram is passed through a post-net consisting of five convolution layers to predict the residual. Each of these layers has 512 kernels with 5x1 shape and 1x1 stride, followed by tanh activation and batch normalization. The residual is added back to the original prediction to form the final prediction:

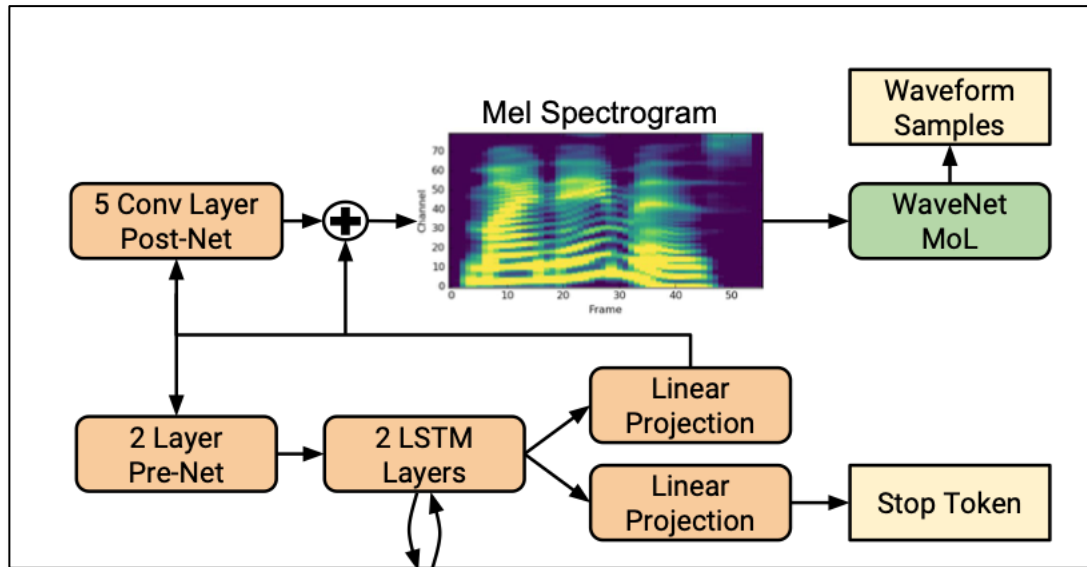
$$\hat{y}^t = \hat{y}_{\text{pre}}^t + \text{PostNet}(\hat{y}_{\text{pre}}^t) \text{” EX1008-Accentron, 15.}$$

208. After this step, 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

209. 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 X.Q.1.

210. 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*. A PHOSITA would have appreciated how common in the prior art this was. As an example, *Shen* teaches the “encoder converts a character sequence into a hidden feature representation which the decoder consumes to predict a spectrogram...Finally, the predicted mel spectrogram is passed through a 5-layer convolutional post-net which predicts a residual to add to the prediction to improve the overall reconstruction....We minimize the summed mean squared error (MSE) from before and after the post-net to aid convergence.” EX1036-Shen, 2.



EX1036-Shen, Fig. 1 (excerpt). A PHOSITA would have been motivated to train the modified *Feinauer* system to operate in the manner taught by *Accentron* such that the encoder, decoder, and post-net module would have operated in the same manner as *Accentron*.

211. 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, [0001] (“accent and dialect modification”), Fig. 7 (step 714); [0110] (“modifying vocal features 714 can include modifying the dialect and/or accent of at least a portion of the voice communication.”).

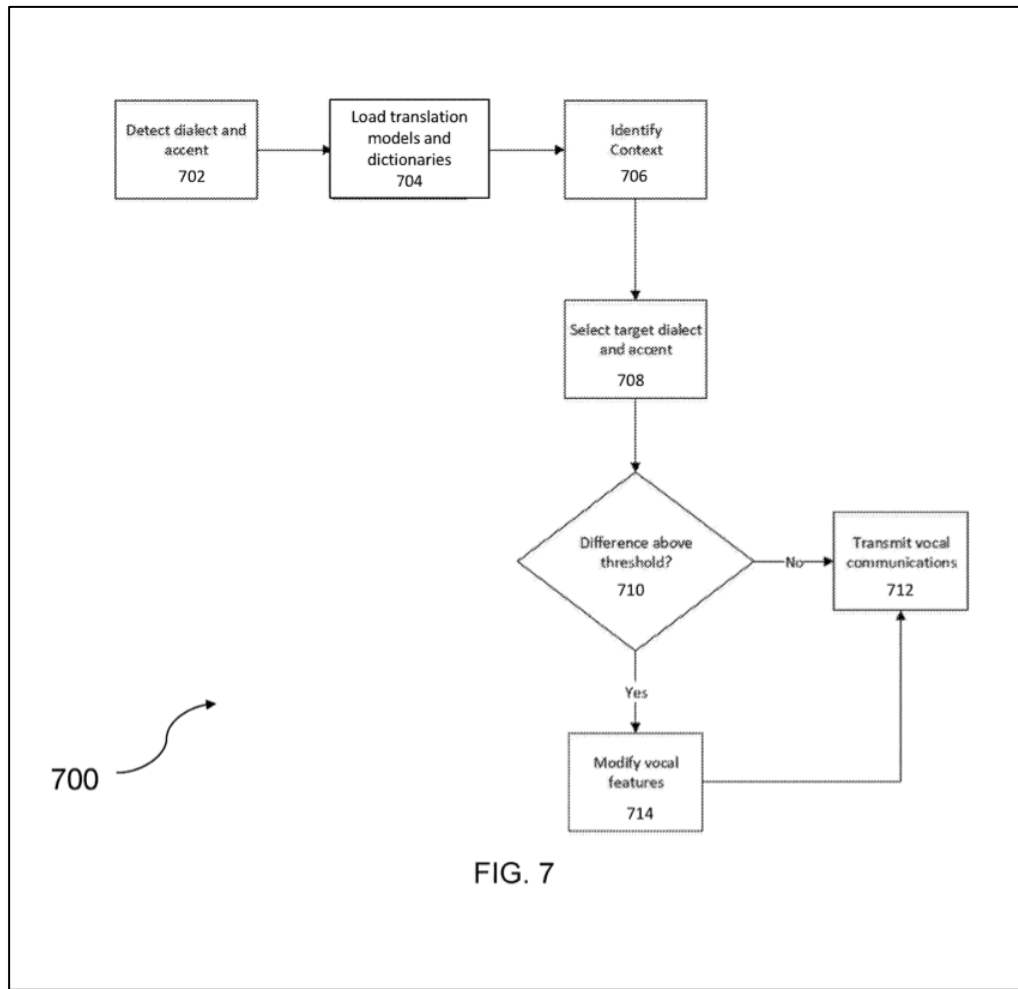


FIG. 7

EX1004-Feinauer, Fig. 7.

212. 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; (“model is optimized”). A PHOSITA would have appreciated that this step would have ensured that the generated audio representation accurately matched the target

mel spectrogram. For example, U.S. Patent Appl. No. 2022/0246132 to Zhang et al. (“*Zhang*”) teaches “determining a Euclidean distance between the corresponding quantized embedding and the latent feature extracted from the corresponding portion of the reference spectrogram aligned with the speech embedding...Here, assigning the quantized embedding to the latent feature includes assigning the selected quantized embedding associated with the minimum Euclidean distance.” EX1037-Zhang, [0007].

213. A PHOSITA would have appreciated that one way this modification would have improved the overall accent conversion accuracy is by improving the “overall reconstruction” of the signal was well known in the prior art when using a post-net. As an example, *Shen* teaches “the predicted mel spectrogram is passed through a 5-layer convolutional post-net which predicts a residual to add to the prediction to improve the overall reconstruction...We minimize the summed mean squared error (MSE) from before and after the post-net to aid convergence.” EX1036-Shen, 2. As another example *Zhao* teaches an “advantage of the PostNet is that it can see the entire decoded sequence. Therefore, the PostNet can use both past and future information to correct the prediction error for each individual frame [49].” EX1045-Zhao, 2371hc.

214. As discussed in Claim 9, a PHOSITA would have been motivated to implement Feinauer’s “accent modifier 318” to be a neural network. A PHOSITA

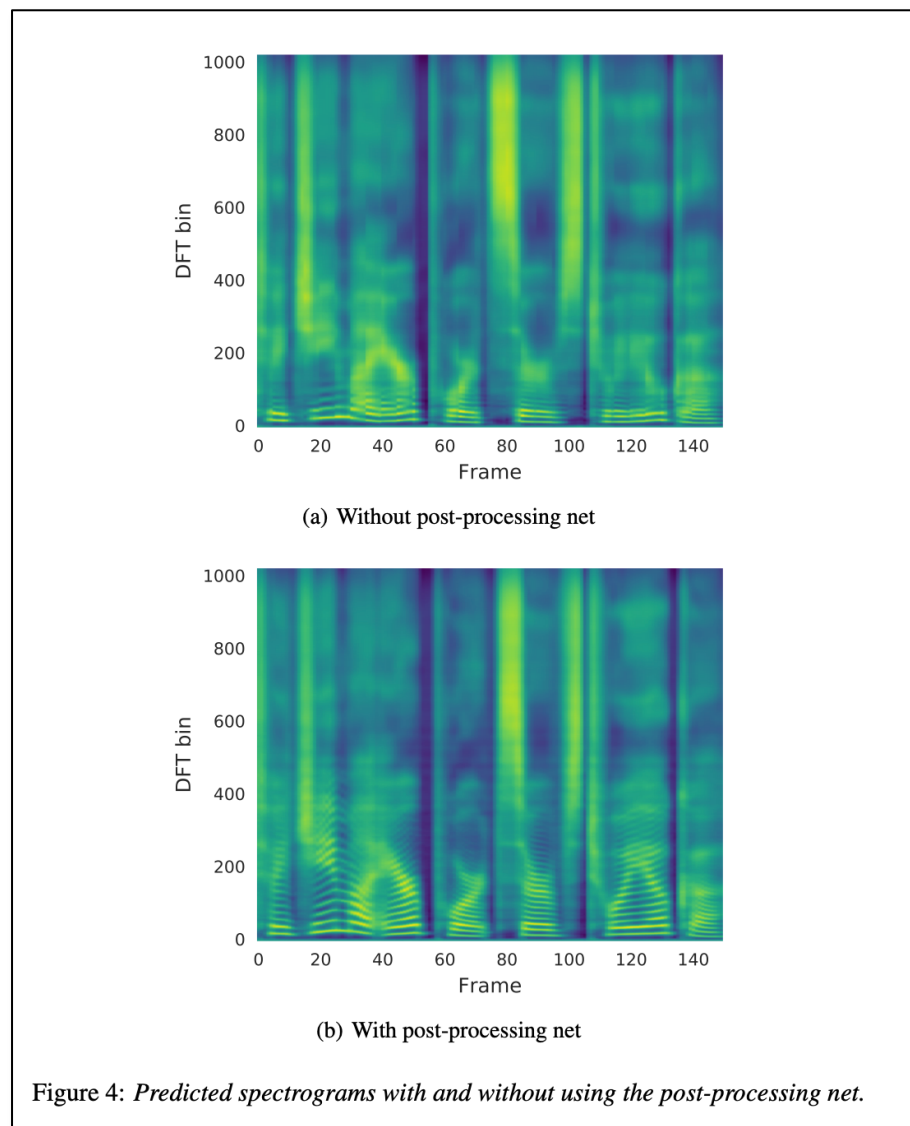
would have been motivated to make the combination since it was well known to convert an accent to include an encoder, decoder, and a post-net module as demonstrated by Accentron.

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

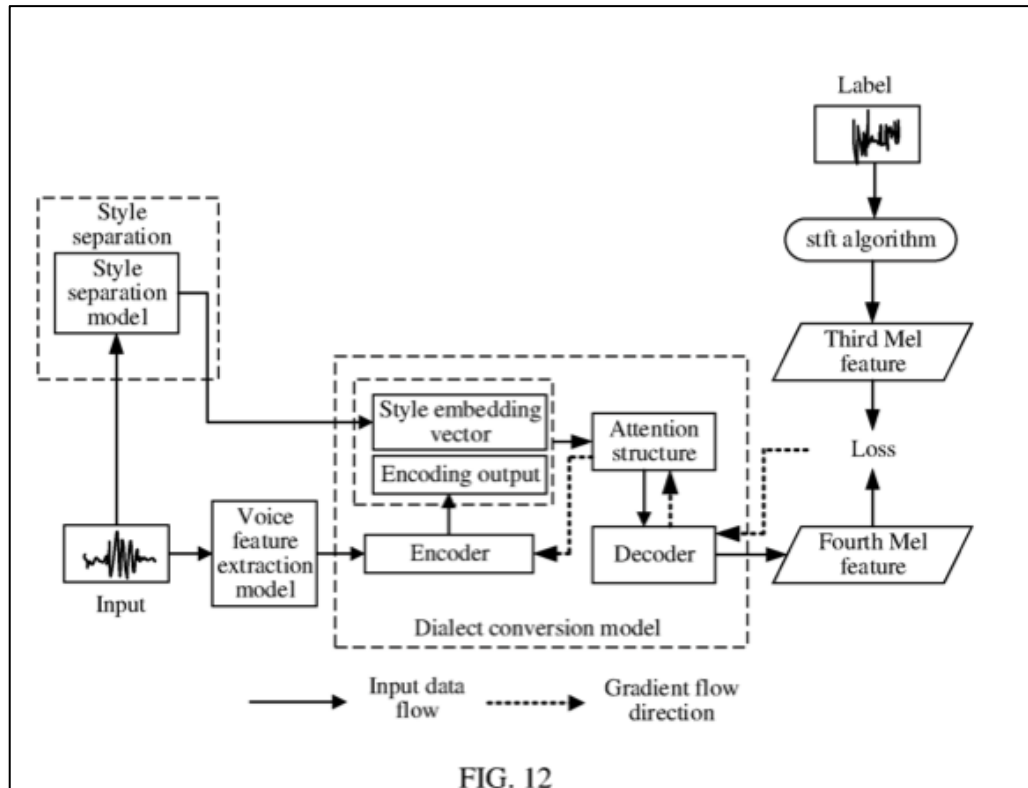
EX1008-Accentron, Table 1 (annotated). As another example *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches one example of “an encoder-decoder architecture.” EX1045-Zhao, 2370. This encoder-decoder architecture uses an encoder to “consume” input sequences and generate a high-level hidden representation sequence. EX1045-Zhao, 2371. Then, *Zhao* passes “the hidden linguistic representation sequence to the decoder, which consists of a location-sensitive attention mechanism [47] and a decoder LSTM, to predict the raw

Mel- spectrogram.” EX1045-Zhao, 2370. *Zhao* also teaches a PostNet (“PostNet: Five stacked 1-D convolutional layers”) EX1045-Zhao, 2370. As an example, *TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS* to Wang et al. (“*Wang*”) teaches it was well known that using post-net processing improves the predicted spectrogram such as by resulting in “better resolved harmonics...and high frequency formant structure, which reduces synthesis artifacts” EX1038-Wang, 7.



EX1038-Wang, Fig. 4.

215. 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. This was very common in the prior art, for example *Fan* also teaches a sequence-to-sequence neural network that includes an encoder-decoder structure. EX1006-Fan, [0047] (“The seq2seq neural network means that both the encoder and the decoder include a neural network. The encoder (encoder) is configured to encode the input sequence according to a specific rule to generate the vector. The decoder (decoder) is configured to convert the generated vector into the output sequence.”). An example of this setup in *Fan's* accent conversion system is depicted in Figure 12 below. *See also*, EX1006-Fan, Abstract.



EX1006-Fan, 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.

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."***

216. 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.

217. As discussed in Claims 1(b)-(c), *Feinauer-Prabhavalkar-Fan* teach linguistic features with the reduced accent, acoustic features, and speaker embeddings. See Sections IX.I.3, IX.I.4. *Accentron* further teaches *generating hidden features to generated further features then processing the further features by the decoder and the post-net module to generate the melspectrogram.*

218. The linguistic features (BNF) is input into the encoder, where the “encoder converts a BNF sequence to a hidden representation sequence.” EX1008-Accentron, 14, 8 (“BNFs which we assume capture only the linguistic content”). 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.* As an example, *Ding* teaches most methods “implicitly assume that PPGs are speaker-independent and contain only linguistic information in an utterance. In practice, however, PPGs carry speaker individuality cues, such as accent, intonation, and speaking rate. As a result, these cues can leak into the voice conversion, making it sound similar to the source

speaker.” EX1012-Ding, Abstract. Therefore, it was well known that PPGs will still retain some accent within the PPG, though at a reduced amount. *See* Section IX.I.4.

219. 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*.

220. 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*.

221. 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

222. 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 X.Q.1. 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, [0001] (“accent and dialect modification”), Fig. 7 (step 714); [0110] (“modifying vocal features 714 can include modifying the dialect and/or accent of at least a portion of the voice communication.”)

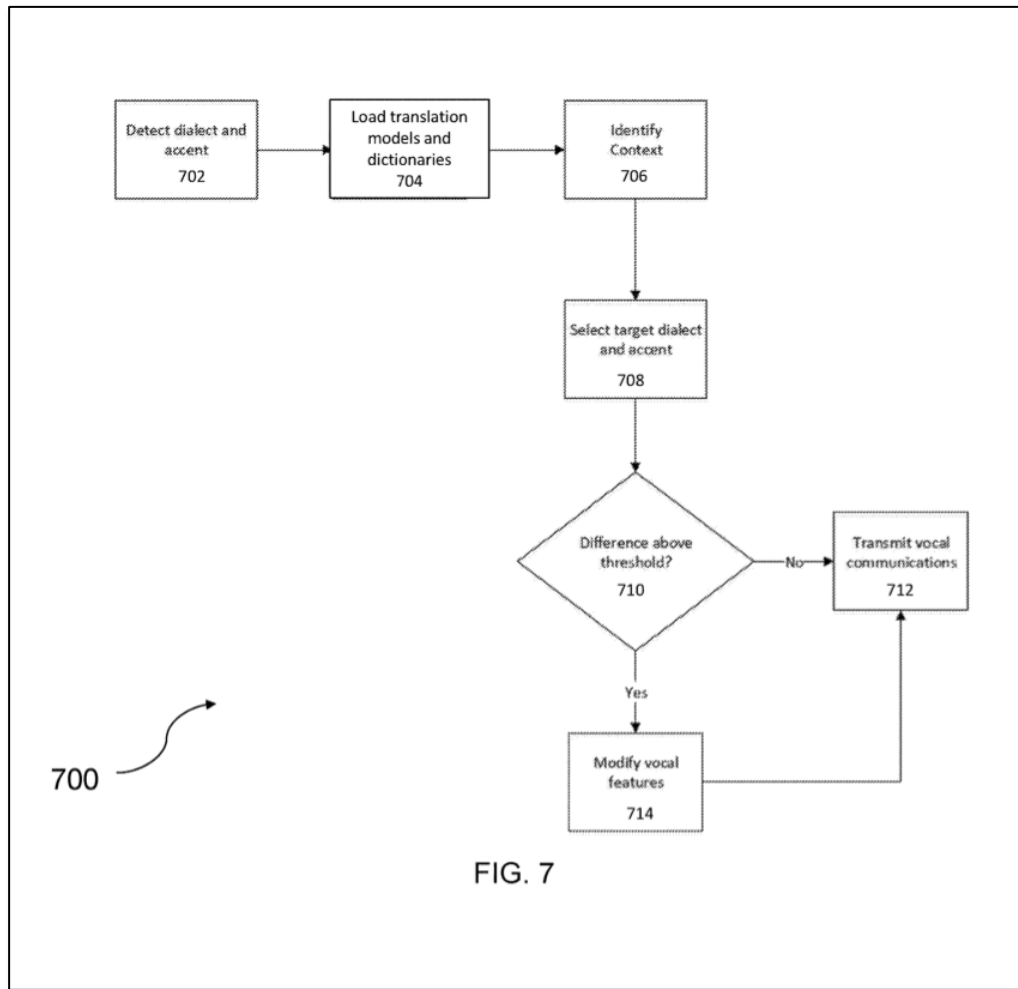


FIG. 7

EX1004-Feinauer, Fig. 7. In the proposed modified method, a PHOSITA would have been motivated and found it obvious to modify *Feinauer's* accent modifier 318 to processes the linguistic features with the reduced accent by the encoder to generate hidden features as taught by *Accentron*. It was well known in the prior art that neural networks generate hidden features. EX1045-Zhao, 3370. (“Conceptually, an encoder-decoder architecture uses an encoder (usually a recurrent neural network; RNN) to ‘consume’ input sequences and generate a high-level hidden representation sequence. Then, a decoder (an RNN with an attention mechanism) processes the

hidden representation sequence.”) A PHOSITA would have appreciated that by generating hidden features, this would improve the overall accent conversion processing time in *Feinauer* since the hidden features allow automation of feature modeling via the neural network. EX1045-Zhao, 3370. (“Conceptually, an encoder-decoder architecture uses an encoder (usually a recurrent neural network; RNN) to ‘consume’ input sequences and generate a high-level hidden representation sequence. Then, a decoder (an RNN with an attention mechanism) processes the hidden representation sequence.”) As discussed in Claim 9, a PHOSITA would have been motivated to implement Feinauer’s “accent modifier 318” to be a neural network. A PHOSITA would have been motivated to modify *Feinauer*’s accent modifier 318 to combine the hidden features, acoustic features, and speaker embeddings to generate further features as taught by *Accentron*, improving the overall accent conversion accuracy in *Feinauer*. As another example *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches one example of “an encoder-decoder architecture.” EX1045-Zhao, 2370. This encoder-decoder architecture uses an encoder to “consume” input sequences and generate a high-level hidden representation sequence. EX1045-Zhao, 2370. Then, *Zhao* passes “the hidden linguistic representation sequence to the decoder, which consists of a location-sensitive attention mechanism [47] and a decoder LSTM, to predict the raw

Mel-spectrogram.” EX1045-Zhao, 2370. *Zhao* also teaches a PostNet (“PostNet: Five stacked 1-D convolutional layers”) EX1045-Zhao, 2370.

223. 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*. A PHOSITA would have appreciated that this was common practice. For example, *Zhao* passes “the hidden linguistic representation sequence to the decoder, which consists of a location-sensitive attention mechanism [47] and a decoder LSTM, to predict the raw Mel-spectrogram.” EX1045-Zhao, 2370. *Zhao* also teaches a PostNet (“PostNet: Five stacked 1-D convolutional layers”) EX1045-Zhao, 2370. 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. *See* Section X.Q.1.

224. A PHOSITA would have been motivated to make the combination since these were all well-known concepts in the accent conversion space. *See* EX1045-Zhao, 2370-2371. 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. This was very common in the prior art, for example *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches one example of “an encoder-decoder architecture.” EX1045-Zhao, 2370. This encoder-decoder architecture uses an encoder to “consume” input sequences and generate a high-level hidden representation sequence. EX1045-Zhao, 2370. Then, Zhao passes “the hidden linguistic representation sequence to the decoder, which consists of a location-sensitive attention mechanism [47] and a decoder LSTM, to predict the raw Mel-spectrogram.” EX1045-Zhao, 2370. 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.

C. Claim 11

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.”

225. 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. *See* Section IX.I.3. Further, as discussed regarding Claim 1(c), *Feinauer-Prabhavalkar-Fan* teaches extracting linguistic features from input chunks and context. *See* Section IX.I.4. Claim 1(d-e) explains that acoustic features in chunks with context – output chunks. *See* Section IX.I.5- IX.I.6. As explained regarding Claims 9-10, *Accentron* further teaches generating output chunks comprising melspectrograms. *See* Section X.Q.1-X.R.2.

226. 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, 17:33-51; *See* Section IX.I.6. Because the ’457 specification describes each chunk as being “provided to modules” and the “linguistic feature extraction module” as shown in Fig 3 is a parallel path for the audio signal with the acoustic feature extraction module, the system as-described operates on the same chunks and produces linguistic and acoustic feature output corresponding to each chunk. *EX1001*, 6:45-50, Fig. 3. 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 X.Q.1.

XI. 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.”*

227. 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. For example, *Ding* teaches most methods “implicitly assume that PPGs are speaker-independent and contain only linguistic information in an utterance. In practice, however, PPGs carry speaker individuality cues, such as accent, intonation, and speaking rate. As a result, these cues can leak into the voice conversion, making it sound similar to the source speaker.” EX1012-Ding, Abstract. Therefore, it was well known that PPGs will still retain some accent within the PPG, though at a reduced amount. *Fan’s* voice feature extraction model is a “deep neural network” and outputs

the linguistic features with the reduced accent. EX1006-Fan, [0085]; 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*

228. *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]. *Singh* use a loss function for making the features invariant to acoustic variability, where acoustic variability (a.k.a. a domain shift) may include “speaker accents.” EX1007-Singh, [0007]. 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[.]*” See Section IX.I.4. *Singh* additionally describes embodiments recognizing English spoken in a variety of accents, including a British, Australian, or American accent. EX1007-Singh, [0009], [0023]; (“domain adaptation can be performed by taking American accent and British accent as source domain and target domain respectively and vice-versa.”); [0034] (“target domain can be Australian speech”) *Singh* teaches “domain invariant features can be learned directly from raw speech with significant improvement over the baseline acoustic models trained without domain adaptation.”

EX1007-Singh, [0009]. Within this example for the domain of an American accent, *Singh's* system is trained to account for at a plurality of American sub-accents (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.*” A PHOSITA would have appreciated that if the domain is an American accent and *Singh* learns “domain invariant features” that *Singh* would be reducing *contributions due to a further accent present in the audio data.*

229. 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]. This neutral network is “trained to minimize the cross-entropy cost of the label classifier and at the same time maximize the cross-entropy cost of the domain classifier.” EX1007-Singh, [0008]; [0020].

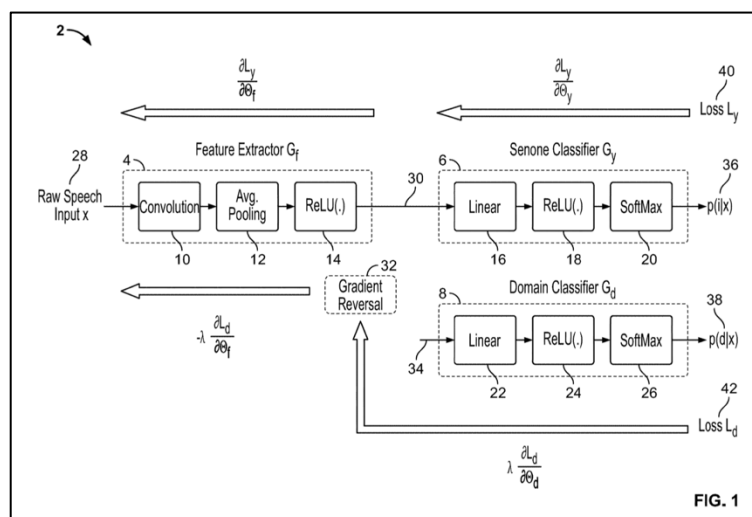


FIG. 1

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]. In examples, an utterance spoken with an American source accent may be adapted to a target British accent. EX1007-Singh, [0023] (“domain adaptation can be performed by taking American accent and British accent as source domain and target domain respectively and vice-versa.”) *Singh's* feature extractor “is a multi-layer CNN and takes the raw speech input vector x_i and generates a d -dimensional feature vector[.]” EX1007-Singh, [0019]. *Singh's* neural network “can be optimized by the standard stochastic gradient descent” at the end of the feature extractor and at the beginning of the domain classifier. EX1007-Singh, [0022]; Fig. 1.

230. *Singh* teaches using “an additional domain classifier along with the regular senone classifier and forcing the network during training to learn features from raw speech that are sufficiently discriminative for the senone classifier and invariant enough to fool the domain classifier.” EX1007-Singh, [0033]. Specifically, the senone “classifier generates an output 36 and the domain classifier generates an output 38. The system of the present disclosure can calculate a loss L_y based on the output 36 of the label classifier 6 and a loss L_d 42 based on the output 38 of the domain classifier 8. At training time, the label classifier's loss can be computed

only over labeled samples from $S(x,y)$, whereas the domain classifier's loss can be computed over both, labeled samples from $S(x,y)$ and unlabeled samples from $T(x)$.” EX1007-Singh, [0018]. Overall “the features f_i can be discriminative enough to perform good prediction on samples from the source domain. At the same time, the extracted features can be invariant enough to the shift in domain. In order to obtain domain invariant features, the parameters of feature extractor Θ_f can be optimized to maximize the domain classification loss $L_y L_y (\Theta_f \Theta_d)$ while, at the same time, domain classifier Θ_d can classify the input features. In other words, the domain classifier of the trained network can be configured to not be able to correctly predict the domain labels of the features coming from the feature extractor.” EX1007-Singh, [0021].

231. In sum, *Singh's* feature extraction neural network is trained to reduce contributions due to the speaker's American accent. Thus, when *Singh's* approach is used to train the *Feinauer-Prabhavalkar-Fan* LFEM, *Feinauer-Prabhavalkar-Fan-Singh* teaches the *linguistic feature* “neural network being trained... to reduce... contributions due to a further accent present in the audio data.” A PHOSITA would have appreciated that if the domain is an American accent and *Singh* learns “domain invariant features” that *Singh* would be reducing *contributions due to a further accent present in the audio data*. EX1007-Singh, [0009].

b. Motivation to Combine Singh

232. A PHOSITA would have been motivated to utilize *Singh's* teachings regarding neural network training to train *Feinauer-Prabhavalkar-Fan's* LFEM (linguistic features extraction module) in order to increase the accuracy and efficiency of the accent modification. *Fan* already taught “the voice feature extraction model may be a deep neural network[,]” EX1006-*Fan*, [0085]. It was well known to use a domain loss function in order to train neural networks as demonstrated by *Singh*. EX1007-*Singh*, [0009], [0018], and [0020]-[0023]. As another example, U.S. Patent Appl. No. 2020/0134425 titled *Systems and methods for domain adaptation in neural networks using cross-domain batch normalization* to Yoo et al. (“*Yoo*”) teaches in “speech recognition domains” to use a “loss function...for the model to reduce the difference of overall distribution between source and target data.” EX1039-*Yoo*, [0062]-[0063], [0093]. *Yoo's* loss function is performed “via back-propagation from the second layer of the second neural network (the output/activation layer of the second neural network) using a first discrepancy/loss function.” EX1039-*Yoo*, [0066].

233. 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. It was well known that accent (domain) invariant features, reduced accents, improve the overall linguistic feature recognition

and extraction. For example, an article titled *Adversarial Learning Of Raw Speech Features For Domain Invariant Speech Recognition* to Tripathi et al. (“Tripathi”) teaches to “mitigate speaker accent” due to “domain invariant features...learned directly from raw speech to **mitigate acoustic variabilities due to** speaker gender and **accent** which often adversely affect ASR performance” EX1040-Tripathi, 2rhc, 4rhc. 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.

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.”*

234. See Claim 6.

XII. GROUND 4: CLAIMS 7 IS OBVIOUS OVER FEINAUER, PRABHAVALKAR, FAN, AND WU

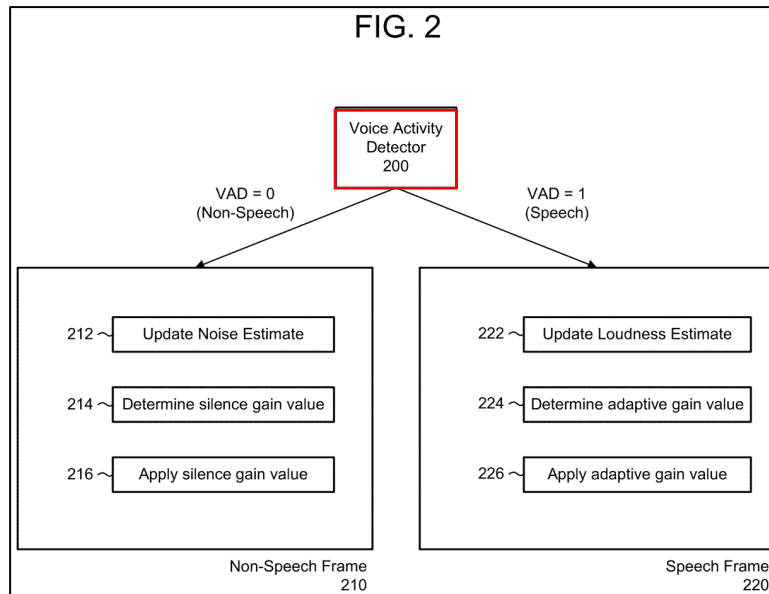
A. Claim 7

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.”*

235. 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. To calculate loudness or volume, the energy of a signal is measured over time. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“*Fletcher*”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point.” EX1025-Fletcher, 83. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. Therefore, there is a direct mathematical correlation between volume and energy. *See* Section IX.I.3. *Feinauer-Prabhavalkar-Fan* 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*

236. *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, “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.

237. *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[.]” EX1009-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. Like the ’457 Patent, *Wu* repetitively defines “energy level” as “amplitude.” EX1009-Wu, 14:15-16 *with* ’457 Patent, 5:61. As an example, *Wu* discloses “illustrat[ing] the raw audio data 512 as an audio waveform, with an upper and lower envelope...corresponding to an amplitude (e.g., energy level)[.]” EX1009-Wu, 14:13-16. Moreover, *Wu* describes that “[e]nergy levels may be illustrated using decibels (dB)” which a PHOSITA would have understood is a

measure of amplitude. EX1009-Wu, 4:31-34; EX1025-Fletcher, 108 (“amplitude of the components of about ± 0.5 db”).” *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

238. A PHOSITA would have found it obvious and been motivated to further modify *Feinauer-Prabhavalkar-Fan* to include *Wu*’s VAD to “not only improves the performance of downstream components by discarding non-speech signals, but also significantly reduces the overall computational cost due to its relatively small size.” EX1026-Chang, 1. A PHOSITA would have also appreciated including a VAD to “eliminate dead space” and “improve utilization efficiency of...storage resources.” See U.S. Patent No. 11,138,334 to Garrod et al., EX1017-Garrod, 10:30-33. 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). Voice activity detection was very common in the prior art. EX1017-Garrod, 10:30-34 (“[v]oice activity detection...main function is to eliminate dead space, to improve utilization efficiency of more compute-intensive resources, such as the ASR engine, or of storage resources. VAD algorithms are well known in the art.” EX1026-Chang,

1 (“In modern speech processing systems, voice activity detection (VAD) usually lives in the upstream of other speech components such as speech recognition and speaker recognition.”). As discussed above it was understood prior to the ’457 Patent that “VAD algorithms are well known in the art.” 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.

XIII. 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.”*

239. 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*

240. *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.

i. “prior to dividing the speech audio signal”

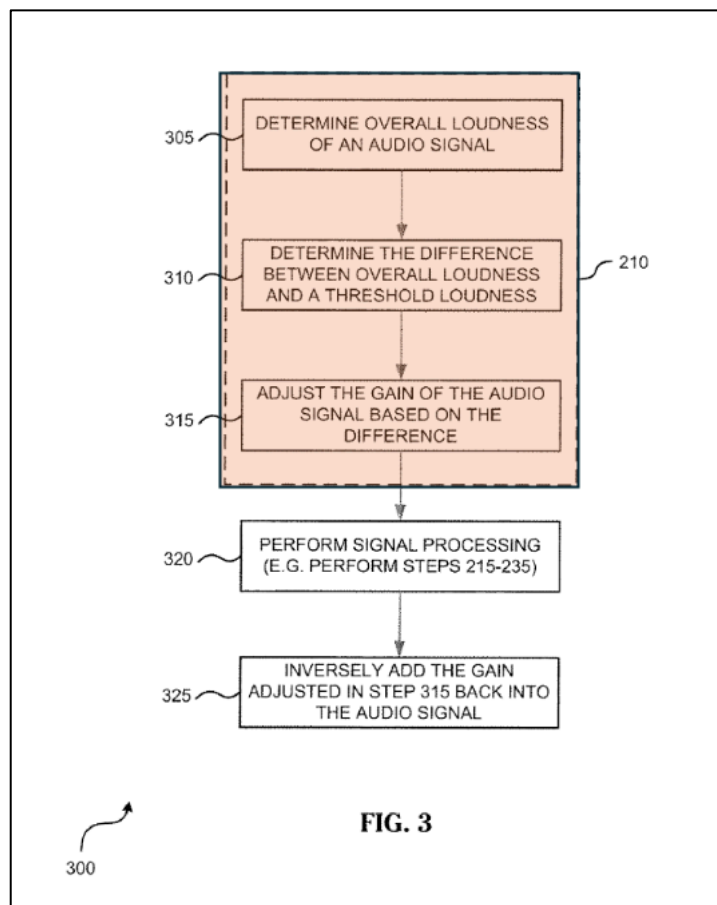
241. As discussed in Claim 1(a), *Feinauer-Prabhavalkar* teaches dividing a speech utterance into chunks of 5 frames each. *See* Section IX.I.2 *citing* EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. As will be discussed, a PHOSITA would have implemented *Bongiovi's* teachings of pre-processing the audio's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

ii. “processing the speech audio signal...”

b. *Bongiovi's Teachings*

242. *Bongiovi* teaches digitally processing an audio signal using a computer chip, such as...digital signal processor, or DSP. EX1010-Bongiovi, Abstract, [0052]. *Bongiovi* teaches digitally processing an audio signal using a computer chip, such as...digital signal processor, or DSP. EX1010-Bongiovi, Abstract. *Bongiovi* specifically teaches “digitally processing an audio signal.” EX1010-Bongiovi, Abstract. *Bongiovi's* digitally processing an audio signal “methods 300...are

executable on a computer chip, such as...digital signal processor, or DSP.” EX1010-Bongiovi, [0052]. A PHOSITA would have recognized a DSP satisfies the claimed *digital signal processing module*.



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

243. When the audio comes into *Bongiovi's* device, the overall loudness of the received signal is determined. EX1010-Bongiovi, [0047] (“step 305, the overall loudness of the received signal from step 205 is determined.”) 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*. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“*Fletcher*”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point” which can occur in a “speech” signal. EX1025-Fletcher, 83; 103; 82. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. *Bongiovi* then teaches “[a]fter the gain is adjusted in step 315, the signal can be further processed in step 320.” EX1010-Bongiovi, [0050].

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

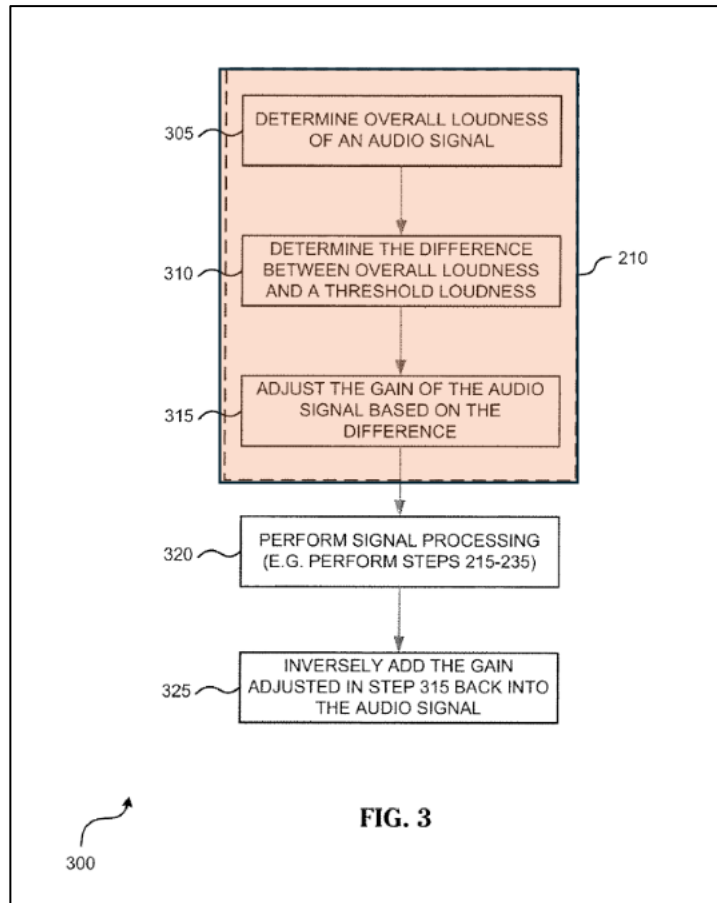
244. *Bongiovi* teaches performing steps 215-235 adjusting the volume of the input audio signal to reduce distortion. EX1010-Bongiovi, [0046]. (“when the volume is substantially lowered or raised, downstream signal processing done by unit 165 may be distorted because the loudness is not at the expected level...Thus, method 300 can be used to generate an output signal with a predetermined loudness regardless of the volume of the input audio signal...unit 165 may perform steps 215-235 as designed and without added

distortion.) A PHOSITA would have appreciated that reduced distortion would have *improved* the quality for *extraction of the linguistic features and the acoustic features*. For example, feature extraction is improved by improving the signal to noise ratio of the speech. For example, U.S. Patent No. 11,632,456 to Mevorah et al. (“*Mevorah*”) teaches “the audio preprocessing module 822 may use a...gain and **improve** the signal to noise ratio (SNR) of the speech 816 from the user 802 with respect to speech 816 or noise from other sources.” EX1041-Mevorah, 10:42-46. (emphasis added) A PHOSITA would have appreciated that the signal to noise ratio is directly correlated to a “higher risk of corrupting speech (e.g., distortion) is presented when attempting to remove noise.” EX1042-Jin, [0018]. It was well known that the “effects of distortion and interference are more damaging to speech recognition.” EX1043-Pereira, 2:21-24. Severe distortions “due to interfering sounds and speaker interference reduce accuracy of identification and increase cost. Major improvements in identification are possible if the sources of interference are identified and their signals or impact is reduced.” EX1043-Pereira, 7:2-6. A PHOSITA would have appreciated that it was common to preprocess audio to reduce the distortion and then perform feature extraction on linguistic and acoustic features on the clean/reduced distortion audio. EX1043-Pereira, 15:21-22. (“performed on the clean audio to detect the speech phonemes.”); EX1043-Pereira, Figs. 2A, 3B;

EX1043-Pereira, 7:36-37 (“a pitch calculation function is performed.”); and EX1043-Pereira, 7: 37-38 (“the energy of the incoming audio signal is calculated”).

c. *Motivation to Combine Bongiovi*

245. 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. *Bongiovi* teaches figures 2 and 3 are related steps, for example in “step 305, the overall loudness of the received signal from step 205 is determined.” EX1010-Bongiovi, [0047]. Therefore, a PHOSITA would have appreciated that teachings from one figure apply to the other. *Bongiovi*’s digitally processing an audio signal “methods 300...are executable on a computer chip, such as...digital signal processor, or DSP.” EX1010-Bongiovi, [0052]. *Bongiovi* teaches that *Bongiovi*’s digital signal processor may perform processing “without added distortion.” EX1010-Bongiovi, [0046]. *Bongiovi* already teaches the gain is adjusted before processing happens. EX1010-Bongiovi, [0050]. (“After the gain is adjusted in step 315, the signal can be further processed in step 320.”).



EX1010-Bongiovi, Fig. 3. (annotated). *Prabhavalkar* already teaches processing audio utterances with “noisy environmental recordings.” EX1005-Prabhavalkar, 15:4-19. *Prabhavalkar* determines “from the audio data for the utterance, speech frames each representing audio characteristics over different periods having a same duration; successively processing chunks that each include different sets of speech frames with the speech recognition model” EX1005-Prabhavalkar, 6:56-62. Further, it was well known to preprocess the audio before splitting the audio into fragments. For example, *Mevorah* teaches in “real-time” the “audio data may be initially processed to remove non-speech audio data” and subsequently “multiple audio data

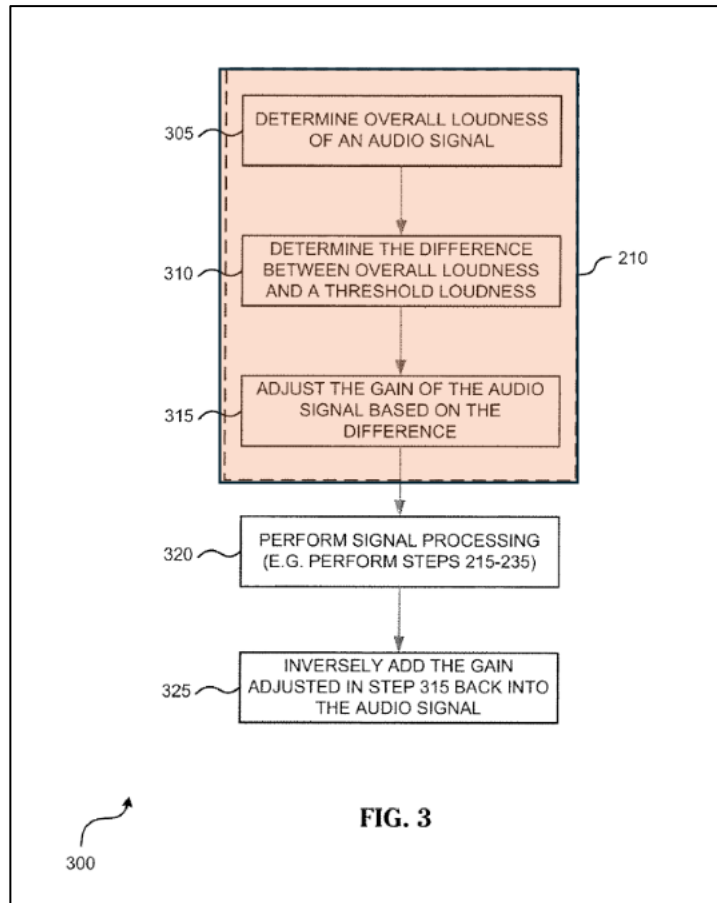
segments” are generated and the “segment of audio data is then processed to generate...character data.” EX1041-Mevorah, 5:29-35, 2:51-53; 27:9-18; 4:47-51.

246. 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 signals being processed. A PHOSITA would have appreciated that such a combination would have been a combination of known elements (DSP) that would yield predictable results, including the use of a DSP to adjust characteristics of a speech audio signal. For example, *Prabhavalkar* teaches to use “digital signal processing circuitry[.]” EX1005-Prabhavalkar, 26:43-45. As another example U.S. Patent Appl. No. 2013/0282373 to Visser et. al (“*Visser*”) teaches audio signal processing using a “digital signal processor” and that processor is used to perform functions on the electronic device” such as “voice activity detection” and “noise suppression” EX1044-Visser, [0260], [0271], Abstract, [0186]. 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.

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]; *Bongiovi* already teaches this. EX1010-Bongiovi, [0050]. (“After the gain is adjusted in step 315, the signal can be further processed in step 320.”) Further, it was well known within the art to pre-process audio data including, but not limited to, conditioning an audio signal with a constant loudness and/or gain level prior to the main processing. EX1010-Bongiovi, [0034]-[0035]. A PHOSITA would have found it obvious to incorporate such pre-processing into the methodology prior to fragmentation of the audio data in order to improve the rate at which the main audio data process occurs and to minimize a required overall processing time. Correspondingly, a PHOSITA would have appreciated that reducing distortion of *Feinauer’s* speech audio signal before further processing would result in improved accent modification quality, speed, and computational efficiency by reducing the amount of unwanted signal being processed.

248. A PHOSITA would have appreciated that such a combination would have been obvious to try in *Feinauer’s* accent conversion system, since *Bongiovi* already teaches adjusting the gain prior to processing the signal. Bongiovi, [0050]. (“After the gain is adjusted in step 315, the signal can be further processed in step 320.”)



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

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 as *Bongiovi* teaches. EX1010-Bongiovi, Abstract (“digitally process an audio signal in a manner such that studio-quality sound that can be reproduced.”); and EX1010-Bongiovi, [0002] (“presenting enhanced quality sound in a high noise environment.”). It was well known to pre-process audio before further speech processing to desirably improve performance using a DSP, such as by improving the

signal to noise ratio of the speech. For example, feature extraction is improved by improving the signal to noise ratio of the speech. For example, U.S. Patent No. 11,632,456 to Mevorah et al. (“*Mevorah*”) teaches “the audio preprocessing module 822 may use a...gain and **improve** the signal to noise ratio (SNR) of the speech 816 from the user 802 with respect to speech816 or noise from other sources.” EX1041-Mevorah, 10:42-46. (emphasis added) A PHOSITA would have appreciated that the signal to noise ratio is directly correlated to a “higher risk of corrupting speech (e.g., distortion) is presented when attempting to remove noise.” EX1042-Jin, [0018]. It was well known that the “effects of distortion and interference are more damaging to speech recognition.” EX1043-Pereira, 2:21-24. Severe distortions “due to interfering sounds and speaker interference reduce accuracy of identification and increase cost. Major improvements in identification are possible if the sources of interference are identified and their signals or impact is reduced.” EX1043-Pereira, 7:2-6.

250. A PHOSITA would have appreciated that it was common to preprocess audio to reduce the distortion and then perform feature extraction on linguistic and acoustic features on the clean/reduced distortion audio. EX1043-Pereira, 15:21-22. (“performed on the clean audio to detect the speech phonemes.”); EX1043-Pereira, Figs. 2A, 3B; EX1043-Pereira, 7:36-37 (“a pitch calculation function is performed.”); and EX1043-Pereira, 7: 37-38 (“the energy of the incoming audio

signal is calculated”). 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.

B. Claim 13

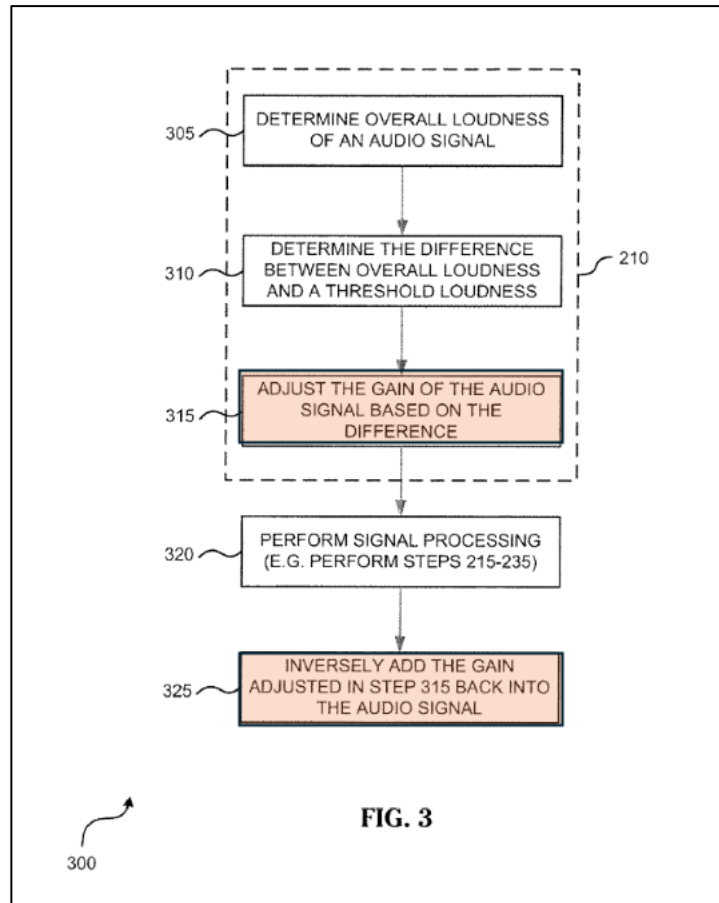
1. Claim 13: “The method for real-time correction of claim 1, further comprising: prior to dividing the speech audio signal, processing the speech audio signal by a digital signal processing module to adjust a loudness of the speech audio signal from a first level to a second level; after generating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

251. *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 XIII.X.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.

252. *Bongiovi* teaches digitally processing an audio signal using a computer chip, such as...digital signal processor, or DSP. EX1010-Bongiovi, Abstract, [0052]. *Bongiovi* specifically teaches “digitally processing an audio signal.”

EX1010-Bongiovi, Abstract. *Bongiovi's* digitally processing an audio signal “methods 300...are executable on a computer chip, such as...digital signal processor, or DSP.” EX1010-Bongiovi, [0052]. A PHOSITA would have recognized *Bongiovi's* DSP satisfies the claimed *digital signal processing module*.

253. *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*. As taught by *Bongiovi*, the digital signal processor performs the methods described within *Bongiovi*, such as the gain adjustment. EX1010-Bongiovi, [0052].

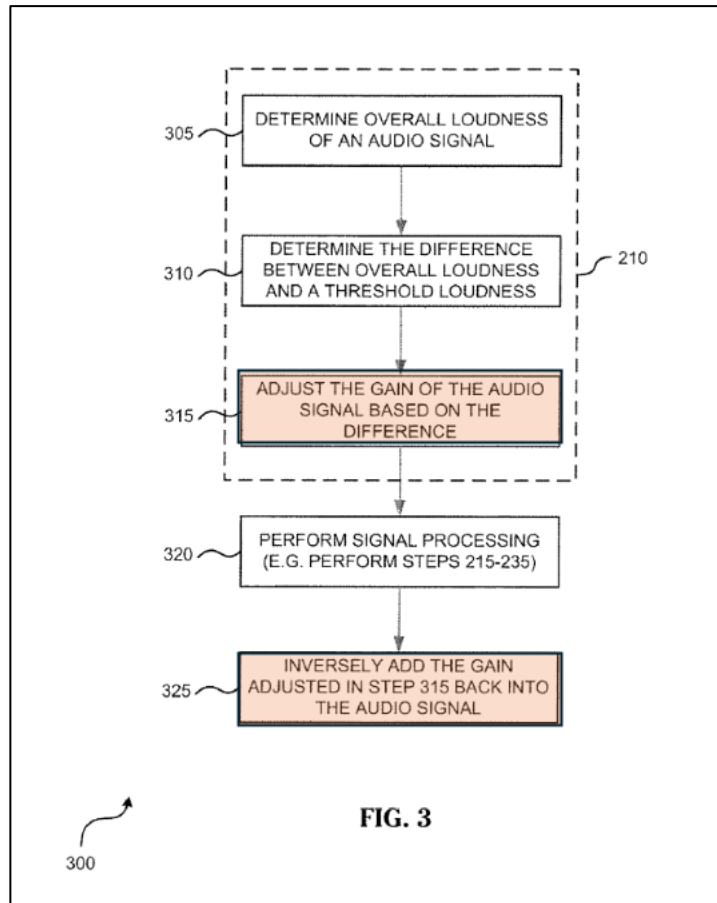


EX1010-Bongiovi, Fig. 3. (annotated).

a) Motivation to Combine Bongiovi

254. 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. *Bongiovi* teaches figures 2 and 3 are related steps, for example in “step 305, the overall loudness of the received signal from step 205 is determined.” EX1010-Bongiovi,

[0047]. Therefore, a PHOSITA would have appreciated that teachings from one figure apply to the other. *Bongiovi's* digitally processing an audio signal "methods 300...are executable on a computer chip, such as...digital signal processor, or DSP." EX1010-Bongiovi, [0052]. *Bongiovi* teaches that *Bongiovi's* digital signal processor may perform processing "without added distortion." EX1010-Bongiovi, [0046]. *Bongiovi* already teaches after processing the audio signal, adding the gain back into the signal. EX1010-Bongiovi, [0050]. ("step 325, the gain add 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.")



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

255. A PHOSITA would have been motivated to add the volume (gain) back into each output chunk until all the chunks are combined to enable real-time processing. *Prabhavalkar* teaches the “model may shift attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance while a second chunk or segment of the utterance is still being spoken.” EX1005-Prabhavalkar, 17:33-51. Bongiovi already teaches the last step of method 300 is that the gain is added back into the signal after processing.

EX1010-Bongiovi, [0050], Fig. 3. 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 IX.I.2.

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 in improved efficiency and user experience. *Bongiovi* already teaches the last step of method 300 is that the gain is added back into the signal after processing, and a PHOSITA would have appreciated that the “volume control is mimicked” improving sound quality which would improve the user’s experience. EX1010-Bongiovi, [0009], [0050], Fig. 3. 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. For example, *Prabhavalkar* teaches to use “digital signal processing circuitry.” EX1005-Prabhavalkar, 26:43-45. As another example U.S. Patent Appl. No. 2013/0282373 to Visser et. al (“*Visser*”) teaches audio signal processing using a “digital signal processor” and that processor is used to perform functions on the electronic device” such as “voice activity detection” and “noise suppression” EX1044-Visser, [0260], [0271], Abstract, [0186]. It was well known in the prior art to preprocess audio by adjusting the gain

using a DSP. For example, U.S. Patent No. 11,632,456 to Mevorah et al. (“*Mevorah*”) teaches “the audio preprocessing module 822 may use a...gain and improve the signal to noise ratio (SNR) of the speech 816 from the user 802 with respect to speech 816 or noise from other sources.” EX1041- Mevorah, 10:42-46. It was well known that the “effects of distortion and interference are more damaging to speech recognition.” EX1043-Pereira, 2:21-24. Severe distortions “due to interfering sounds and speaker interference reduce accuracy of identification and increase cost. Major improvements in identification are possible if the sources of interference are identified and their signals or impact is reduced.” EX1043-Pereira, 7:2-6.

XIV. 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.”*

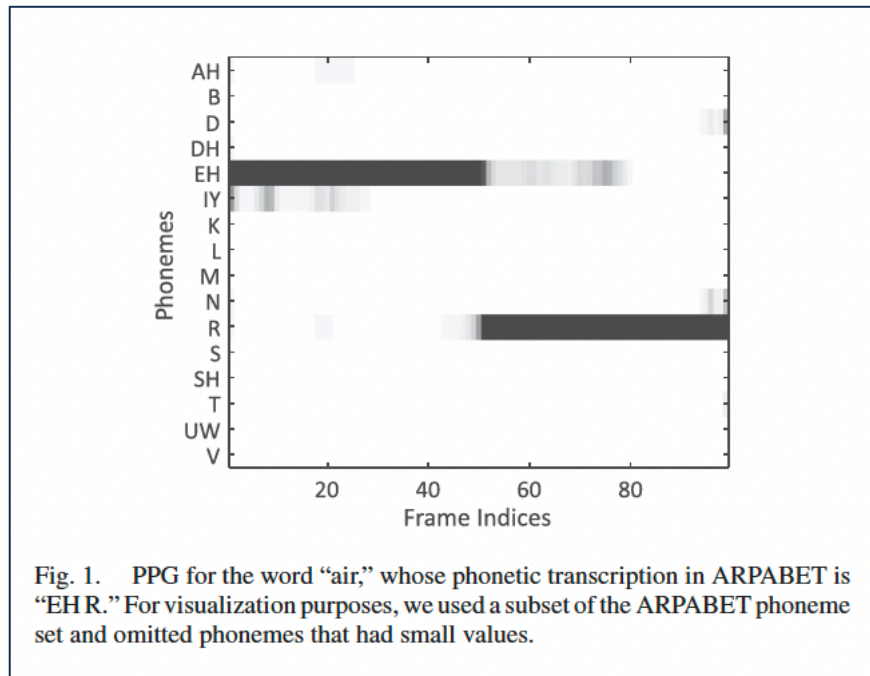
257. *Feinauer-Prabhavalkar-Fan* in view of *Osuna* teach phonetic posteriorgrams with a standardized phonetic dictionary (per *Fan*’s PPGs and *Osuna*’s ARPAbet dictionary when using PPGs).

258. 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 IX.I.4 citing EX1004-Feinauer, [0083], [0094], [0099]; EX1006-Fan, [0007], [0085]-[0086], Fig. 5. 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’s* dictionaries may “be compiled of existing *linguistic databases*[,]” that may be downloaded into memory teaching or rendering obvious *a data-driven phonetic library*. EX1004-Feinauer, [0066], [0081]. As *Feinauer* teaches, a linguistic database would be made up of words. EX1004-Feinauer, [0038] (“determine the equivalent word by referencing an uploaded dialect dictionary”); [0066] (“dictionaries 336 can be used to replace a word with another word, such as in the implementation of dialect modification.”), [0081]. *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*

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



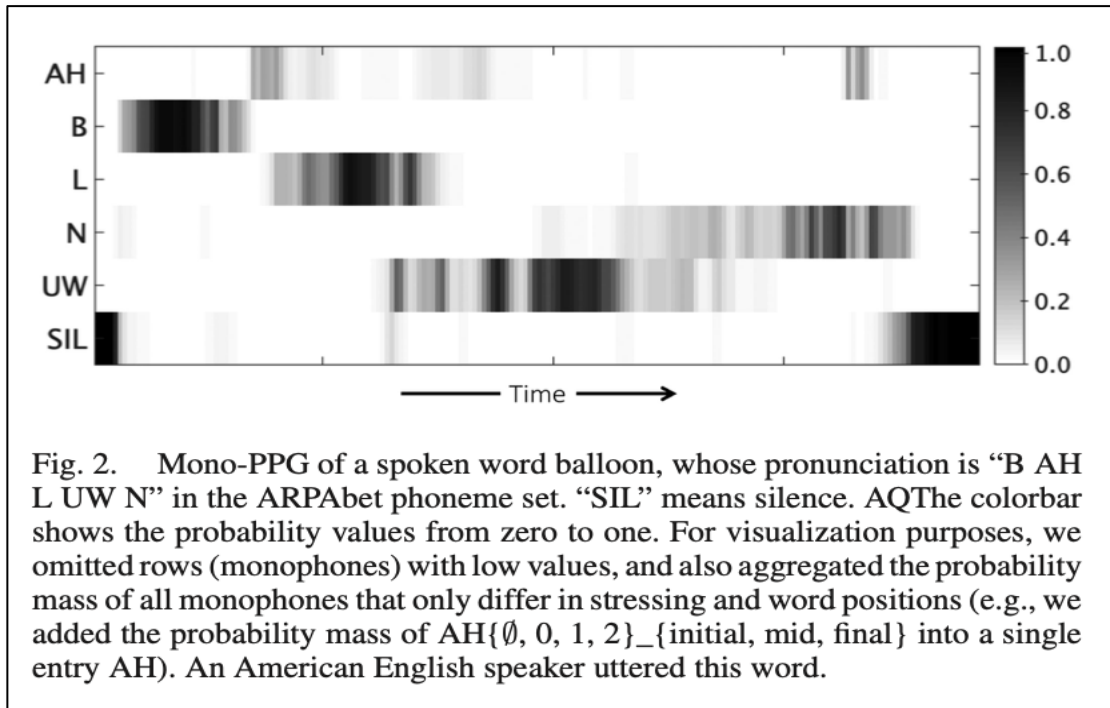
EX1011-Osuna, Fig. 1.

260. A PHOSITA would understand that any PPG will have a representation of phonemes because the first P in PPG refers to phonemes. A PHOSITA would realize that ARPAbet was a well-known standard way of representing phonemes. For example, ARPAbet, uses “ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters.” EX1021-Schmandt, 16. ARPAbet was developed “for the convenience of computer manipulation and representation of speech using ASCII-printable characters.” EX1021-Schmandt, 16.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[æ s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oo]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɜ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

Figure 27.2 ARPAbet symbols for American English vowels, with IPA equivalents.

EX1029-Jurafsky, 2. This was well known in the prior art. For example, *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) uses the ARPAbet standard phoneme representation for its PPG, as shown in the PPG that is depicted in Fig 2.



EX1045-Zhao, Fig. 2. A PHOSITA would have recognized *Osuna’s* ARPAbet dictionary satisfies the claimed *standardized phonetic dictionary*.

b. *Motivation to Combine Osuna*

261. A PHOSITA would have found it obvious and would have been motivated to use *Osuna’s* teachings of using a ARPAbet dictionary to *Feinauer’s* processing steps of determining PPGs. 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. This was well known. As an example, *Speech and Language Processing*, Chapter 27 “Phonetics” to Jurafsky et al. (“*Jurafsky*”), one phonetic alphabet is the International Phonetic

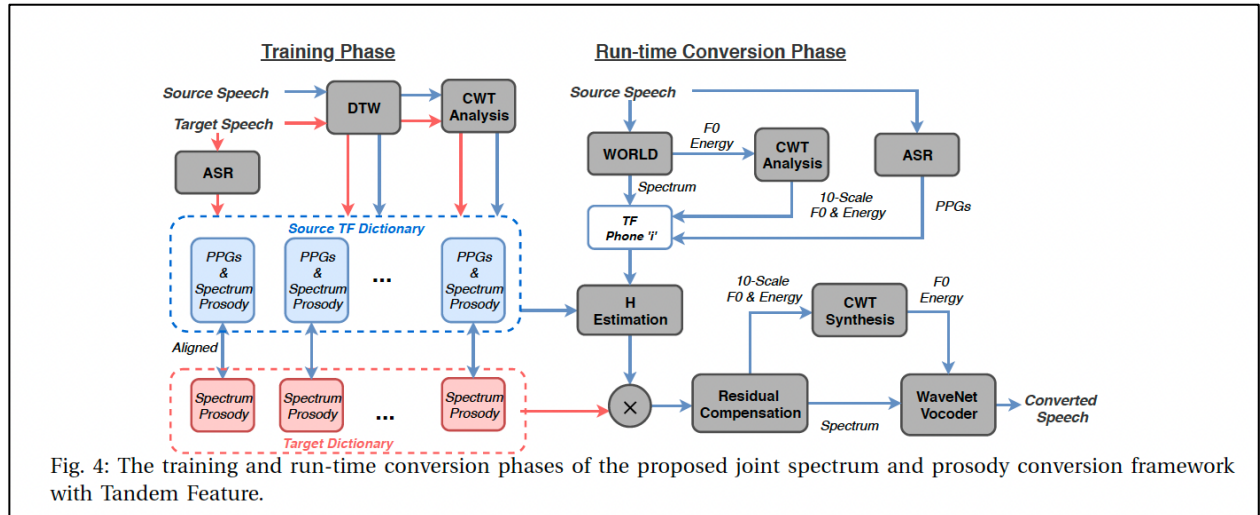
Alphabet (IPA), which is an evolving standard originally developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. EX1029-Jurafsky, 2. Another phonetic alphabet is the ARPAbet, which uses “ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters.” EX1021-Schmandt, 16. ARPAbet was developed “for the convenience of computer manipulation and representation of speech using ASCII-printable characters.” EX1021-Schmandt, 16.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[æ s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɜ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

Figure 27.2 ARPAbet symbols for American English vowels, with IPA equivalents.

EX1029-Jurafsky, 2. 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.

262. 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. This was well known in the prior art, for example is *Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion* to *Sisman* et al. ("Sisman") teaches "to convert all speaker's voice into PPGs, and construct dictionaries that consist of source spectrum and prosody features as well as PPGs." EX1023-Sisman, 6. *Sisman* teaches "a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary[.]" EX1023-Sisman, 6.



EX1023-Sisman, Fig. 4. *Sisman*'s PPG dictionary has "multiple coupled dictionaries $[A_i; B_i]$, one for each phone i , where $i = 1, \dots, N$, A_i is the source phonetic dictionary, and B_i the target phonetic dictionary. The source phonetic dictionary here

consists of both PPGs, source spectrum features as well as CWT-based prosody features, called source TF-dictionary while the target phonetic dictionary B_i only includes spectrum features and prosody features, called target dictionary.” EX1023 Sisman, 8. As another example, Li teaches “a phonetic segment classifier to generate segmental phonetic posterior-grams (SPPGs) to represent phone segment-level information.” EX1031-Li, Abstract.

Table 4. The explored non-category set comparison between this work and the previous work [11].

explored additional non-categories			explored but existing non-categories			missing
ah_ax	ax_ix	ih_ix	ax_er	aa_ao	ey_ih	aa_ax
er_r	ch_t	g_k	eh_ey	ae_ay	d_t	aw_ax
r_w	dh_l	s_z	l_n	b_p	f_v	
			m_n			

Table 5. Non-categories caused by articulation absence from L1 language. The absence types are adapted from [26].

	non-categories		
Missing voiced plosives	b_p	d_t	g_k
Missing affricates	ch_t		
Missing fricatives	f_v	dh_l	s_z
Missing and confused approximants	r_w		
Confusion among [l] and [n]	l_n		

EX1031-Li, Tables 4-5.

XV. 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.”

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

XVI. CONCLUSION

264. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: March 18, 2026

By: 
Christopher Schmandt

Christopher Schmandt

4 Longfellow
Winchester, MA 01890
+1-617-230-4257

Education

M.I.T., Master of Science, Visual Studies (Computer Graphics), 1980

M.I.T., Bachelor of Science, Computer Science, 1978

Professional Experience -- MIT

Media Laboratory, Principal Research Scientist, 1985-2018 (retired)
Director, Living Mobile Research Group (formerly Speech + Mobility)

Architecture Machine Group, Research Associate, 1980-1984

Architecture Machine Group, Research Assistant, 1979-1980

Architecture Machine Group, Graphics Programmer, 1977-1979

Departmental Undergraduate Research Opportunities Program Coordinator, 1984-2018

Laboratory Intellectual Property Committee 2001-2016, chair 2002-2009

Departmental Committee on Graduate Studies, 1996-2001, 2007-2018

Sponsored Research Activities

Alerting and Mobile Messaging, Digital Life Consortium, MIT Media Lab, 1997-2018

Acoustical Cues to Discourse Structure, National Science Foundation, Principal Investigator, 1995-1998

Parsing Radio, News in the Future Consortium, MIT Media Lab, 1993-1999

Desktop Audio, SUN Microsystems, Inc., Principal Investigator, 1989-1996

Voice Interaction in Hand Held Computers, Apple Computer, Principal Investigator, 1991-1993

Voice Interfaces for Network Services, AT&T, Principal Investigator, 1989-1991

Back Seat Driver, NEC, Principal Investigator, 1988-1991

Acoustic and Visual Cues for Speech Recognition, DARPA, co-Principal Investigator, 1986-1988

Personal Computers and Telephony, NTT Public Corporation, Principal Investigator, 1984-1989

Home Telecomputing, Atari, Inc., Principal Investigator, 1983

Professional experience -- intellectual property

Testifying legal engagements within the last five years, representing party in italics

- *Microsoft v. SpeakWare*, 2019
- *Bumble Trading LLC v. Match Group LLC*, 2019
- *Tile Inc. v. Cellwitch Inc.*, 2019
- *Blackberry Ltd v. Facebook, Inc* 2020
- *EROAD Ltd v. PerDiem Co.*, 2020
- *Snap Inc. v. SRK Technology LLC*, 2020
- *Shopify Inc Ltd v. Express Mobile Inc.*, 2020
- *Express Mobile Inc, v Wix Ltd* 2020
- *Motorola Mobility LLC v. Ironworks patents, LLC.* 2021
- *Express Mobile Inc, v GoDaddy Inc* 2021
- *Bumble Trading LLC v. Kinectus, LLC.* 2021
- *Tile, Inc v. Linquet Technologies, Inc.* 2021
- *Facebook, Inc. v. Palo Alto Research Center, Inc.* 2021
- *Unified Patents, LLC v. Gesture Technology Partners, LLC.* 2021
- *Quantum Metric, Inc v. Content Square Isreal, Ltd.* 2021
- *Meta Platforms Inc. v. AlmondNet, Inc* 2022
- *Apple Inc. v. Zentian Limited* 2022
- *Google LLC v. Mira Advanced Technology Systems Inc* 2022
- *Google LLC v. Robocast* 2023
- *Google LLC v. Buffalo Patents LLC* 2023
- *Amazon v. AlmondNet* 2023
- *Google LLC v. Valtrus* 2023
- *Samsung Electronics Co., Ltd v. Stanton Techiya, LLC* 2024
- *Roku, Inc. v. AlmondNet, Inc.* 2024
- *LinkedIn v. AlmondNet, Inc* 2024
- *Google, LLC v. Dialect, LLC* 2024
- *SAP America, Inc. v. Cyandia, Inc.* 2024
- *Tableau Software LLC v. iCharts LLC* 2024
- *Google LLC and Samsung Electronics Company., Ltd. v. Cerence Operating Company* 2024
- *Samsung Electronics Company., Ltd. v. Cerence Operating Company* 2025
- *Amazon.com, Inc. v. KAIFI* 2025

Co-inventor on U.S. Patents

- 5,177,685 for "Automobile navigation system using real time spoken driving instructions"
- 6,728,348 for "System for storing voice recognizable identifiers using a limited input device such as a telephone key pad"
- 6,937,986 for "Automatic dynamic speech recognition vocabulary based on external sources of information"
- 7,098,776 for "Methods and apparatus for vibrotactile communication"
- 7,392,280 for "Method for summarization of threads in electronic mail"
- 7,738,637 "Interactive voice message retrieval"
- 7,443,283 for "Methods and apparatus for connecting an intimate group by exchanging awareness cues and text, voice instant messages, and two-way voice communications"
- 7,865,560 for "System for summarization of threads in electronic mail"
- 8,121,653 for "Methods and apparatus for autonomously managing communications using an intelligent intermediary"
- 8,135,128 for "Animatronic creatures that act as intermediaries between human users and a telephone system"

Co-inventor on U.S. Patent Applications

- 20020076009 "International dialing using spoken commands"
- 20020087328 "Automatic dynamic speech recognition vocabulary based on external sources of information"
- 20020064257 "System for storing voice recognizable identifiers using a limited input device such as a telephone keypad"
- 20030023688 "Voice-based message sorting and retrieval method"
- 20030081738 "Method and apparatus for improving access to numerical information in voice messages"
- 20030144846 "Method and system for modifying the behavior of an application based upon the application's grammar"
- 20030158903 "Calendar bar interface for electronic mail interaction"

Publications

Field Study of a Tactile Sound Awareness Device for Deaf and Hard of Hearing Users ISWC 2020. (with D. Jain, B. Chiu, S. Goodman, L. Findlater, J. Froehlich)

SkinMorph: Texture-Tunable On-Skin Interface Through Thin, Programmable Gel ISWC 2018. (with Cindy Kao, M. Banforth, D. Kim)

Technical Interventions to Detect, Communicate, and Deter Sexual Assault. ISWC 2017. (with Manisha Mohan)

Exploring Interactions and Perceptions of Kinetic Wearables. DIS 2017. (with Cindy Hsin-Liu Kao, D. Ajilo, O. Anilonyte, A. Dementyev, I. Choi and S. Follmer)

Leveraging User-made Predictions to Help Understand Personal Behavior. MobileHCI 2017. (with Miriam Greis, Tilman Dangler and Albrecht Schmidt)

Rovables: On-Body Robots as Mobile Wearables. UIST 2016. (with Cindy Hsin-Liu Kao, A. Dementyev, I. Choi, D. Ajilo, M. Xu, and S. Follmer)

DuoSkin: Rapidly Prototyping On-Skin User Interfaces Using Skin-Friendly Materials. ISWC 2016. (with Cindy Hsin-Liu Kao, Christian Holz, Asta Roseway, and Andres Calvo)

Immersive Terrestrial Scuba Diving Using Virtual Reality (with Dhruv Jain, Misha Sra, Jingru Go, Rodrigo Margues, Raymond Wu and Justin Chiu) Proceedings, UIST 2016

Expanding social mobile games beyond the device screen (with Misha Sra) Journal of Personal and Ubiquitous Computing, 2015

NailO: Fingernails as an input surface (with Cindy Hsin-Liu Kao, Artem Dementyev, Joseph Paradiso) CHI 2015

Mugshots: A mug display for front and back stage social interaction in the workplace (with Cindy Hsin-Liu Kao) TEI (Tangible and Embedded Interfaces) 2015

Mime: compact, low power 3D gesture sensing for interaction with head mounted displays (with Andrea Colaco, Ahmend Kirmani, Hye Soo Yang, Nan-Wei Gong, and Vivek Goyal) Proceedings of UIST 2013.

Spotz: A location-based approach to self-awareness (with Misha Sra) Proceedings of Persuasive 2013.

Setting the stage for interaction: A tablet application to augment group discussion in a seminar class (with Drew Harry and Eric Gordon) Proceedings of CSCW 2012.

Indoor Location Sensing using Geo-Magnetism (with Jaewoo Chung, Matt Donahoe, Ig-Jae Kim, Pedram Razavai and Micaela Wiseman) Proceedings of International Conference on Mobile Systems, Applications, and

Services (Mobisys) 2011.

My second bike: a TV-enabled social and interactive riding experience (with Jaewoo Chung, Kuang Xu, Andrea Colaco, and Victor Li) Proceedings of IEEE Communications and Networking Conference, Jan 2010.

Going my way?: User-aware route planner (with Jaewoo Chung) Proceedings CHI 2009.

Globetoddler: Designing for remote interaction between preschoolers and their traveling parents (with Paulina Modlitba) CHI 2008 Extended Abstracts

Are we there yet? - a temporally aware media player (with Matt Adcock and Jaewoo Chung), Australian User Interface Conference (AUIIC) 2008

Physical embodiments for mobile communication agents (with Stefan Marti), UIST 2005

Giving the caller the finger: collaborative responsibility for cellphone interruptions (with Stefan Marti) Extended Abstracts, CHI 2005

Active Messenger: email filtering and delivery in a heterogeneous network (with Stefan Marti) Human-Computer Interaction Journal (HCI) Volume 20 (2005)

WatchMe: communication and awareness between members of a closely-knit group (with Natalia Marmasse) Proceedings of Ubicomp 2004

An audio-based personal memory aid (with S. Vemuri, W. Bender, S. Tellex and B. Lassey) Proceedings of Ubicomp 2004

Improving speech playback using time-compression and speech recognition (with Sunil Vemuri, Philip DeCamp, and Walter Bender) Proceedings of CHI 2004

Impromptu: managing networked audio applications for mobile users (with Kwan Lee, Jang Kim, and Mark Ackerman), Proceedings of MobiSys 2004

TalkBack: a conversational answering machine (with Vidya Lakshminpathy and Natalia Marmasse) Proceedings of UIST 2003

"ListenIn" to domestic environments from remote locations (with Gerardo Vallejo) Proceedings of the 2003 International Conference on Auditory Display (ICAD)

Safe & Sound: a wireless leash (with Natalia Marmasse) Extended Abstracts, Proceedings of CHI 2003

Mediated voice communication via mobile IP (with Jang Kim, Kwan Lee, Gerardo Vallejo, and Mark Ackerman), Proceedings of UIST 2002.

The Audio Notebook: Paper and pen interaction with structured speech (with Lisa Stifelman and Barry Arons), Proceedings of CHI 2001

Synthetic News Radio (with Keith Emnett) IBM Systems Journal, Vol. 39 Nos. 3-4, pp. 646-659, 2000.

Everywhere messaging (with Natalia Marmasse, Stefan Marti, Nitin Sawhney, and Sean Wheeler) IBM Systems Journal, Vol. 39 Nos. 3-4, pp. 660-677, 2000.

Location-aware information delivery with comMotion (with Natalia Marmasse), Proceedings of the Second International Symposium on Handheld and Ubiquitous Computing, pp. 157-171, Springer, 2000.

Nomadic Radio: Scalable and contextual notification for wearable audio messaging (with Nitin Sawhney), Proceedings of CHI 1999.

Speaking and listening on the run: Design for wearable audio computing (with Nitin Sawhney), Proceedings of International Symposium on Wearable Computing, 1998.

Audio Hallway: A virtual acoustic environment for browsing, Proceedings of UIST 1998.

Dynamic Soundscape: Mapping time to space for audio browsing (with Minoru Kobayashi), Proceedings of CHI 1997.

CLUES: Dynamic personalized message filtering (with Matt Marx), Proceedings of CSCW 1996.

Using acoustic structure in a hand-held audio playback device (with Deb Roy), IBM Systems Journal, Vol 35, Nos. 3 and 4, 1996.

Mailcall: Message presentation and navigation in a nonvisual environment (with Matt Marx), Proceedings of CHI 1996

AudioStreamer: Exploiting simultaneity for listening (with Atty Mullins), short paper, CHI 1995.

Multimedia momadic services on today's hardware, IEEE Network, September/October 1994.

Putting people first: Specifying proper names in speech interfaces (with Matt Marx), proceedings of UIST 1994.

Chatter: A conversational learning speech interface (with E. Ly) AAAI Workshop on Intelligent Multi-Media Multi-Modal Systems, 1994.

Voice Communication with Computers: Conversational Systems. New York: Van Nostrand Reinhold. 1994.

Capturing, structuring, and representing ubiquitous audio (with D. Hindus and C. Horner), ACM Transactions on Information Systems, Vol. 11, No. 4, October 1993.

Speech Recognition Architectures for Multimedia Environments, (with E. Ly and B. Arons), Proceedings of the 1993 AVIOS Conference, September 1993.

Phoneshell: the Telephone as Computer Terminal, Proceedings of the ACM Multimedia Conference, August 1993.

Voicenotes: A Speech Interface for a Hand-Held Voice Notetaker (with L. Stifelman, B. Arons, and E. Hulteen), Proceedings of INTERCHI'93, April 1993.

From Desktop Audio to Mobile Access: Opportunities for Voice in Computing, book chapter in Advances in Human-Computer Interaction, Vol. 4, H.R. Hartson and D. Hix editors. 1992.

Ubiquitous Audio: Capturing Spontaneous Collaboration (with D. Hindus), Proceedings of CSCW'92, November 1992.

Integrating Audio and Telephony in a Distributed Workstation Environment (with S. Angebrannt, R. Hyde, D. Luong, and N. Siravara), Proceedings of the Summer 1991 USENIX Conference, June 1991.

Augmenting a Window System with Speech Input (with M. Ackerman and D. Hindus), Computer, IEEE Computer Society, Vol. 23, No. 8, August 1990.

Observations on Using Speech Input for Window Navigation (with D. Hindus, M. Ackerman, and S. Manandhar), Proceedings, Human-Computer Interaction, Interact '90, IFIP, August 1990.

Phonetool: Integrating Telephones and Workstations (with S. Casner), Proceedings, GLOBECOM '89, IEEE Communications Society, November 1989.

Desktop Audio (with B. Arons), UNIX Review, October 1989.

Synthetic Speech for Real Time Direction-Giving (with J. Davis), IEEE Transactions on Consumer Electronics, IEEE, September 1989.

The Back Seat Driver: Real Time Spoken Driving Instructions (with J. Davis), Proceedings, IEEE Vehicle Navigation and Information Systems Conference, IEEE, Toronto, Canada, September 1989.

A Voice and Audio Server for Multimedia Workstations (with B. Arons, C. Binding, K. Lantz), Proceedings, Speech Tech 1989

An Audio and Telephone Server for Multi-media Workstations (with M. McKenna), Proceedings, Second IEEE Conference on Workstations, IEEE, Palo Alto, CA., 1988.

Employing Voice Back Channels to Facilitate Audio Document Retrieval, Proceedings, ACM Conference on Office Information Systems (COIS), Santa Clara, CA, 1988.

Conversational Telecommunications Environments, Proceedings, Second International Conference on Human-Computer Interaction, 1987.

Understanding Speech Without Recognizing Words, Proceedings, American Voice Input/Output Society Conference, AVIOS, 1987.

A Robust Parser and Dialog Generator for a Conversational Office System (with B. Arons and C. Simmons), Proceedings, American Voice Input/Output Society Conference, AVIOS, Palo Alto, CA, 1987.

Integrated Messages and Network Services for a Personal Workstation, IEEE Workshop on Telematics and Message Handling Systems, IEEE, 1986.

Voice Interaction in an Integrated Office and Telecommunications Environment, Proceedings, American Voice Input/Output Society Conference, AVIOS, San Francisco, CA, 1985.

Voice Communication with Computers, book chapter in Advances in Human-Computer Interaction, H. R. Hartson ed., 1985.

Speech Synthesis Gives Voiced Access to an Electronic Mail System, Speech Technology, Vol. 2, No. 3, Aug/Sept 1984.

A Conversational Telephone Messaging System (with B. Arons), IEEE Transactions on Consumer Electronics, IEEE, Vol CE-30, August 1984.

Phone Slave: A Graphical Telecommunications Interface (with B. Arons), Proceedings, Society for Information Display International Symposium, SID, San Francisco, CA, June 1984.

Input/Display Registration in a Stereoscopic Workstation, Displays, April 1984.

Remote Access to Voice and Text Messages, Proceedings, American Voice Input/Output Society Conference, AVIOS, Washington D.C., 1984.

Fuzzy Fonts: Analog Models Improve Digital Text Quality, Proceedings, National Computer Graphics Association Conference, Chicago, IL, 1983.

Greyscale Fonts Designed From Video Signal Analysis, Society of Applied Learning Technology, Houston, TX, 1983.

Spatial Input/Display Correspondence in a Stereoscopic Computer Graphic Work Station, Proceedings, ACM/SIGGRAPH, Detroit, MI, 1983.

A Programmable Virtual Vocabulary Speech Processing Peripheral (with W. Bender), Proceedings, American Voice Input/Output Society Conference on Voice Data Entry Systems Applications, AVIOS, 1983.

The Intelligent Voice Interactive Interface (with E. A. Hulteen), Proceedings, Human Factors in Computer Systems, National Bureau of Standards/ACM, Gaithersburg, MD, 1982.

Interactive Three-Dimensional Computer Space, Proceedings, SPIE Conference on Processing and Display of Three-Dimensional Data, SPIE, San Diego, CA, 1982, Vol. 367.

Speech Communications, a Systems' Approach, Proceedings, American Voice Input/Output Society Conference on Entry Systems Applications, 1982.

Voice Interaction: Putting Intelligence into the Interface, Proceedings, IEEE International Conference on Cybernetics and Society, IEEE, Seattle, WA, 1982.

The Intelligent Ear: A Graphical Interface to Digital Audio, Proceedings, IEEE International Conference on Cybernetics and Society, IEEE, Atlanta, GA, 1981.

Soft Typography, Information Processing 1980, IFIPS, S. Lavington ed., North-Holland Publishing Co., 1980.