

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. IPR2026-00272
U.S. Patent No. 11,948,550

DECLARATION OF CHRISTOPHER SCHMANDT

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS.....	13
A.	Educational Background and Professional Experience	13
II.	METHODOLOGY AND MATERIALS CONSIDERED	17
III.	OVERVIEW AND LEGAL STANDARDS	21
A.	Person of Ordinary Skill in the Art.....	22
B.	Obviousness	23
C.	Analogous Art	28
D.	Claim Construction.....	29
IV.	LEVEL OF A PERSON OF ORDINARY SKILL.....	29
V.	OVERVIEW OF TECHNOLOGY	30
A.	Machine Learning and Neural Networks	30
B.	Phonemes	32
C.	Phonetic Posteriorgrams (PPGs).....	35
D.	Prosodic Features.....	38
E.	Dialect and Accent	39
F.	Prior Art Accent Conversion Systems	40
1.	Encoder and Decoder	41
2.	Mel Spectrograms.....	42
3.	Vocoder.....	45
VI.	OVERVIEW OF THE '550 PATENT.....	46
A.	Description of the '550 Patent.....	46

B.	Field of Endeavor of the '550 Patent.....	48
C.	Problem Purportedly Solved by the Inventors of the '550 Patent	48
D.	Prosecution History of the '550 Patent	49
VII.	OVERVIEW OF THE PRIOR ART	49
A.	Overview of Fan	50
B.	Overview of Sonsaat	57
C.	Overview of Loots	58
D.	Overview of Feinauer	60
E.	Overview of Osuna.....	61
F.	Overview of Dalce	62
VIII.	SUMMARY OF UNPATENTABILITY	63
IX.	OPINIONS REGARDING GROUND I: <i>FAN, SONSAAT, AND LOOTS</i>	64
A.	Claim 1	64
1.	<i>Claim 1[Pre]: A system, comprising:</i>	64
2.	<i>Claim 1[a]: at least one processor; and</i>	66
3.	<i>Claim 1[b]: non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor to cause the system to:</i>	67
4.	<i>Claim 1[c]: train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;</i>	71

5. *Claim 1[d]: apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content; .84*
6. *Claim 1[e]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent, ...94*
7. *Claim 1[e][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and 104*
8. *Claim 1[f]: convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content. 112*

- B. Claim 2:** The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent. .114

- C. **Claim 3:** The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content. 117
- D. **Claim 4:** The system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.....118
- E. **Claim 7** 120
 - 1. *Claim 7[a]: The system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions are executable by the at least one processor to further cause the system to: 120*
 - 2. *Claim 7[b]: align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and 123*
 - 3. *Claim 7[c]: determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent. 126*
- F. **Claim 10:** The system of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor to further cause the system to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the

synthesized version of the received speech content having the second accent has the set of prosodic features.....127

G. Claim 11 130

1. *11[Pre]: A non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to: 130*
2. *11[a]: train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers; 130*
3. *11[b]: apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content; 131*
4. *11[c]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent, ...131*
5. *11[c][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content,*

	<i>wherein the first and second phonemes are different phonemes; and</i>	131
6.	<i>11[d]: convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.</i>	131
H.	Claim 12: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.....	132
I.	Claim 13: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.	132
J.	Claim 16	132
1.	<i>Claim 16[a]: The non-transitory computer-readable medium of claim 11, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions, when executed by the at least one processor, further cause the computing device to:</i>	132

2.	<i>Claim 16[b]: align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and</i>	<i>133</i>
3.	<i>Claim 16[c]: determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.</i>	<i>133</i>
K.	Claim 19	133
1.	<i>19[pre]: A method comprising:</i>	<i>133</i>
2.	<i>19[a]: training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;</i>	<i>133</i>
3.	<i>19[b]: applying the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;</i>	<i>134</i>
4.	<i>19[c]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent, ...</i>	<i>134</i>
5.	<i>19[c][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second</i>	

pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and 134

6. *19(d): converting the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content. 135*

L. Claim 21: The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content. 135

M. Claim 22 135

1. *Claim 22[a]: The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the method further comprises: 135*
2. *Claim 22[b]: aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and 135*
3. *Claim 22[c]: determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent. 135*

X.	OPINIONS REGARDING GROUND II: <i>FAN, SONSAAT, LOOTS, AND FEINAUER</i>	136
A.	Claim 8: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to transmit the synthesized version of the received speech content having the second accent to a second computing device.	136
B.	Claim 9	142
	1. <i>Claim 9[a]: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to:</i>	142
	2. <i>Claim 9[b]: receive, in real time, continuous speech content having the first accent; and continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.</i>	142
C.	Claim 17: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device.	151
D.	Claim 18	151
	1. <i>Claim 18[a]: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to:</i>	151
	2. <i>Claim 18[b]: receive, in real time, continuous speech content having the first accent; and</i>	151
	3. <i>Claim 18[c] continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after</i>	

	<i>receiving the continuous speech content having the first accent.</i>	151
E.	Claim 20	151
1.	Claim 20[a]: The method of claim 19, further comprising:	151
2.	Claim 20[b]: receiving, in real time, continuous speech content having the first accent; and	152
3.	Claim 20[c] continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.	152
XI.	OPINIONS REGARDING GROUND III: <i>FAN, SONSAAT, LOOTS, AND OSUNA</i>	152
A.	Claim 5: The system of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.	152
B.	Claim 14: The non-transitory computer-readable medium of claim 11, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.....	155
XII.	OPINIONS REGARDING GROUND IV: <i>FAN, SONSAAT, LOOTS, AND DALCE</i>	156
A.	Claim 6: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.	156
B.	Claim 15: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to	

receive a first user input indicating a selection of the first accent
and a second user input indicating a selection of the second
accent.....162

XIII. CONCLUSION.....163

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 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. 11,948,550 (“’550 Patent”) (EX1001), including the specification and claims, and the ’550 Patent’s prosecution history (“’550 File History”) (EX1002). In developing my opinions relating to the ’550 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. 11,948,550 (“’550 Patent”)
Exhibit 1002	Prosecution History for the 11,948,550 Patent (“’550 File History”)
Exhibit 1004	U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>Fan</i> ”)

Exhibit 1005	Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna, <i>L2-ARCTIC: A Non-Native English Speech Corpus</i> , Interspeech (2018) (“ <i>Sonsaat</i> ”)
Exhibit 1006	Linsen Loots and Thomas Niesler, <i>Automatic conversion between pronunciations of different English Accents</i> , Speech Communication 53 (2011) 75–84 (“ <i>Loots</i> ”)
Exhibit 1007	U.S. Patent Application No. 2013/0030789 to Dalce (“ <i>Dalce</i> ”)
Exhibit 1008	U.S. Patent Publication No. 2020/0193971 to Feinauer et al. (“ <i>Feinauer</i> ”)
Exhibit 1009	Guanlong Zhao and Ricardo Gutierrez-Osuna, <i>Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion</i> , IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 27, Issue: 10, October 2019) (“ <i>Osuna</i> ”)
Exhibit 1010	Declaration of June Munford (“ <i>Munford</i> ”)
Exhibit 1011	Priority Chinese Patent Publication No. 202010996501 to Fan et al. and Certified English Translation.
Exhibit 1012	Excerpts of the Prosecution History for the U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>FanHistory</i> ”)
Exhibit 1013	Arlie Coles, <i>Forced Phonetic Alignment Using Deep Neural Networks</i> , School of Computer Science, McGill University (April 16, 2018) (“ <i>Coles</i> ”)
Exhibit 1014	Micheal McAuliffe, Michaela Socolof, Sarah Mihuc, Micheal Wagner and Morgan Sonderegger, <i>Montreal Forced Aligner: trainable test-speech alignment using Kaldi</i> , Interspeech 2017, 498-502 (“ <i>McAuliffe</i> ”)
Exhibit 1015	Mehrayar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, <i>Foundations of Machine Learning second edition</i> , The MIT Press, 2018 (“ <i>Mohri</i> ”)
Exhibit 1016	Priority PCT 2022/057759 to Fan et al. and Certified English Translation
Exhibit 1017	Su-Hyun Han, Ko Woon Kim, SangYun Kim and Young Chul Youn, <i>Artificial Neural Network: Understanding the Basic Concepts without Mathematics</i> , Dement Neurocognitive Disord. 2018 Sep;17(3):83-89 (“ <i>Han</i> ”)
Exhibit 1018	Jason Brownlee, <i>Difference Between Algorithm and Model in Machine Learning</i> (August 19, 2020),

	https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/ (“Brownlee”)
Exhibit 1019	Maurya Vijayaramachandran, and Siddique Afraaz N, <i>Impact of Hidden Layer in Artificial Neural Networks</i> , IOSR Journal of Engineering, Vol. 10, Issue 11, November 2020, Series -I, 33-38 (“Siddique”)
Exhibit 1020	Maad M. Mijwel, Adam Esen and Aysar Shamil, <i>Overview of Neural Networks</i> , Babylonian Journal of Machine Learning, 1:2, April 2019 (“Mijwel”)
Exhibit 1021	Caroline Wiltshire, <i>Chapter 13 English Sounds in Context: The Pronunciation of Phonemes and Morphemes</i> , in <i>Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners</i> , 303-324, (2019) (“Wiltshire”)
Exhibit 1022	XiuSong Sun, Qun Yang, ShaoHan Liu and Xin Yuan, <i>Improving Low-Resource Speech Recognition Based on Improved NN-HMM Structures</i> , IEEE Access, vol. 8, pp. 73005-73014 (2020). (“Yang”)
Exhibit 1023	Nina Kazanina, Jeffery S. Bowers and William Idsardi, <i>Phonemes: Lexical access and beyond</i> , Psychon Bull Rev 25, 560–585, (2018) (“Kazanina”)
Exhibit 1024	Daniel Jurafsky and James H. Martin, <i>Speech and Language Processing, Chapter 27 Phonetics</i> , (2019) (“Jurafsky”)
Exhibit 1025	Christopher Schmandt, <i>Voice Communication with Computers Conversational Systems</i> , Van Nostrand Reinhold, (1994) (“Schmandt”)
Exhibit 1026	U.S. Patent No. 5,502,790 to Jie Yi (“Yi”)
Exhibit 1027	Tianze Shi, Shun Kasahara, Teeraphon Pongkittiphan, Nobuaki Minematsu, Daisuke Saito and Keikichi Hirose, <i>A Measure of Phonetic Similarity to Quantify Pronunciation Variation by Using ASR Technology</i> , International Congress of Phonetic Sciences, (2015) (“Shi”)
Exhibit 1028	Helene Francois and Olivier Boeffard, <i>Design of an optimal continuous speech database for text-to-speech synthesis considered as a Set Covering Problem</i> , Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 829-832 (“Francois”)
Exhibit 1029	Lifa Sun, Kun Li, Hao Wang, Shiyin Kang and Helen Meng, <i>Phonetic Posteriorgrams for Many-to-One Voice Conversion Without Parallel Data Training</i> , 2016 IEEE International

	Conference on Multimedia and Expo (ICME), pp. 1-6, (2016) (“Sun”)
Exhibit 1030	Guanlong Zhao, Shaojin Ding and Ricardo Gutierrez-Osuna, <i>Converting Foreign Accent Speech Without a Reference</i> , IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2367-2381, (2021) (“Zhao”)
Exhibit 1031	Berrak Sisman, Mingyang Zhang and Haizhou Li, <i>Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion</i> , IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27, (2019) (“Sisman”)
Exhibit 1032	David Crystal, <i>A Dictionary of Linguistics and Phonetics, Sixth Edition</i> , Blackwell Publishing, (2008) (“Crystal”)
Exhibit 1033	Walter Wolfram, <i>Everyone Has an Accent</i> , Southern Poverty Law Center (2000), https://archive.learningforjustice.org/magazine/fall-2000/everyone-has-an-accent (“Wolfram”)
Exhibit 1034	U.S. Patent Application Publication No. 2022/0122579 to Biadsy et al. (“Biadsy”)
Exhibit 1035	Shaojin Ding, Guanlong Zhao and Ricardo Gutierrez-Osuna, <i>Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning</i> , Journal of LATEX Templates, (2021) (“Accentron”)
Exhibit 1036	Leland Roberts, <i>Understanding the Mel Spectrogram</i> , Medium, (2020), https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53 (“Roberts”)
Exhibit 1037	Eleanor Chodroff, <i>Corpus Phonetics Tutorial</i> , arXiv 1811.05553, (2018) (“Chodroff”)
Exhibit 1038	Amrianto Saragih, <i>4 Methods to Boost the Accuracy of a Neural Network Model</i> , Medium. (2019), https://medium.com/@amrianto.saragih/4-methods-to-boost-the-accuracy-of-a-neural-network-model-bb694e650a69 (“Saragih”)
Exhibit 1039	Wei Zhang, <i>Alternation of [n] and [l] in Sichuanese Dialect, Standard Mandarin and English: A Single-Case Study</i> , Leeds Working Papers in Linguistics and Phonetics, Issue 12, (2007) (“Zhang”)
Exhibit 1040	G. Ayorkor Mills-Tettey and David Kotz, <i>Mobile Voice Over IP (MVOIP): An Application-level Protocol for Call Hand-off in Real Time Applications</i> , Dartmouth Scholarship, 3453, (2002) (“Ayorkor”)

Exhibit 1041	Noriko Umeda, <i>Consonant duration in American English</i> , J. Acoust. Soc. Am. 61, 846–858 (1977) (“ <i>Umeda</i> ”)
Exhibit 1042	Noriko Umeda, <i>Vowel duration in American English</i> , J. Acoust. Soc. Am. 58, 434–445 (1975) (“ <i>UmedaVowels</i> ”)
Exhibit 1043	Songxiang Liu, Disong Wang, Yuewen Cao, Lifa Sun, Xixin Wu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, <i>End-To-End Accent Conversion Without Using Native Utterances</i> , https://www1.se.cuhk.edu.hk/~hccl/publications/pub/ICASSP2020_e2eAC%20(1).pdf (2020) (“ <i>Liu</i> ”)
Exhibit 1044	Declaration of June Munford Regarding Sonsaat (“ <i>Munford-Sonsaat</i> ”)
Exhibit 1045	Declaration of Mina Ching (“ <i>Ching</i> ”)

14. I have considered these materials from the viewpoint of a PHOSITA as of the priority date of the ’550 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the ’550 Patent is May 6, 2021. I note that my opinions provided in this Declaration are made from the perspective of a PHOSITA as of this priority date of the ’550 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 ’550 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 ’550 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 ordinary and customary 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. I have been instructed to construe *“fourth audio data representative of”* as **“synthesized acoustic feature data generated as an output of a speech conversion process and used as input to a subsequent conversion process that produces output waveform audio.”**

IV. LEVEL OF A PERSON OF ORDINARY SKILL

27. Based on my review and analysis of the '550 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a PHOSITA in the field of the '550 Patent at the time of the earliest possible priority date (May 6, 2021) 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 May 6, 2021, 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. *See, e.g.,* the '550 Patent (EX1001) at 1:14-2:46 (describing the “Background” and “Overview”). A person with less relevant practical experience but with additional education may also qualify as a PHOSITA in the field of the '550 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 May 6, 2021.

V. OVERVIEW OF TECHNOLOGY

28. I was asked to briefly summarize the background of the technology from the standpoint of a PHOSITA prior to May 6, 2021.

A. Machine Learning and Neural Networks

29. Machine learning was a well-known concept in the art. Machine learning includes computational methods that utilize experience, such as collected data, to improve performance or make accurate predictions, and consists of designing prediction algorithms. EX1015-Mohri, 14. Input and output data is provided, and the machine determines the process by which the given input produces the given output data. EX1017-Han, 1. It was known that machine learning was used in speech processing applications, including “speech recognition, speech synthesis,

speaker verification, speaker identification, as well as sub-problems such as language modeling and acoustic modeling.” EX1015-Mohri, 15.

30. Machine learning involves procedural computations including algorithms, which utilizes training set data to define or otherwise establish a machine learning model, in order to learn from data and to perform pattern recognition. EX1018-Brownlee, 2. That is, a machine learning model is the output of training a machine learning algorithm with data. *Id.* Thus, a machine learning model consists of algorithm-specific data structures that allow the model to make predictions. *Id.*

31. An artificial neural network is a machine learning algorithm inspired by the connections of human neurons. EX1017-Han, 1. A neural network seeks to recognize the underlying relationship of a set of data with a series of algorithms to simulate the working of a human brain. EX1019-Siddique, 1. Accordingly, neural networks are therefore “capable of learning and can be trained to produce desired results.” EX1019-Siddique, 2. For instance, the learning process of an artificial neural network includes “updating the connection strength [] of a node” via error minimization between a predicted value and a correct value. EX1017-Han, 6. Specifically, each connection between nodes in the neural network has a weight that controls the signal between those nodes. EX1019-Siddique, 2. In training, the weights are altered so that the model provides more accurate results. EX1019-Siddique, 2; EX1017-Han, 3-4. Typical applications of neural networks (and, thus,

their learning processes) include voice recognition and speech synthesis. EX1020-Mijwel, 3. For example, *Liu* teaches an end-to-end accent conversion approach “composed of four independently trained neural networks[.]” EX1043-Liu, 2; *see generally* EX1043-Liu.

B. Phonemes

32. A phoneme is the smallest unit of sound that distinguishes meaning in a language. EX1021-Wiltshire, 304. *Improving Low-Resource Speech Recognition Based on Improved NN-HMM Structures* to Yang et al. (“*Yang*”) teaches that phonemes are also known as phones. EX1022-Yang, 2. *Phonemes: Lexical access and beyond* to Kazanina et al. (“*Kazanina*”) teaches that phonemes have two essential properties: “they are ‘segment-sized’ (the size of a consonant or vowel) and abstract (a single phoneme may have different acoustic realizations).” EX1023-Kazanina, 1. *English Sounds in Context: The Pronunciation of Phonemes and Morphemes* to Wiltshire (“*Wiltshire*”) teaches that changing one phoneme changes the meaning of a word. EX1021-Wiltshire, 304. 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.” EX1021-Wiltshire, 304. Such pairs of words are called “minimal pairs”, because they only differ by a single sound but have completely different meanings. EX1021-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]”. EX1021-Wiltshire, 304.

33. 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’).” EX1023-Kazanina, 2. 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/).” EX1023-Kazanina, 2.

34. 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. EX1024-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.” EX1025-Schmandt, 40. ARPAbet was developed “for the convenience of computer manipulation and representation of speech using ASCII-printable characters.” EX1025-Schmandt, 40.

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]	[ɔɪ]	soil	[s oy l]

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

EX1024-Jurafsky, 2.

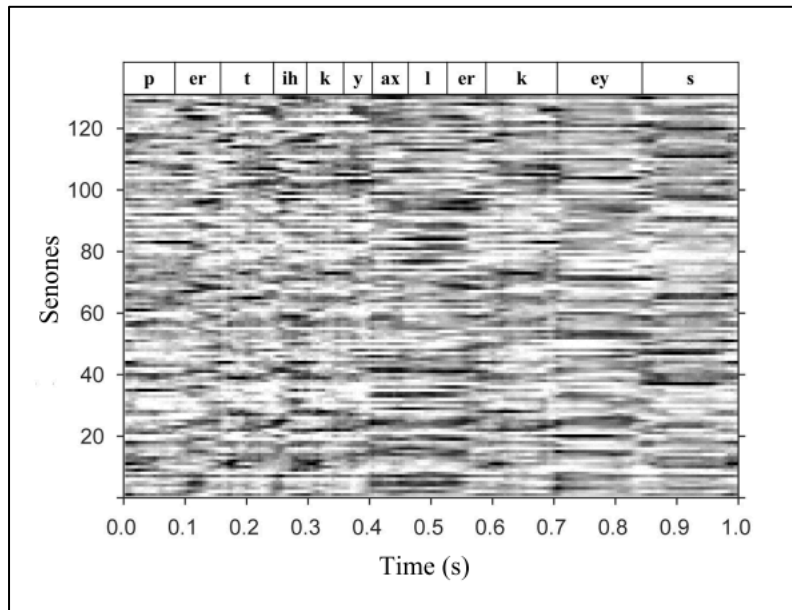
35. For a given language the length of a phoneme varies depending on the speaker and the position of the phoneme in a words. *Consonant Duration in American English* to Umeda (“*Umeda*”) and *Vowel Duration in American English* to Umeda (“*UmedaVowels*”) each provide exemplary phoneme duration for American English. See generally EX1041-Umeda; see generally EX1042-UmedaVowels.

36. Yang teaches that a monophone refers to a single phone, while a triphone is “a group of 3 phones in the form “L – X + R”, where the “L” phone (i.e. the left-hand phone) precedes “X” phone and the “R” phone (i.e. the right-hand phone) follows it.” EX1022-Yang, 2, 3. Similarly, U.S. Patent No. 5,502,790 to Yi

(“Yi”) teaches that triphones are “three consecutive phonemes,” and diphones are “two consecutive phonemes.” EX1026-Yi, 1:52-55. It was known that a triphoneme is synonymous with a triphone. In *A MEASURE OF PHONETIC SIMILARITY TO QUANTIFY PRONUNCIATION VARIATION BY USING ASR TECHNOLOGY* to Shi et al. (“Shi”), Shi describes a type of HMM model known as triphone models in the ASR community that is also referred to as triphoneme models. EX1027-Shi, 3. It was known that a triphoneme is essentially identical to a triphone. A triphoneme is “a sequence of three phonemes and presented as the phonetic description of an allophone,” as taught by *Design of an optimal continuous speech database for text-to-speech synthesis considered as a Set Covering Problem* to Francois et al. (“Francois”). EX1028-Francois, 1.

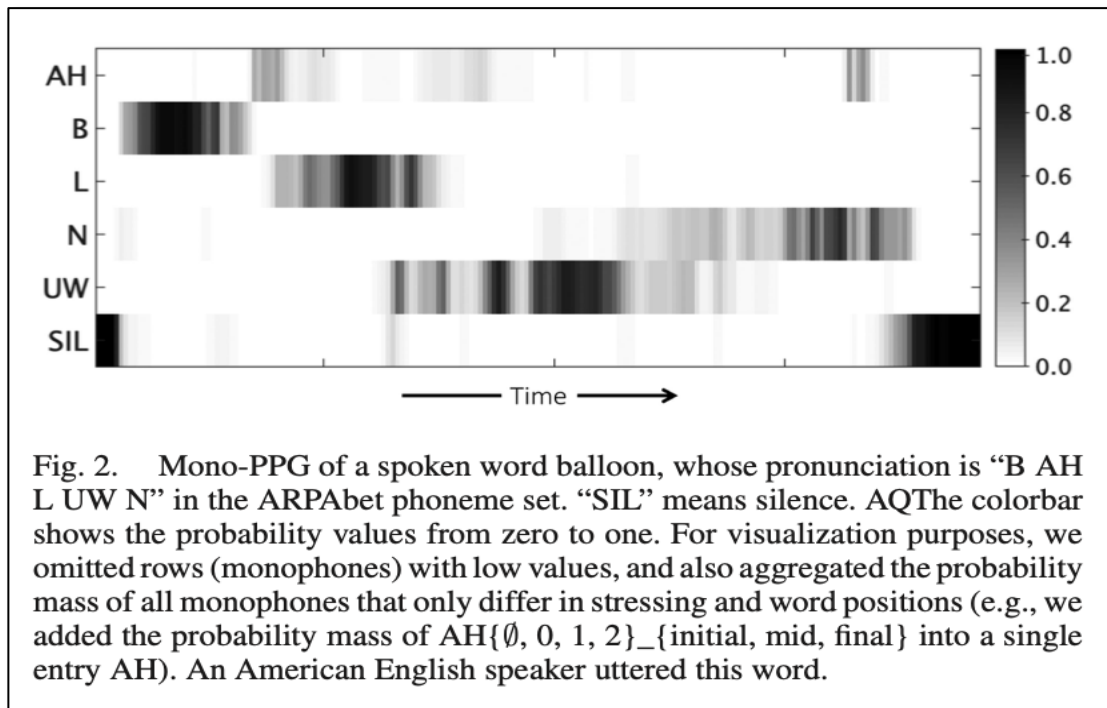
C. Phonetic Posteriorgrams (PPGs)

37. 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. EX1029-Sun, 2-3. “A phonetic class may refer to a word, a phone or a senone.” EX1029-Sun, 3. An example of a PPG (using senones) is as follows:



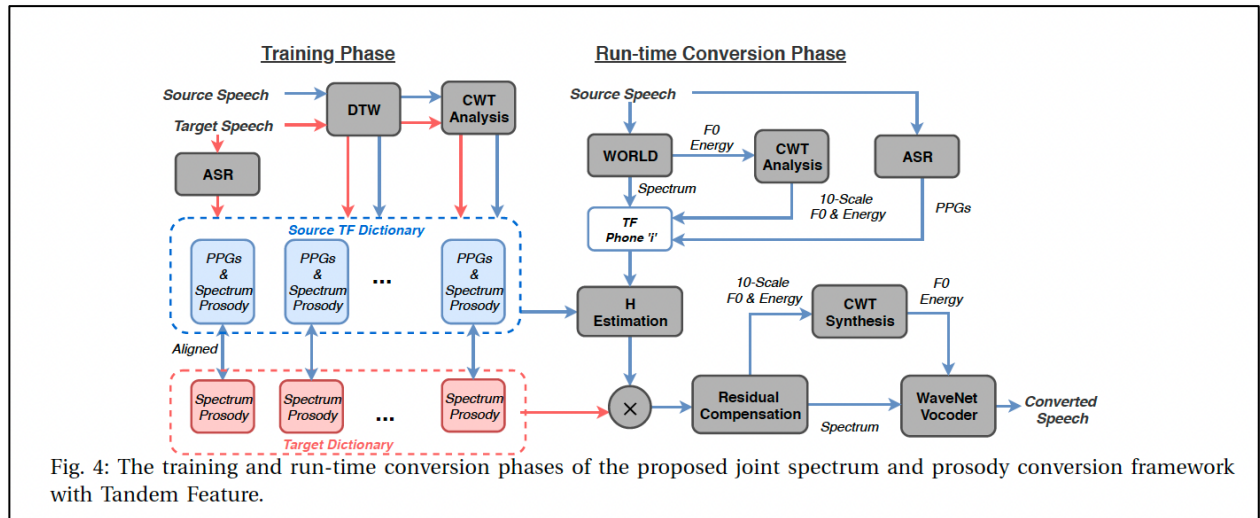
EX1029-Sun, Fig. 4. In this example, the “horizontal axis represents time in seconds and the vertical one contains indices of phonetic classes.” EX1029-Sun, 3. A “[d]arker shade implies a higher posterior probability.” EX1029-Sun, 3.

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



EX1030-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.” EX1030-Zhao, 4.

39. As another example, EX1031, *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.” EX1031-Sisman, 7. *Sisman* teaches “a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary.” EX1031-Sisman, 7.



EX1031-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 CWTbased prosody features, called source TF dictionary while the target phonetic dictionary B_i only includes spectrum features and prosody features, called target dictionary.” EX1031-Sisman, 8.

D. Prosodic Features

40. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches that prosody “refers to the spoken style of discourse independent of lexical content,” and “encompasses the majority of information lost when comparing an utterance to its transcription, and much of the richness of speech is conveyed by exactly this nonlexical means of expression.” EX1025-Schmandt, 325-326.

Schmandt further teaches that intonation, stress, pitch, phrasing, and duration are all aspects of prosody. *Id.*

41. Specifically, *Schmandt* teaches that “intonation is the tune of an utterance,” and “helps to convey the stress of words and syllables within a sentence as stressed syllables are spoken with a pitch higher or lower than normal.” EX1025-*Schmandt*, 325-326. *Schmandt* also teaches that “phrasing is the breaking of speech into groups, ... how we may speak a few words much more slowly for emphasis.” *Id.* *Schmandt* teaches that “the duration of individual syllables, stressed syllables being generally longer than unstressed ones,” results “in a meter scheme across an entire phrase.” *Id.*

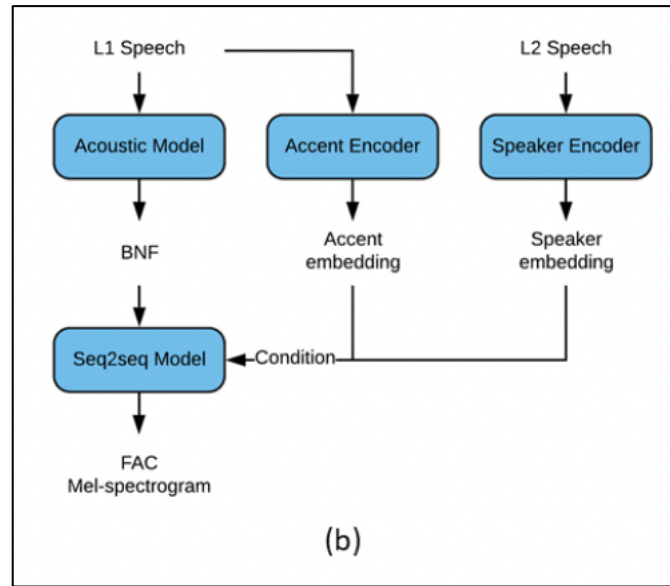
E. Dialect and Accent

42. Each individual speaker has their own unique linguistic system, one’s personal dialect, and *Crystal* teaches this to be the linguistic term idiolect. EX1032-*Crystal*, 261. “A dialect can be seen as an abstraction deriving from the analysis of a large number of idiolects.” *Id.* *Everyone Has an Accent* to Wolfram (“*Wolfram*”) aligns with *Crystal* and teaches that “[l]inguists use the term *dialect* to denote patterns in the way people use language. These patterns include pronunciation (or “accent”), vocabulary and grammatical structures that reflect the user's cultural and regional background.” EX1033-Wolfram, 2.

F. Prior Art Accent Conversion Systems

43. 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.” EX1034-Biadsy, [0023]; *see also* [0021].

44. 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) [.]” EX1035-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.” EX1035-Accentron, 1.



EX1035-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.” EX1035-Accentron, 1.

1. Encoder and Decoder

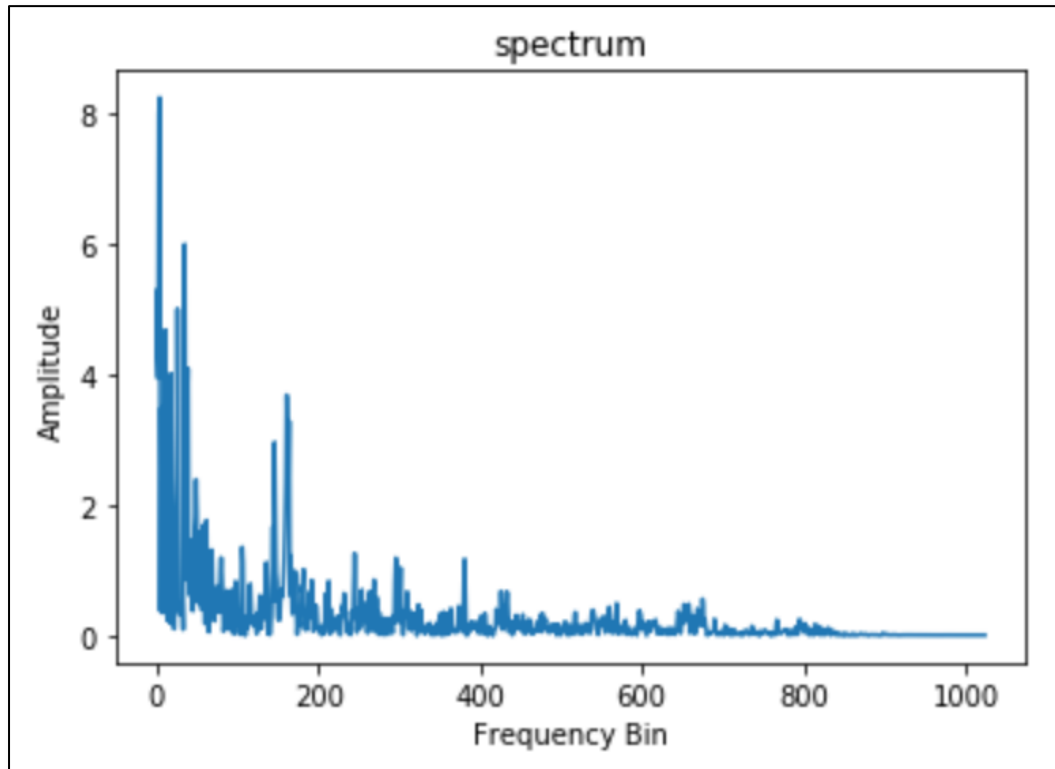
45. In the context of a neural network, an encoder takes as input the data or signal to be analyzed, such as spectrograms in the case of speech processing, and converts it into a higher dimensional vector suitable for processing within the network; these encodings often have little meaning outside the context of the network which operates on them. Similarly a decoder takes the result of processing

the input within the neural network, as a similarly high dimensional vector representation, and converts it to a more meaningful output, generally in the same form as the original input to the encoder.

46. *Zhao* teaches one example of “an encoder-decoder architecture.” EX1030-Zhao, 5. This encoder-decoder architecture uses an encoder to “consume” input sequences and generate a high-level hidden representation sequence. EX1030-Zhao, 5. 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.” EX1030-Zhao, 5.

2. Mel Spectrograms

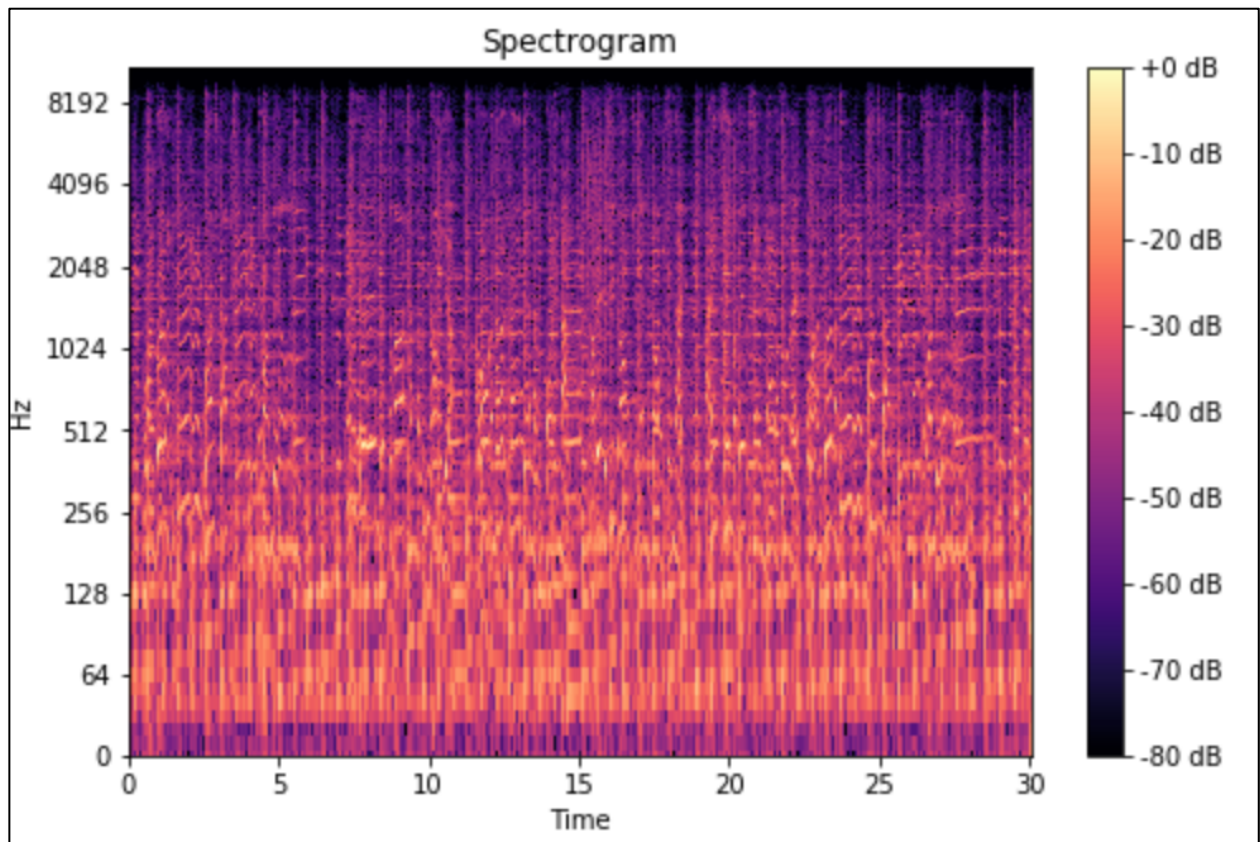
47. An audio signal is “comprised of several single-frequency sound waves...The Fourier transform is a mathematical formula that allows us to decompose a signal into its individual frequencies and the frequency’s amplitude. In other words, it converts the signal from the time domain into the frequency domain. The result is called a spectrum.” EX1036-Roberts, 3.



EX1036-Roberts, 4.

48. 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.” EX1036-Roberts, 4-6. *Schmandt* teaches a “spectrogram is a transformation of the waveform into the frequency domain.” EX1025-Schmandt, 37. 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

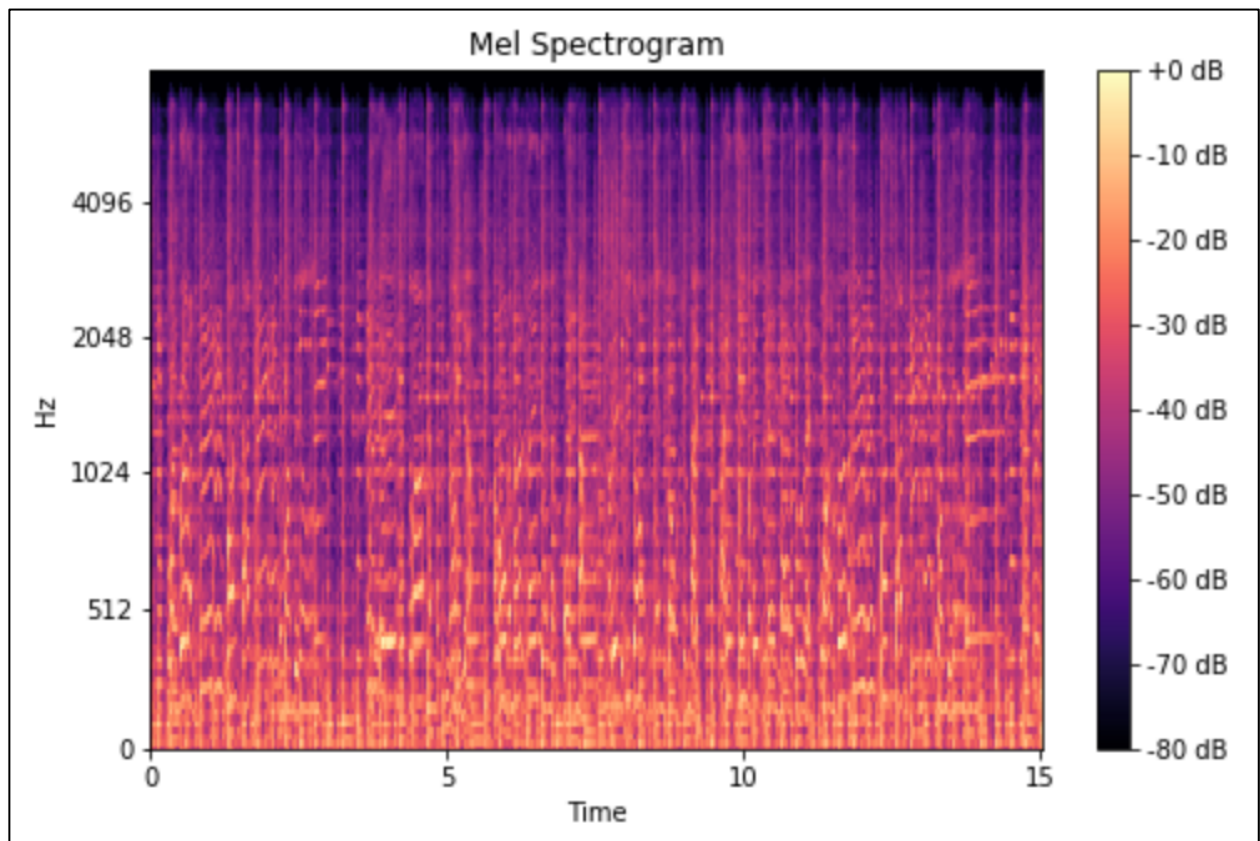
that frequency or time.” EX1025-Schmandt, 37. A spectrogram “requires much less storage than the original waveform and reveals acoustic features specific to speech.” EX1025-Schmandt, 37.



EX1036-Roberts, 6.

49. 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.” EX1036-Roberts, 7. To address 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.” EX1036-Roberts, 7.

From the spectrogram in the prior paragraph, you “perform a mathematical operation on frequencies to convert them to the mel scale.” EX1036-Roberts, 7. A Mel-spectrogram is “is a spectrogram where the frequencies are converted to the mel scale.” EX1036-Roberts, 8.



EX1036-Roberts, 8.

3. Vocoder

50. A vocoder converts a computed representation of speech, such as a spectrogram or LPC coefficients, into an audible speech waveform. *Zhao* teaches that to obtain hearable synthesized audio, mel spectrograms are transformed into waveforms using a real-time vocoder. EX1030-Zhao, 6 (“The predicted Mel-

spectrograms are converted back to audio waveforms using a WaveGlow neural vocoder trained on the L2 utterances[.]”); EX1030-Zhao, 2. (“we use a high-quality WaveGlow [15] real-time neural vocoder to convert Mel-spectrograms to speech waveform.”).

51. Another type of vocoder is a “WaveNet vocoder [21] to generate the target speech waveforms.” EX1031-Sisman, 9.

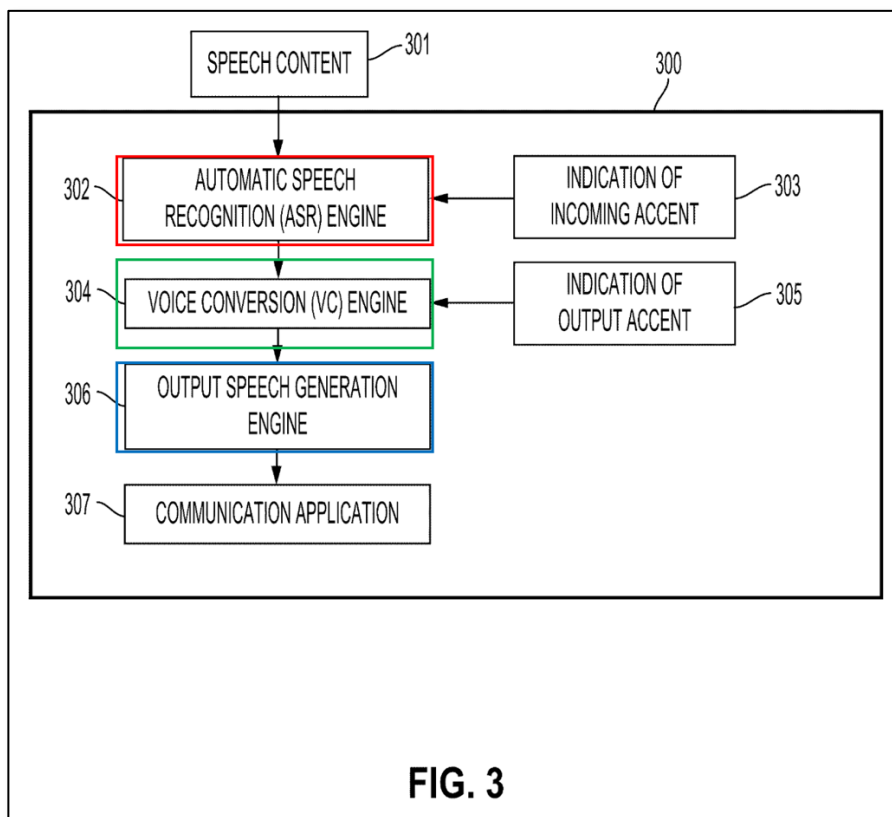
VI. OVERVIEW OF THE ’550 PATENT

52. I have reviewed the ’550 Patent (EX1001). Descriptions of the ’550 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the ’550 Patent are intended only as a summary of the ’550 Patent and/or intended as being pertinent to the Challenged Claims of the ’550 Patent. My opinions provided herein are not intended as a discussion of my entire understanding of the ’550 Patent or any other claims in any related patent to the ’550 Patent.

A. Description of the ’550 Patent

53. The ’550 Patent purports to describe systems and methods for real-time accent conversion using machine learning. EX1001, 2:4-14. The alleged invention provides a machine-learning framework that changes speech from one accent to another. EX1001, 1:60-66, 2:4-14, 8:16-21. To achieve this goal, the alleged invention utilizes a multi-stage machine learning architecture comprising an

automatic speech recognition (ASR) engine, a voice conversion (VC) engine, and an output speech generation engine. EX1001, 5:45-50, claim 1, Fig. 3.



EX1001, Fig. 3.¹

¹ Claim language is *italicized*. All drawing annotations added. Claim language in colored font maps to same-colored regions in the annotated drawings. Citations to a claim limitation and/or section number are to the limitation's mapping in this Declaration and are incorporated by reference, including the Motivations to Combine (MTC) and showing of reasonable expectation of success (REOS) provided in the cross-referenced section. All REOS include the benefits described for the respective MTC. All emphases added unless otherwise noted.

B. Field of Endeavor of the '550 Patent

54. The '550 Patent describes “[t]echniques for real-time accent conversion.” EX1001, Abstract. The '550 Patent discloses techniques such as “software technology that utilizes machine-learning models to” convert speech of a first accent to a second accent, “a method that involves a computing device” to convert speech of a first accent to a second accent. EX1001, 2:4-10, 2:15-30. After reviewing the '550 Patent, it is my opinion that a PHOSITA would understand that the field of endeavor of the '550 Patent includes speech audio processing methods, devices, and systems.

C. Problem Purportedly Solved by the Inventors of the '550 Patent

55. I have been asked to analyze the '550 Patent and determine the problem that the inventors were trying to solve. The '550 Patent purportedly solves “problems with existing solutions for performing accent conversion.” EX1001, 2:4-14.

56. One such problem the '550 Patent purportedly solves is that existing solutions do not account for the phonetic differences between accents, “and therefore these aspects of the accent remain in the output speech.” EX1001, 1:49-59, 7:55-64. Another problem the '550 Patent addresses is that existing solutions cannot capture speaker characteristics and nuances, “such as the prosody or emotion of the speaker[,]” in accent conversion. EX1001, 1:60-66, 8:16-21. The '550 Patent notes that existing solutions are “impractical for use in real-time communication scenarios

such as an ongoing conversation (e.g., a phone call).” EX1001, Title, Abstract, 2:4-14, 5:24-30.

D. Prosecution History of the ’550 Patent

57. I have reviewed the prosecution history of the ’550 Patent. The ’550 Patent claims priority to U.S. Provisional Application No. 63/185,345, filed on May 6, 2021. Because the ’550 Patent’s earliest possible priority date is after March 16, 2013, Petitioner applies AIA versions of §§ 102, 103. For this Proceeding only, Petitioner applies the earliest possible May 6, 2021 priority date but does not concede that any of the Challenged Claims are entitled to such date.

58. The Challenged Claims were allowed because the prior art of record did not teach limitations “such as the training and application of the machine learning model and the synthesizing of fourth audio data using the method described in the claims.” EX1002, 660.

VII. OVERVIEW OF THE PRIOR ART

59. The prior art is analogous to the claimed invention of the ’550 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 ’550 Patent, namely speech audio processing. EX1001, Abstract, 2:4-10, 2:15-30. Moreover, a PHOSITA would have found the prior art to be of the same field of endeavor and thus analogous to the ’550 Patent.

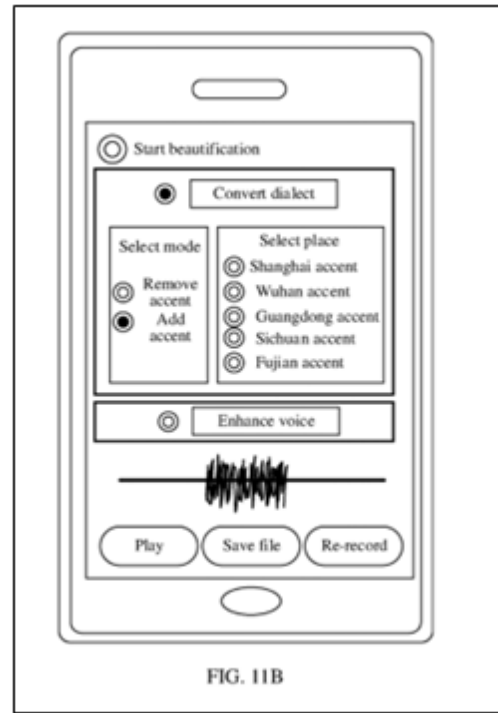
60. The prior art is reasonably pertinent to at least one problem facing the '550 Patent inventors, namely at least one of: performing accent conversion in real-time, capturing speaker characteristics and nuances in accent conversion, accounting for phonetic differences between accents in accent conversion, as described further below with respect to each denoted prior art reference.

A. Overview of Fan

61. U.S. Patent Appl. No. 2023/0223006 to Fan et al. ("*Fan*")² was filed on March 20, 2023, published July 13, 2023, and is a continuation of PCT/CN2021/117945 (EX1016), which was filed on September 13, 2021, and claims priority to Chinese application 202010996501.5 (EX1011) filed September 21, 2020. *See* EX1012-FanHistory. The US application incorporates both the PCT application (EX1016) and the Chinese application (EX1011) by reference. The PCT application (EX1016) incorporates the Chinese application (EX1011) by reference.

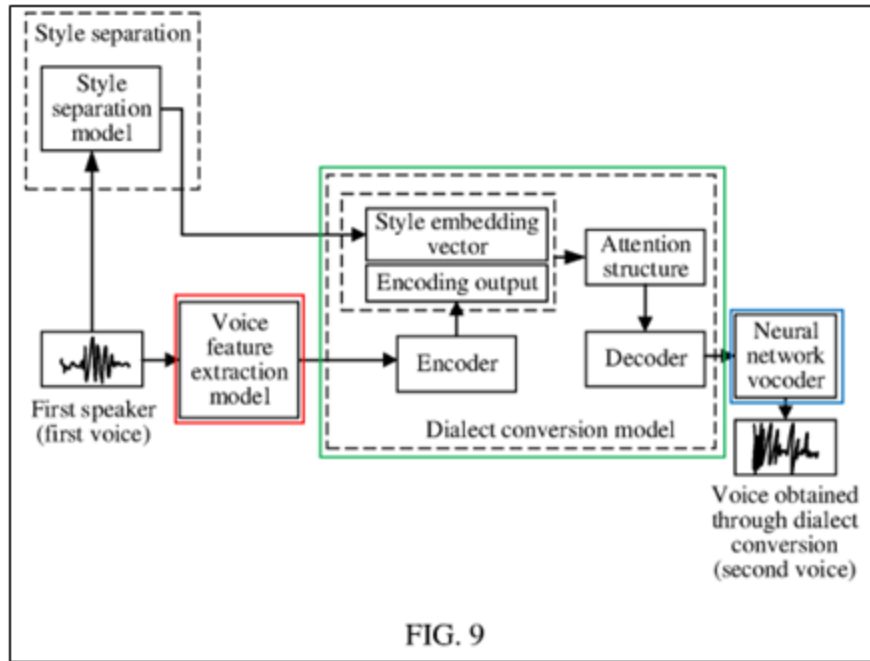
² EX1004 is the U.S. Patent Appl. to *Fan*. Also included is the Chinese priority application (EX1011, 64), the English translation of the Chinese priority application (EX1011, 1), and a certification that the translation is true and correct (EX1011, 63). I have been informed that *Fan* made a foreign priority claim within 12 months of the foreign priority application (EX1016, 3-4) and the Office acknowledged same (EX1016, 10).

62. *Fan* teaches “a dialect conversion mode, for adding an accent to or removing an accent from the first voice[.]” EX1004-Fan, Abstract.



EX1004-Fan, Fig. 11B.

63. The dialect conversion mode uses a voice feature extraction model to extract a PPG feature from a first voice (input voice) having a first accent and a style separation model to extract a timbre feature from the input voice. EX1004-Fan, [0113], [0127], Fig. 9. The PPG feature and timbre feature are provided to a dialect conversion model to produce a mel feature representing the style and content of the input voice but with a second accent. EX1004-Fan, [0126]-[0127]. The mel feature is provided to a neural network vocoder and is converted to a voice signal so that it can be played. EX1004-Fan, [0099], Fig. 9.



EX1004-Fan, Fig. 9. The voice feature extraction model and dialect conversion model are trained neural networks. EX1004-Fan, [0085], [0126], [0140]-[0141].

64. A PHOSITA would have identified *Fan* as being within the same field of endeavor as the '550 Patent because both *Fan* and the '550 Patent relate to speech audio processing and accent conversion. *Compare* EX1004-Fan, Abstract, [0002], [0134], [0201], *with* EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Fan* discloses a system having the capacity for “speaking style conversion on a to-be-converted first voice...[and] a dialect conversion mode.” EX1004-Fan, Abstract. This dialect conversion mode involves the addition of removal of an accent to a voice. EX1004-Fan, [0056]. Specifically, *Fan* discloses a dialect conversion network that includes a plurality of dialect conversion models “for performing

dialect conversion on an input first voice” dedicated to removing an accent from a first voice and adding an accent to a to-be-converted voice. EX1004-Fan, [0119]-[0020]. For example, *Fan* describes converting a Mandarin accent into a Sichuanese accent by first removing the Sichuanese accent. EX1004-Fan, [0020]. *Fan*’s system utilizes an interface for the selection of voice processing mode, at which a user may select category labels including “Remove accent” and “Add accent” to dictate accent conversion and to generate a “second voice ... obtained after dialect conversion.” EX1004-Fan, [0135], [0201].

65. A PHOSITA would have found *Fan* reasonably pertinent to the problem faced by the inventors of the ’550 Patent because both *Fan* and the ’550 Patent describe an improvement upon existing solutions, that is, capturing speaker characteristics and nuances in accent conversion. *Compare* EX1004-Fan, [0009], [0127], [0132], *with* EX1001, 1:60-66, 8:16-21. For example, *Fan* describes separating “a style feature of a first voice by using the style separation model, to obtain a timbre feature vector of the first voice.” EX1004-Fan, [0127]. A PHOSITA would have understood *Fan*’s style separation model to capture speaker characteristics and nuances in accent conversion.

66. I have compared the translations of both priority PCT and Chinese applications to *Fan* and determined that the disclosures of *Fan* relied on in this

Declaration and the related Petition are fully supported by both the PCT and Chinese applications in the manner set forth below:

<i>Fan</i> Publication (EX1006) (figure or paragraph)	<i>Fan</i> PCT (EX1016) (page, figure or paragraph)	<i>Fan</i> Priority (EX1011) (page ³ , figure or paragraph)
Abstract	47, Abstract and cover page at (57)	39, Abstract
Fig. 3	49, Fig. 3	41, Fig. 3
Fig. 7	54, Fig. 7	46, Fig. 7
Fig. 8	54, Fig. 8	46, Fig. 8
Fig. 9	55, Fig. 9	47, Fig. 9
Fig. 10	56, Fig. 10	48, Fig. 10
Fig. 11A	57, Fig. 11A	49, Fig. 11A
Fig. 11B	58, Fig. 11B	50, Fig. 11B
Fig. 12	59, Fig. 12	51, Fig. 12
Fig. 15	61, Fig. 15	53, Fig. 15
[0002]	1, ¶2	1, ¶1
[0003]	1, ¶3	1, ¶2
[0005]	1, ¶5	1, ¶4
[0006]	2, ¶1	2, ¶1
[0007]	2, ¶2	2, ¶2
[0016]	4, ¶5	4, ¶5
[0017]	5, ¶1	4, ¶6
[0018]	5, ¶2	5, ¶1
[0024]	6, ¶3	6, ¶3
[0036]	7, ¶10	7, ¶8
[0040]	8, ¶1	7, ¶12
[0043]	8, ¶4	8, ¶2
[0045]	8, ¶6	8, ¶4
[0049]	9, ¶4	9, ¶4
[0050]	10, ¶1	9, ¶5
[0051]	10, ¶2	10, ¶1
[0052]	10, ¶3	10, ¶2
[0053]	10, ¶4	10, ¶3
[0054]	10, ¶5	10, ¶4
[0056]	10, ¶7	10, ¶6
[0058]	11, ¶2	11, ¶2
[0063]	11, ¶7	11, ¶7
[0067]	12, ¶4	12, ¶¶3-4
[0070]	12, ¶7	12, ¶7
[0085]	14, ¶5	14, ¶5
[0086]	15, ¶1	15, ¶1; 2, ¶2
[0088]	15, ¶3	15, ¶3
[0091]	16, ¶1	15, ¶6
[0092]	16, ¶2	16, ¶1
[0093]	16, ¶3	16, ¶2
[0094]	16, ¶4	16, ¶3
[0099]	18, ¶2	18, ¶2
[0110]	20, ¶4	20, ¶3

[0111]	20, ¶5	20, ¶4
[0012]	20, ¶6	20, ¶5
[0113]	21, ¶1	20, ¶6
[0014]	21, ¶2	21, ¶1
[0015]	21, ¶3	21, ¶2
[0116]	21, ¶4	21, ¶3
[0118]	22, ¶2	22, ¶1
[0119]	22, ¶3	22, ¶2
[0120]	22, ¶4	22, ¶3
[0121]	22, ¶5	22, ¶4
[0122]	22, ¶6	22, ¶5
[0123]	23, ¶1	22, ¶6
[0124]	23, ¶2	22, ¶7
[0125]	23, ¶3	23, ¶1
[0126]	23, ¶4	23, ¶2
[0127]	23, ¶5	23, ¶3
[0128]	24, ¶1	23, ¶4
[0129]	24, ¶2	24, ¶1
[0130]	24, ¶3	24, ¶2
[0131]	24, ¶4	24, ¶3
[0132]	24, ¶5	24, ¶4
[0133]	25, ¶1	24, ¶5
[0134]	25, ¶2	25, ¶1
[0135]	25, ¶3	25, ¶2
[0136]	25, ¶4	25, ¶3
[0137]	25, ¶5	25, ¶4
[0138]	25, ¶6	25, ¶5
[0139]	26, ¶1	26, ¶1
[0140]	26, ¶2	26, ¶2
[0141]	26, ¶3	26, ¶3
[0142]	27, ¶1	27, ¶1
[0144]	27, ¶3	27, ¶3

³ EX1011 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 EX1011 in this chart and in this Declaration refer to the page numbers assigned to the specification. All page citations to EX1016 in this chart and in this Declaration refer to the page numbers assigned to the specification.

[0148]	28, ¶2	27, ¶7
[0150]	28, ¶4	28, ¶2
[0162]	29, ¶6	29, ¶4
[0163]	29, ¶7	29, ¶5
[0165]	30, ¶1	29, ¶7
[0167]	30, ¶3	30, ¶1
[0173]	30, ¶9	30, ¶7
[0174]	30, ¶10	30, ¶8
[0189]	32, ¶8	32, ¶6
[0191]	33, ¶2	33, ¶1
[0192]	33, ¶3	33, ¶2
[0193]	33, ¶4	33, ¶3
[0194]	33, ¶5	33, ¶4
[0195]	33, ¶6	33, ¶5
[0200]	35, ¶2	35, ¶2
[0201]	35, ¶3	35, ¶3
[0203]	35, ¶5	35, ¶5
[0206]	36, ¶3	36, ¶2
[0211]	37, ¶1	36, ¶7

B. Overview of Sonsaat

67. “*L2-ARCTIC: A Non-Native English Speech Corpus*” to Zhao et al. (“*Sonsaat*”) was publicly available at least as early as September 20, 2019. EX1045-Ching, 1-2, 6-10, 42-47 (declaration of Mina Ching, Records Request Processor at the Internet Archive); EX1043-Munford-Sonsaat, ¶¶15-21 (declaration of June Munford establishing public availability of *Sonsaat*).

68. *Sonsaat* teaches a dataset and process used to create “L2-ARCTIC, a speech corpus of non-native English that is intended for research in voice conversion, accent conversion, and mispronunciation detection.” EX1005-Sonsaat, 1.

69. A PHOSITA would have identified *Sonsaat* as being within the same field of endeavor as the '550 Patent because both *Sonsaat* and the '550 Patent relate to speech audio processing and accent conversion. *Compare* EX1005-Sonsaat, 1, 4, with EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Sonsaat* discloses “L2-ARCTIC, a speech corpus ... that is intended for research in ... accent conversion[,]” that is “a valuable resource ... for research in ... accent conversion.” EX1005-Sonsaat,1.

70. A PHOSITA would have found *Sonsaat* reasonably pertinent to the problem faced by the inventors of the '550 Patent because both *Sonsaat* and the '550 Patent describe an improvement upon existing solutions, that is, accounting for phonetic differences between accents in accent conversion. *Compare* EX1005-Sonsaat, Abstract, 3, 4, Fig. 3, with EX1001, 1:49-59, 7:55-64.

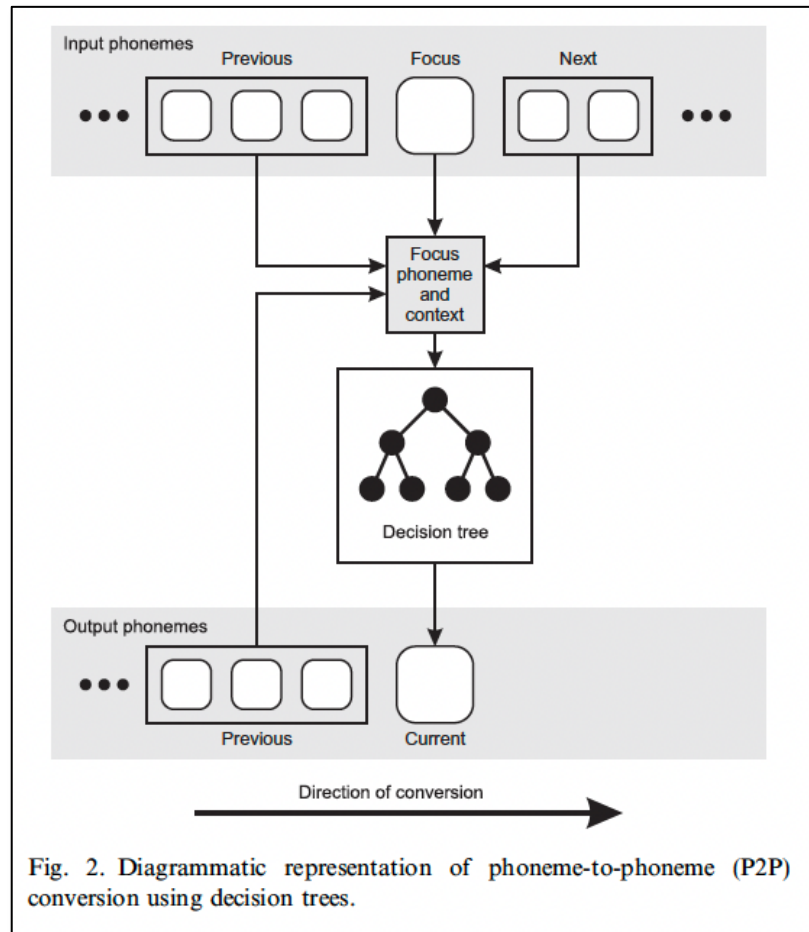
C. Overview of Loots

71. “*Automatic conversion between pronunciations of different English accents*” to Loots et al. (“*Loots*”) was publicly available at least as early as January 2011. EX1010-Munford, ¶¶40-43 (declaration of June Munford establishing public availability of *Loots*).

72. *Loots* teaches “phoneme-to-phoneme (P2P) conversion using decision trees” to align two pronunciations. EX1006-Loots, 6. P2P is “the conversion of one

phoneme string (the source pronunciation) to another (the target pronunciation)[.]”

EX1006-Loots, 6.



EX1006-Loots, Fig. 2.

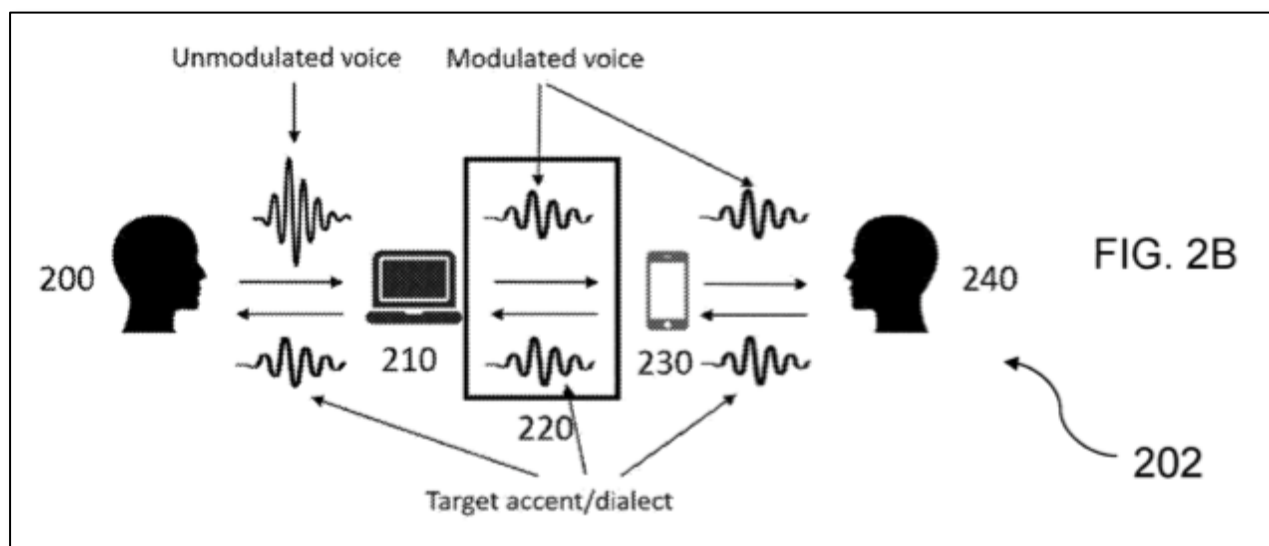
73. A PHOSITA would have identified *Loots* as being within the same field of endeavor as the '550 Patent because both *Loots* and the '550 Patent relate to speech audio processing and accent conversion. Compare EX1006-Loots, 6, Fig. 2, with EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Loots* describes “the

application of decision trees to the automatic conversion of pronunciations between American, British and South African English accents.” EX1006-Loots, Abstract.

74. A PHOSITA would have found *Loots* reasonably pertinent to the problem faced by the inventors of the '550 Patent because both *Loots* and the '550 Patent describe an improvement upon existing solutions, that is, accounting for phonetic differences between accents in accent conversion. *Compare* EX1006-Loots, Abstract, 3, 6, *with* EX1001, 1:49-59, 7:55-64.

D. Overview of Feinauer

75. 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[.]” EX1008-Feinauer, Abstract.



EX1008-Feinauer, Fig. 2B.

76. A PHOSITA would have identified *Feinauer* as being within the same field of endeavor as the '550 Patent because both *Feinauer* and the '550 Patent relate to speech audio processing and accent conversion. *Compare* EX1008-Feinauer, Abstract, [0001], [0091], *with* EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Feinauer* describes “[s]ystems and methods for accent and dialect modification[.]” EX1008-Feinauer, Abstract.

77. A PHOSITA would have found *Feinauer* reasonably pertinent to the problems faced by the inventors of the '550 Patent because both *Feinauer* and the '550 Patent describe multiple improvements upon existing solutions. First, both *Feinauer* and the '550 Patent describe performing accent conversion in real-time. *Compare* EX1008-Feinauer, Abstract, [0002]-[0004]; *with* EX1001, Titles, Abstract, 2:4-14, 5:24-30. Second, both *Feinauer* and the '550 Patent describe capturing speaker characteristics and nuances in accent conversion. *Compare* EX1008-Feinauer, [0038], *with* EX1001, 1:60-66, 8:16-21. Additionally, both *Feinauer* and the '550 Patent describe accounting for phonetic differences between accents in accent conversion. *Compare* EX1008-Feinauer, [0038], [0040], *with* EX1001, 1:49-59, 7:55-64.

E. Overview of Osuna

78. “*Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion*” to Osuna et al. (“*Osuna*”) was publicly available at least as early

as October 2019. EX1010-Munford, ¶¶44-47 (declaration of June Munford establishing public availability of *Osuna*).

79. *Osuna* teaches “accent conversion” that “pairs frames between the native and non-native speaker.” EX1009-*Osuna*, Abstract, 4.

80. A PHOSITA would have identified *Osuna* as being within the same field of endeavor as the ’550 Patent because both *Osuna* and the ’550 Patent relate to speech audio processing and accent conversion. Compare EX1009-*Osuna*, Title, Abstract, 4, with EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Osuna* describes performing “accent conversion,” where frames are paired “based on the phonetic similarity between source and target frames.” EX1009-*Osuna*, 4.

81. A PHOSITA would have found *Osuna* reasonably pertinent to the problem faced by the inventors of the ’550 Patent because both *Osuna* and the ’550 Patent describe an improvement upon existing solutions, that is, capturing speaker characteristics and nuances in accent conversion. Compare EX1009-*Osuna*, Abstract, 1-2, 10, with EX1001, 1:60-66, 8:16-21.

F. Overview of Dalce

82. U.S. Patent Appl. No. 2013/0030789 to Dalce (“*Dalce*”) was filed on July 26, 2012, published on January 31, 2013. *Dalce* teaches “[a] universal language translator” EX1007-*Dalce*, Abstract.

83. A PHOSITA would have identified *Dalce* as being within the same field of endeavor as the '550 Patent because both *Dalce* and the '550 Patent relate to speech audio processing and accent conversion. *Compare* EX1007-Dalce, Abstract, [0010], [0013], [0015]-[0016], *with* EX1001, Abstract, 1:33-41, 2:4-14, 5:24-30, Fig. 4. *Dalce* teaches a system for dialect conversion between a speaker dialect and a target dialect. EX1007-Dalce, [0010], [0013], [0015]-[0016].

84. A PHOSITA would have found *Dalce* reasonably pertinent to the problems faced by the inventors of the '550 Patent because both *Dalce* and the '550 Patent describe multiple improvements upon existing solutions. First, both *Dalce* and the '550 Patent describe performing accent conversion in real-time. *Compare* EX1007-Dalce, [0017], [0033], *with* EX1001, Titles, Abstract, 2:4-14, 5:24-30. Second both *Dalce* and the '550 Patent describe an improvement upon existing solutions, that is, capturing speaker characteristics and nuances in accent conversion. *Compare* EX1007-Dalce, [0029], [0035], *with* EX1001, 1:60-66, 8:16-21.

VIII. SUMMARY OF UNPATENTABILITY

85. 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, 7, 10-13, 16, 19, and 21-22 are obvious over Fan (EX1004), Sonsaat (EX1005), and Loots (EX1006).	EX1004, EX1005, EX1006.

Ground 2: Claims 8-9, 17-18, 20 are obvious over Fan, Sonsaat, Loots, and Feinauer (EX1008).	EX1004, EX1005, EX1006, EX1008.
Ground 3: Claims 5 and 14 are obvious over Fan, Sonsaat, Loots, and Osuna (EX1009).	EX1004, EX1005, EX1006, EX1009.
Ground 4: Claims 6 and 15 are obvious over Fan, Sonsaat, Loots, and Dalce (EX1007).	EX1004, EX1005, EX1006, EX1007.

IX. OPINIONS REGARDING GROUND I: *FAN, SONSAAT, AND LOOTS*

A. Claim 1

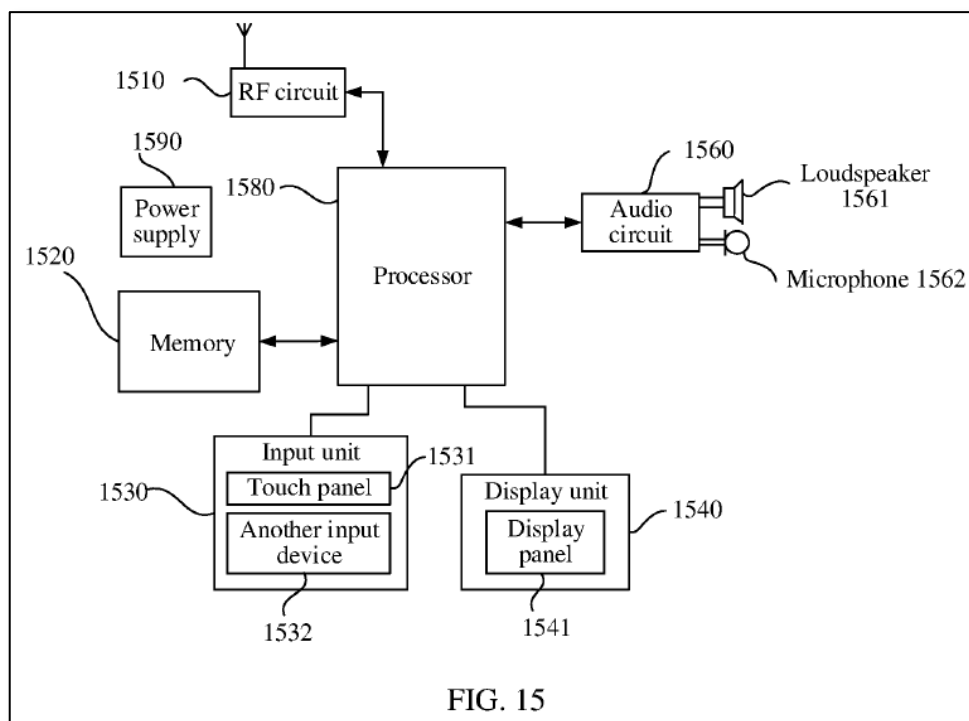
1. *Claim 1[Pre]: A system, comprising:*

86. To the extent the preamble is limiting, *Fan* discloses, or at least renders obvious, *a system* (e.g., *Fan*’s terminal), as claimed.

87. For example, *Fan* discloses a voice conversion apparatus “to implement diversified human voice beautification” such as dialect conversion (i.e., “adding an

accent or removing an accent” from human speech). EX1004-Fan, Abstract, [0003], [0005]-[0006], [0144] Figs. 10-12.⁴

88. *Fan* expressly teaches “a sound conversion method” that “is applied to a terminal. The terminal may be an intelligent terminal having a voice function, such as a mobile phone, a tablet computer, a notebook computer, or a smartwatch.” EX1004-Fan, [0045]. *Fan* further teaches a “voice conversion apparatus,” as shown in Figure 15.



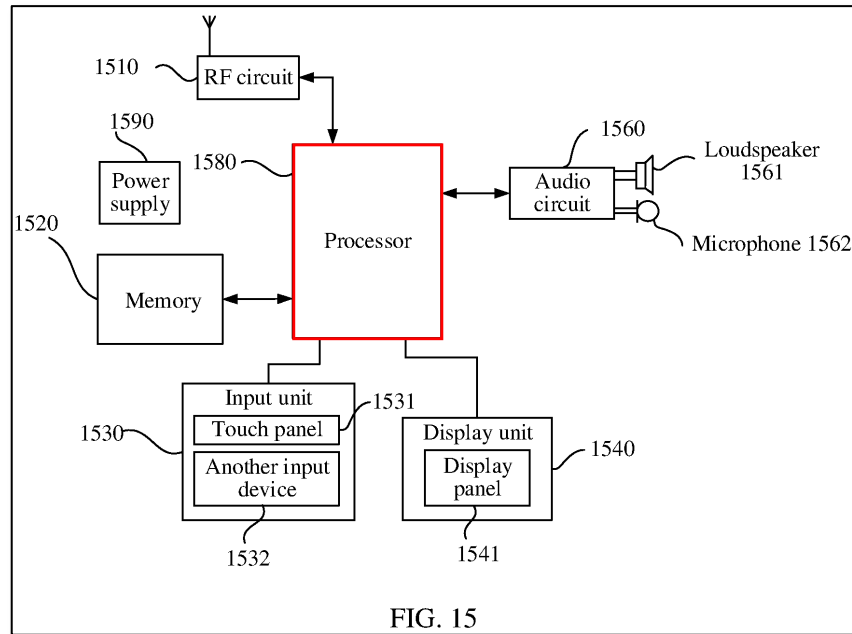
⁴ In addition to the citations to *Fan*, I rely on parallel citations to the certified translation of Fan’s Chinese application (EX1011) and PCT application (EX1016) as set forth in the above chart in Section VI.A.

EX1004-Fan, Fig. 15. “The voice conversion apparatus may be a terminal. The terminal may be a mobile phone, a tablet computer, a notebook computer, a smartwatch, or the like.” EX1004-Fan, [0191]. Therefore, a PHOSITA would have understood that *Fan* teaches *a system*, that is, *Fan*’s terminal.

2. Claim 1[a]: at least one processor; and

89. *Fan* discloses, or at least renders obvious, that its *system* (e.g., *Fan*’s terminal) comprises *at least one processor* (e.g., *Fan*’s processor 1580), as claimed.

90. For example, *Fan* discloses the terminal includes processor 1580. EX1004-Fan, [0192]. *Fan* teaches that processor 1580 “may be a general-purpose central processing unit (CPU), a microprocessor, an application-specific integrated circuit (ASIC), or one or more integrated circuits configured to control program execution...” EX1004-Fan, [0211]. “The mobile phone includes components such as ... a processor 1580[.]” EX1004-Fan, [0192]. The processor 1580 “is a control center of the mobile phone and is connected to each part of the entire mobile phone by using various interfaces and lines. ...the processor 1580 implements various functions of the mobile phone and processes data, to perform overall monitoring on the mobile phone. Optionally, the processor 1580 may include one or more processing units.” EX1004-Fan, [0192], [0203], Fig. 15.



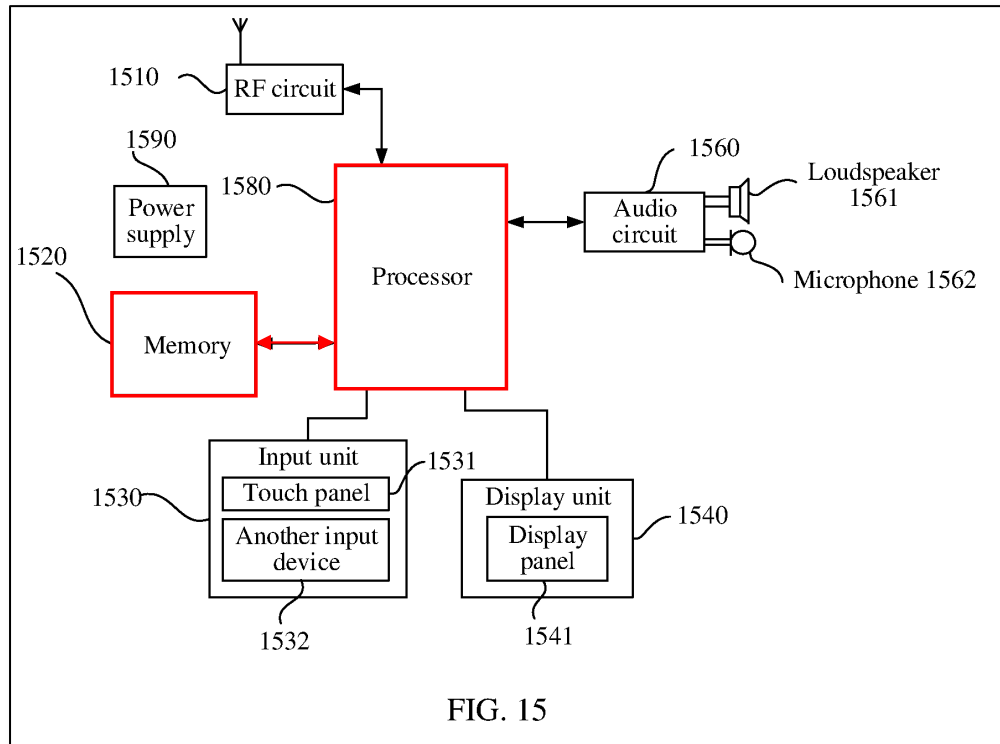
EX1004-Fan, Fig. 15, [0043].

91. For example, *Fan* discloses “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.” EX1004-Fan, [0024].

3. Claim 1[b]: non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor to cause the system to:

92. *Fan* teaches a non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor (e.g., *Fan*’s memory 1520 running the software program on processor 1580), as claimed.

93. *Fan* teaches “memory 1520 may be configured to store a software program and a module. ***By running the software program and the module that are stored in the memory 1520, the processor 1580 executes*** various function applications of the mobile phone and processes data.” EX1004-Fan, [0195], [0203] (further describing processor 1580 “***executing*** the software program and/or module that are stored in the memory 1520...”). Additionally, *Fan* teaches that “when program ***instructions*** stored in the memory are ***executed by the processor***, the methods performed by the terminal in the foregoing method embodiments are implemented.” EX1004-Fan, [0206], [0210] (“In another possible design, when the apparatus is a chip in the terminal, the chip includes a processing unit.... The processing unit may execute computer-executable instructions stored in a storage unit...the storage unit may be a storage unit that is in the terminal...”). Such methods include the dialect conversion function. EX1004-Fan, [0005] (“...an embodiment of this application provides a voice conversion method. The method is applied to a voice conversion apparatus. ...the apparatus may be a terminal, and the method may include: A terminal receives a mode selection operation...where the plurality of modes include...a dialect conversion mode...”), [0006] (“This embodiment provides the plurality of selectable modes, for example, ...the dialect conversion mode, for adding an accent or removing an accent”), [0016]-[0018].



EX1004-Fan, Fig. 15 (demonstrating the reciprocal connection between processor 1580 and memory 1520). Figure 15 of *Fan* is a schematic diagram of “a voice conversion apparatus,” including “components such as ...a memory 1520, ...a processor 1580,” where “the processor 1580 ...is connected to each part of the entire mobile phone by various interfaces and lines.” EX1004-Fan, [0043], [0192], [0203]. *Fan* teaches “memory 1520 may include a high-speed random access memory, and may further include a non-volatile memory, for example, at least one magnetic disk storage component, a flash memory component, or another volatile solid-state storage component.” EX1004-Fan, [0195]. Thus, *Fan* teaches a memory 1520 that is a *non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor.*

94. A PHOSITA would have understood that all the functions of *Fan* (and the *Fan*-based combinations) discussed in this Declaration, including training the voice feature extraction model and dialect conversion model, would be performed by *Fan*'s terminal's processor using program instructions stored in the memory. *Fan* teaches that the "voice conversion method ... is performed by a terminal device," and "may be performed by a processor in the terminal device, or the method may be performed by a chip in the terminal device." EX1004-Fan, [0054]. *Fan*'s teachings of "a voice conversion method" at least includes "[a] voice processing neural network [that] is configured in the terminal, and the voice processing neural network mainly includes a style conversion network, a dialect conversion network, and a voice enhancement model," training and execution of the voice feature extraction model, and training and execution of the dialect conversion model. EX1004-Fan, [0054] (regarding a voice processing neural network), [0110]-[0116] (regarding the voice feature extraction model), [0118]-[0138] (regarding the dialect conversion mode), [0139]-[0142], [0144] ("The foregoing describes a voice conversion method"). Furthermore, *Fan* teaches that "[t]he memory is configured to store a computer program stored in the memory, ***to perform corresponding processing and/or steps in any method embodiment.***" EX1004-Fan, [0174]. *Fan* also teaches that "when program instructions stored in the memory are executed by the processor,

the methods performed by the terminal in the foregoing method embodiments are implemented.” EX1004-Fan, [0206].

4. Claim 1[c]: train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

95. *Fan* in view of *Sonsaat* renders obvious *train[ing] a first machine-learning algorithm* (e.g., *Fan*’s training of its voice feature extraction NN model) *with first audio data* (e.g., *Fan*’s training dataset) *comprising speech content captured from a plurality of different speakers having a first accent* (e.g., *Sonsaat*’s English recordings from two speakers with the same non-English native language), *wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers* (e.g., *Sonsaat*’s forced alignment technique which aligns frames of the English recordings of the multiple speakers to the same transcript and classifies phonetic information of the frames at the word and phoneme levels), as claimed.

a) *Fan*’s Teachings

96. *Fan* teaches a voice feature extraction model (*a first machine-learning algorithm*). *Fan* teaches that “[t]he 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...” EX1004-Fan, [0007], [0070]. *Fan* teaches that the “voice feature extraction model may be a deep neural network” and “includes a plurality of convolutional layers, two LSTM layers, and one fully connected layer.” EX1004-Fan, [0085]. A PHOSITA would have understood that a neural network is a type of machine learning algorithm. For example, *Artificial Neural Network: Understanding the Basic Concepts without Mathematics* to Han et al. (“*Han*”) teaches that a neural network is a type of machine learning algorithm. EX1017-Han, 1; *See also*, above Section V.A. Additionally, the ’550 Patent teaches “a neural network, such as a generative adversarial network, one or more Griffin-Lim algorithms, etc.” as examples of machine learning algorithms. EX1001, 8:32-37.

97. *Fan* teaches the voice extraction model is trained in advance on corpora of voice data:

The voice feature extraction model *is trained in advance by using a large quantity of corpuses*⁵. For example, an *input of sample data*

⁵ A PHOSITA would have understood corpuses to mean corpora, the plural form of the term corpus. For example, *Crystal* teaches the plural form of corpus to be corpora. EX1032-Crystal, 143.

included in a training dataset is a voice, and a label is a PPG feature of the input voice. The voice feature extraction model is for extracting the PPG feature, and the voice feature extraction model does not participate in the parameter update of the entire network.

EX1004-Fan, [0113]; *see also* EX1004-Fan, [0085] (“The voice feature extraction model is obtained by performing training in advance based on a large quantity of corpuses.”). A PHOSITA would have understood that voice data is a form of *audio data*. Voice data consists of sound waves generated by a human speaker speaking. Audio data contains any sound waves, e.g., music, speech, noise, etc.

98. *Fan* also teaches that the trained voice feature extraction model is capable of extracting a PPG feature. EX1004-Fan, [0113] (“The voice feature extraction model is for extracting the PPG feature...”); A PPG is the probability that each frame of speech belongs to one of a set of pre-defined phonetic units. EX1004-Fan, [0052]; EX1009-Osuna, 3; *see also* Section V.C.

99. A PHOSITA would have understood that a PPG is a classification of audio data, and that a neural network that produces such a PPG is trained to perform such classification. For example, *Osuna* teaches that “[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.” EX1009-Osuna, 3. Additionally, *Phonetic Posteriorgrams for Many-to-One Voice Conversion Without Parallel Data Training* to Sun et al. (“*Sun*”) teaches that “[a]

PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class *for each specific time frame of one utterance.*” EX1029-Sun, 2.

100. However, *Fan* does not expressly describe all of the details for how its training dataset is compiled or expressly describe all of the details for how its feature extraction model is trained to produce the classifications present in its PPG feature. A PHOSITA would have looked to well-known examples of how to do both, for example as taught by *Sonsaat*.

b) *Sonsaat’s* Teachings

101. *Sonsaat* is titled “L2-ARCTIC: A Non-Native English Speech Corpus” and describes a process to compile, align, and classify a corpus of English speech from a plurality of non-native English speakers to train voice/accent conversion systems. EX1005-Sonsaat, Title, Abstract, 1 (the L2-ARCTIC corpus is used “[t]o train and evaluate voice and accent conversion systems designed for non-native speakers[.]”).

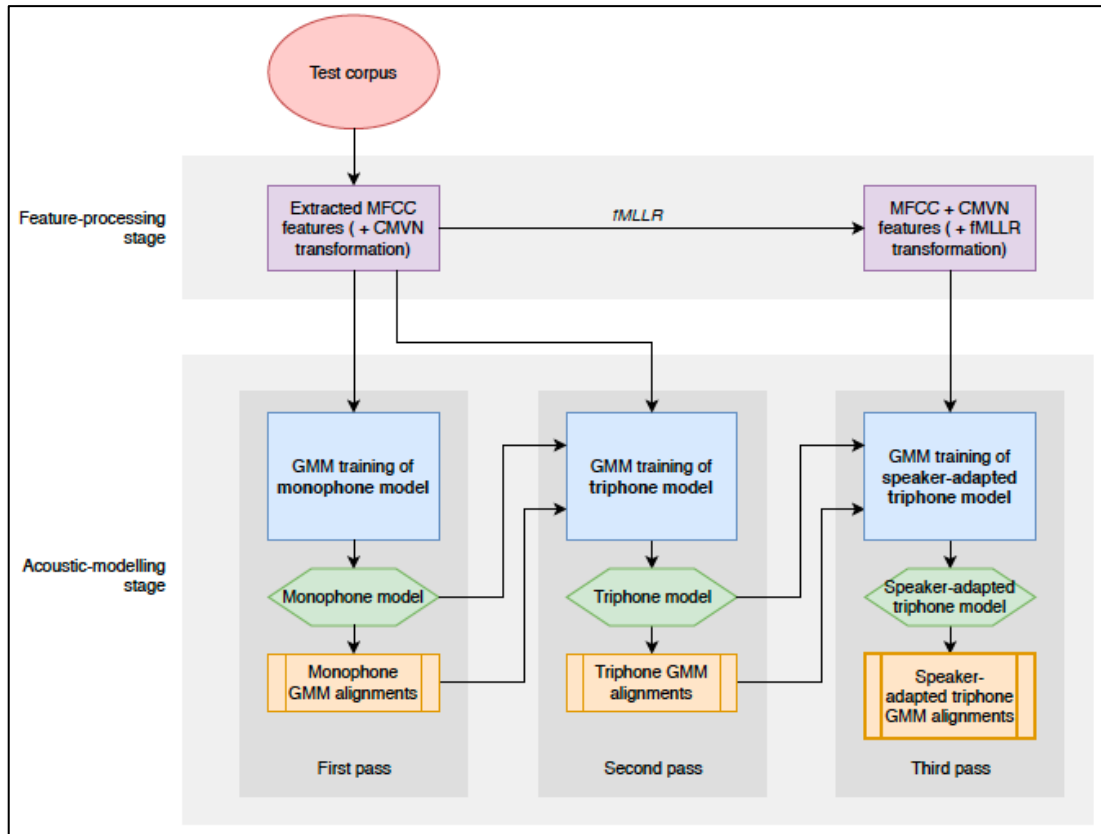
102. Specifically, L2-ARCTIC “contains English speech [recorded from] speakers from five different first languages[.]” EX1005-Sonsaat, 2. It “includes recordings from ten non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish, and Arabic, each L1 *containing recordings from one male and one female speaker.*” EX1005-Sonsaat, 1, 2. Thus, *Sonsaat* teaches *audio data comprising speech content captured from a plurality of different speakers*

having a first accent (e.g., English speech recordings from two Mandarin-native speakers) for use in training accent conversion systems.

103. *Sonsaat* explains that its speakers each follow the same transcript (i.e., “the 1,132 sentences in the CMU ARCTIC prompts”) to create the corpus recordings. EX1005-Sonsaat, 2.

104. *Sonsaat* also explains that its recordings are force-aligned to word-level and phoneme-level transcriptions, via a “Montreal forced-aligner” (MFA). EX1005-Sonsaat, 3. Force-alignment involves aligning each frame of a voice recording to a transcript so that the words/phonemes are aligned and correlated with the corresponding words/phonemes of the transcript. For example, *Corpus Phonetics Tutorial* to Chodroff (“*Chodroff*”) teaches that “[t]he Montreal Forced Aligner is a forced alignment system” that, “[a]s with any forced alignment system,” will “time-align a transcript to a corresponding audio file at the phone and word levels.” EX1037-Chodroff, 27. Specifically, *Sonsaat* teaches using its MFA “to produce phonetic transcriptions in PRAAT’s TextGrid format [42], which contains word and phone boundaries (Figure 1).” EX1005-Sonsaat, 3. A PHOSITA would have understood that *Sonsaat*’s (“MFA”) breaks speech into frames and aligns and classifies the speech of each speaker according to the same transcript. For example, *Montreal Forced Aligner: trainable text-speech alignment using Kaldi* to McAuliffe et al. (“*McAuliffe*”) teaches that MFA breaks speech into frames. Specifically,

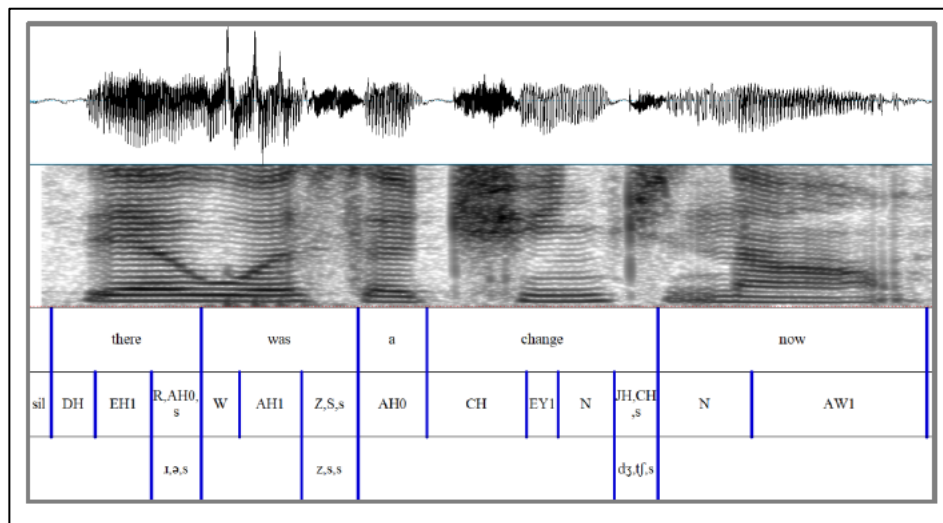
McAuliffe teaches that MFA “uses mel-frequency coefficients (MFCCs) as acoustic features,” where “[t]hirteen MFCCs are calculated with a 25 ms window size and 10 ms frame shift.” EX1014-McAuliffe, 1. *McAuliffe* further teaches that MFA aligns and classifies the speech of each speaker according to a transcript. Specifically, “speech and its corresponding orthographic transcription are automatically aligned at the word and phone level, given a way to map graphemes to phonemes ... and a statistical model of how phones are realized.” EX1014-McAuliffe, 1. Furthermore, *Forced Phonetic Alignment Using Deep Neural Networks* to Coles (“*Coles*”) teaches that MFA breaks speech into frames. Specifically, “[i]n the feature-processing stage” of MFA, “Mel-frequency coefficients (MFCCs) are calculated for **each window** in the audio recording, including delta and delta-delta features (first and second derivatives from the original MFCCs, respectively) **from surrounding frames.**” EX1013-Coles, 4. *Coles* further teaches that MFA “performs forced alignment given an audio recording, an orthographic transcription, and a pronunciation dictionary [26].” *Id.* *Coles* teaches that “forced alignment” is the “alignment of a speech recording and its orthographic transcription at the phone level[,]” and “its ultimate result identifies the position of phonemes in a recording.” EX1013-Coles, 1. *Coles*’s diagram below shows the process of the MFA generating a final alignment:



EX1013-Coles, 4, Fig. 3; *cf.* EX1001, 6:19-22 (“Thus, the ASR engine 302 may align and classify each frame of the captured speech content according to its monophone and triphone sounds, as indicated in the corresponding transcript.”).

105. Accordingly, because *Sonsaat*’s English recordings from two speakers with the same accent are recorded from the same transcript and are force-aligned to such same transcript at the word/phoneme levels via an MFA (where MFAs operate on frames), *Sonsaat* teaches *aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers*, as claimed.

106. A sample TextGrid of processed data with transcriptions is provided in Fig. 1, where the speech waveform and spectrogram are aligned and classified with the words, phonemes, and error tags.



EX1005-Sonsaat, Fig. 1.

107. Although *Sonsaat* is targeted to English accents, a PHOSITA would have understood that the same teachings are readily applied to accents in other languages, for example, the dialects discussed by *Fan*. The method of corpus creation disclosed in *Sonsaat* would be applicable to any language. There is no technical difference in recording speech from multiple speakers in Chinese, nor in aligning and classifying Chinese voice data to a transcript. In fact, *McAuliffe* notes that “MFA has been successfully applied to 29 languages[.]” EX1014-McAuliffe, 1. Thus, in the example of *Fan* discussed below in Claims 1[d]-1[f], the first accent is Sichuanese.

c) Motivation to Combine

108. *Fan* teaches a *first machine-learning algorithm* trained on *first audio data*. But, *Fan* does not expressly disclose the audio data being speech content captured from a plurality of different speakers with a first accent, or of the training comprising aligning and classifying frames of said speech content. However, *Sonsaat* teaches *first audio data comprising speech content captured from a plurality of different speakers having a first accent and aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers*. Namely, the *Fan-Sonsaat* combination uses the method of corpus creation taught by *Sonsaat* as part of training the feature extraction model of *Fan*.

109. A PHOSITA would have been motivated to include *Sonsaat's* speech corpus recordings and force alignment training technique as part of training *Fan's* voice feature extraction model with a reasonable expectation of success for several reasons.

110. Regarding the inclusion of *Sonsaat's* speech corpus recordings, *Fan* expressly discloses using a large corpus of voice data to train the voice feature extraction model but does not specify the makeup of the training data or the training process. EX1004-Fan, [0085] (“The voice extraction model is obtained by performing training in advance based on a large quantity of corpuses.”), [0113]

(“The voice feature extraction model is trained in advanced by using a large quantity of corpuses.”). *Sonsaat* expressly discloses that its speech corpus is intended for “voice conversion, accent conversion, and mispronunciation detection” and that its speech corpus is “a valuable resource not only for research in voice conversion and accent conversion but also in ***computer-assisted pronunciation training***.” EX1005-Sonsaat, 1. Based on these express disclosures, a PHOSITA would have been motivated to use *Sonsaat*’s corpus recordings to train the voice feature extraction model of *Fan*’s computer-assisted pronunciation. The large quantity of speakers in the L2-ARTIC corpus is consistent with the training data necessary to train *Fan*’s voice feature extraction model.

111. Additionally, such a modification would have constituted the use of a known technique (*Sonsaat*’s corpus creation) to improve a similar device (*Fan*’s trained voice conversion system) in the same way (*Fan*’s trained voice conversion system would benefit in the same way as a system using *Sonsaat*’s training corpus—e.g., improving the training dataset by focusing on non-native English speakers). EX1005-Sonsaat, 1-2 (see section titled “The need for a new L2 English corpus”). A PHOSITA would have recognized that using voice data from multiple speakers with the same accent would have improved the accuracy of the model and made it more robust for operating across different speakers. It was known that increasing the size of training data would enhance the accuracy of a neural network model.

EX1038-Saragih, 1; *see also* EX1015-Mohri, 19 (“[T]he amount of training data is often chosen to be larger than that of the test data since the learning performance directly depends on the training sample.”). In fact, it was already considered a “baseline approach” to train an ASR model with a multi-speaker corpus of a single accent. For example, *Liu* teaches that a baseline approach to an accent translation system is to “first train a speaker-independent ASR (SI-ASR) model using a native multi-speaker corpus.” EX1043-Liu, 2.

112. Further, implementing the Montreal forced-aligner of *Sonsaat* into the programming of *Fan’s* terminal allows the terminal to train the voice feature extraction model on the best available machine alignment, increasing the accuracy of the PPG feature produced by the model. EX1013-Coles, 1-2 (“any significant improvements in alignment performance is likely to have a wide range of beneficial consequences. ... The Montreal Forced Aligner (MFA), the Kaldi-based successor to the Prosodylab-Aligner on which this paper is based, distinguished itself from previous aligners in its trainability (it is shipped with pre-trained models to speed up computation, but can also be retrained on new data) and its architecture (which used more elaborate modelling techniques than its predecessors). MFA outperformed Prosodylab-Aligner and FAVE in *producing alignments closest to human-annotated gold standards* on the Buckeye and Phonsay corpora [26].”) ; EX1014-

McAuliffe, Abstract, 1, 4 (“Our evaluation suggests that both MFA’s more complex architecture and the ability to retrain on new data generally improve performance.”).

113. A PHOSITA would have been motivated and found it obvious to implement *Sonsaat’s* Montreal forced-aligner to prepare audio data as part of training *Fan’s* feature extraction model. For example, *Osuna* teaches to that to create a deep neural network for creating PPGs, the neural network must be trained with inputs of “stacked MFCC frames” and target outputs of senone labels obtained from force-alignment using an existing GMM-HMM speech recognizer. EX1009-Osuna, 3-4. MFA is one such GMM-HMM speech recognizer. EX1013-Coles, 2. A PHOSITA would have understood this modification is nothing more than combining prior art elements (i.e., *Sonsaat’s* Montreal forced-aligner and *Fan’s* trained feature extraction model) according to known methods to yield predictable results of increasing accuracy of the feature extraction model by improving the phonetic recognition of the model to output.

114. A PHOSITA would have had a reasonable expectation of success (REOS) in implementing *Sonsaat’s* functionality into *Fan’s* terminal. *Fan* teaches “[t]he voice feature extraction model is ***trained*** in advance” before being used in the accent conversion process. EX1004-Fan, [0113]. A PHOSITA would have understood that training a neural network to recognize phonemes (for example, to extract a PPG feature) requires processing input speech data into phonemes based

on a transcript. A neural network is trained to produce an output, e.g., a PPG feature representing a phoneme. *See* Section V.A. To produce the PPG feature, the voice feature extraction model must be trained on target output data of a PPG feature of a relevant input frame of voice data. Thus, both the output and input data must be aligned to a transcript for the neural network to learn the proper association between the data. Further, a PHOSITA would have a reasonable expectation of success applying *Sonsaat's* training teachings to train *Fan's* voice feature extraction model because *Sonsaat's* corpus is meant (and ready) for implementation in accent conversion systems such as *Fan*. EX1005-Sonsaat, Abstract (“The corpus is publicly accessible at [URL].”). The L2-ARTIC corpus was not only publicly available, but specifically intended for application in the “voice conversion [and] accent conversion” taught by *Fan*. EX1005-Sonsaat, 1. A PHOSITA would have understood that implementing the Montreal forced-aligner of *Sonsaat* into the terminal of *Fan* would have constituted a simple design choice and only required a minor programming adjustment, well within the skill level of a PHOSITA. For example, *Chodroff's* teaches to set up and run the Montreal forced-aligner. EX1037-Chodroff, 27-30. Additionally, the Montreal forced-aligner runs on the Kaldi toolset, a toolset commonly used for developing voice recognition neural networks which PHOSITAs were aware of and understood how to readily implement). EX1014-

McAuliffe, 1 (“MFA is built on top of Kaldi, an actively maintained, open-source automatic speech recognition toolkit[.]”); *see* EX1013-Coles, 2.

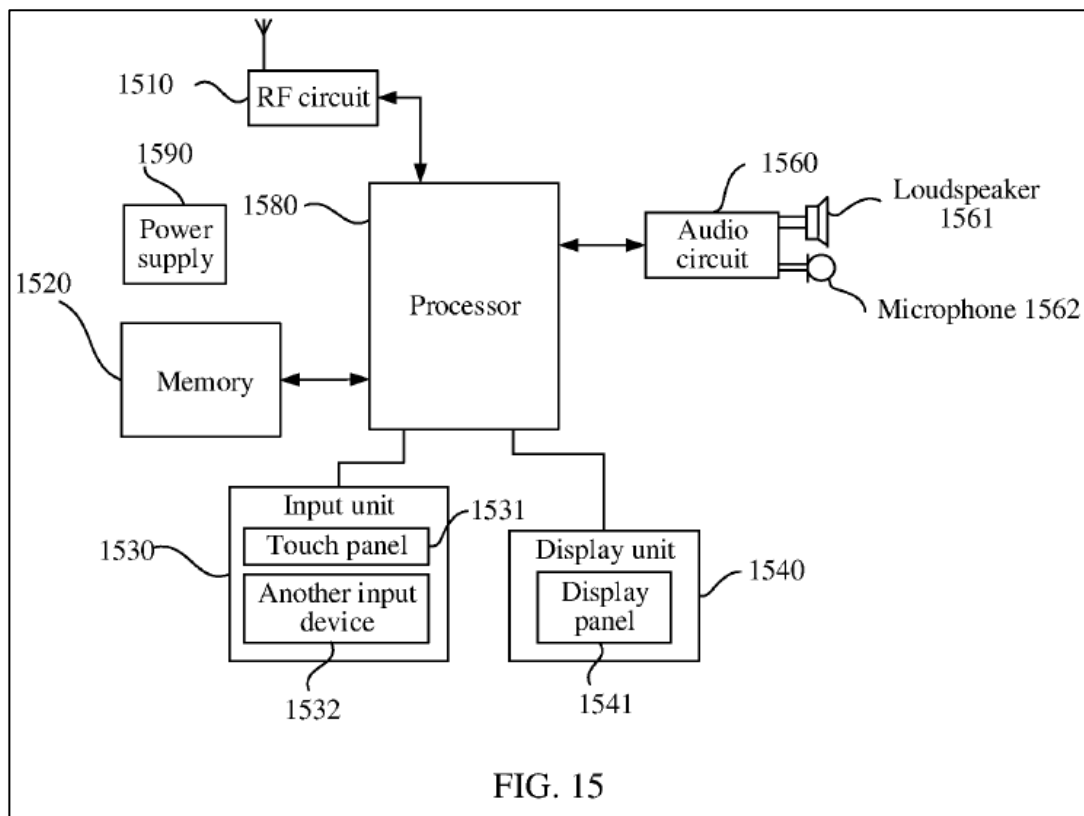
5. Claim 1[d]: apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

115. *Fan teaches apply[ing] the first machine-learning algorithm (e.g., Fan’s voice feature extraction NN model) to speech content received via at least one microphone (e.g., Fan’s first voice input received by a microphone), and comprising a set of phonemes associated with a first pronunciation of the received speech content (e.g., Fan’s phonemes associated with a first accent, e.g., a Sichuanese accent), to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (e.g., Fan’s voice feature extraction NN model extracts a PPG feature from the first voice input), as claimed.*

(1) “apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content”

116. *Fan teaches “[t]he terminal device receives, **by using a microphone, the first voice input by the first speaker**” where the first voice is the voice “the user expects to beautify (or process).” EX1004-Fan, [0063]. Fan further teaches that*

“[t]he audio circuit 1560, a loudspeaker 1561, and *a microphone 1562* may provide an audio interface between the user and the mobile phone[,]” where “the microphone 1562 converts the collected sound signal into an electrical signal, the audio circuit 1560 receives the electrical signal, converts the electrical signal into audio data, and then the audio circuit 1560 outputs the audio data to the processor 1580 for processing.” EX1004-Fan, [0200].



EX1004-Fan, Fig. 15. *Fan* teaches that “*dialect conversion*” is one of the beatification processes performed by the terminal, which comprises “*adding an accent to or removing an accent from the first voice*[.]” EX1004-Fan, Abstract (“...a dialect conversion mode, for adding an accent to or removing an accent from

the first voice; ...”), [0006] (“...the dialect conversion mode, for adding an accent or removing an accent; ...”), [0045] (“A voice processing neural network is configured in the terminal device, and ... mainly includes ... a dialect conversion network.... ...a dialect conversion mode, for adding an accent or removing an accent; ...”), [0056] (“a plurality of modes include ... a dialect conversion mode.... The dialect conversion mode is for dialect conversion, for example, adding an accent or removing an accent.”), [0120] (“The plurality of dialect conversion models include at least two types of dialect conversion models. A first type is removing an accent.... A second type is adding an accent....”), [0124] (“Optionally, when the mode selection operation is for selecting a dialect switching mode, the mode selection operation (a level-1 operation) may further include a plurality of selection operations at a next level (a level-2 operation), for example, a first selection operation and a second selection operation. The first selection operation is for selecting “Remove accent”, and the second selection operation is for selecting “Add accent”. ... When the user selects “Add accent”, the user interface of the terminal displays a level-3 category label of accent options of different places[.]”), [0036] (“Fig. 9 is a schematic diagram of an example of a dialect conversion process of a voice according to an embodiment of this application;”). As explained in the next paragraphs, *Fan* receives the first voice input from a speaker with pronunciations associated with a first accent (e.g., Sichuanese) and extracts phonetic posteriorgrams

(PPGs) from such input (where the PPGs represent phonemes present in the first voice input, which are associated with the pronunciation of the first accent).

117. In one example, the terminal of *Fan* converts Sichuanese into Mandarin. EX1004-Fan, [0120]. *Fan* teaches that the voice feature extraction model derives a PPG feature of this input voice that is input to a dialect conversion model to perform the conversion. EX1004-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.”), [0126] (“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...”), [0127] (“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.”).

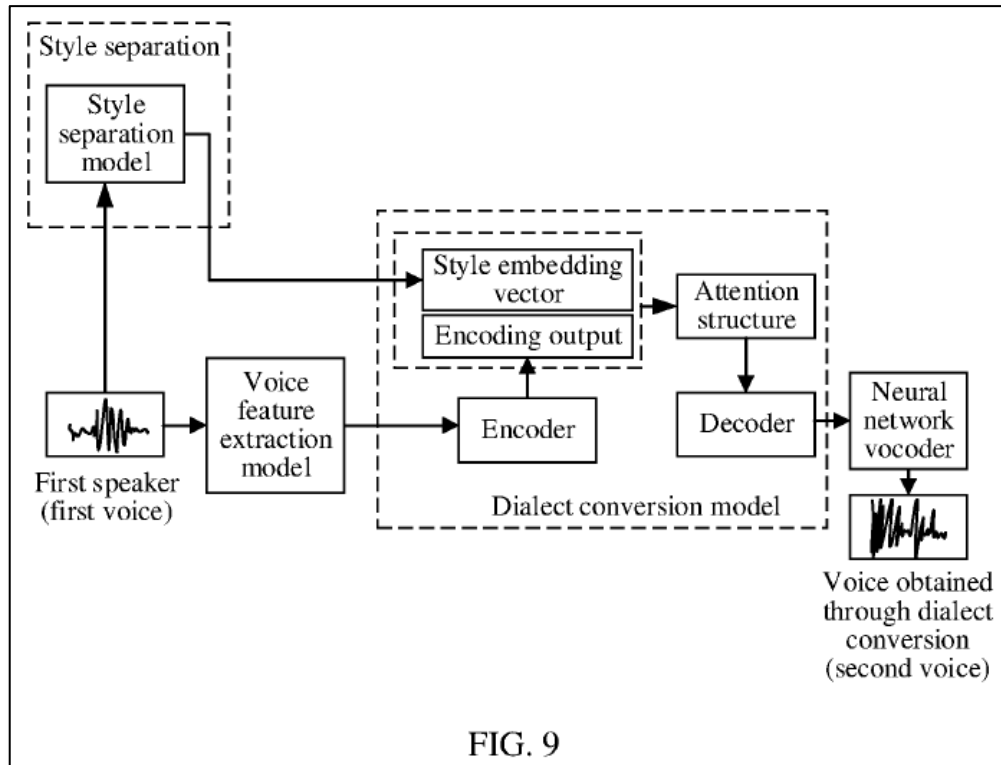


FIG. 9

EX1004-Fan, Fig. 9. A PHOSITA would have understood that the input voice (Sichuanese, in this example) consists of a set of phonemes associated with the Sichuanese dialect.⁶ Speech in every accent contains the phonemes used by speakers

⁶ A PHOSITA would have understood that an accent is a component of dialect, which is consistent with *Fan's* teaching that the “dialect conversion mode [is] for adding an accent or removing an accent[.]” EX1004-Fan, [0075]. An accent is generally a plurality of pronunciations, while a dialect includes these pronunciations as well as other language features such as vocabulary and grammar. *See* Section V.E. Although dialect is more inclusive, a PHOSITA would have understood that the dialect conversion of *Fan* is directed towards pronunciation, as demonstrated by *Fan's* use of the words “dialect” and “accent.” EX1004-Fan, [0120] (“A first type is *removing an accent, that is, converting a dialect* into Mandarin. For example,

of the accent. For example, *Alteration of [n] and [l] in Sichuan Dialect, Standard Mandarin and English: a Single-Case Study* to Zhang (“Zhang”) teaches that “[t]here are many differences between Sichuan dialect and Standard Mandarin in phonetics,” including the “alternation of [n] and [l].” EX1039-Zhang, 1. Specifically, “[i]n Sichuan dialect, the alveolar nasal [n] and the lateral approximant [l] are not distinct. The consonant [n] of Sichuan dialect corresponds to a nasal consonant [n] and a lateral approximate [l] in Standard Mandarin.” EX1039-Zhang, 3. Zhang further teaches that “[w]hen Sichuan dialect-speaking people speak Standard Mandarin, they cannot pronounce correctly if the word’s initial consonant is [n] or [l], for example, they will say [la] for *na* ‘take’, [fən lu] for *fen nu* ‘angry’.” *Id.* Fan teaches that phonemes are “the smallest phonetic unit of a pronunciation action.” EX1004-Fan, [0051]. A PHOSITA would have understood that all speech consists of phonemes. See Section V.B. Further, a PPG (phoneme posteriorgram)

Sichuanese is converted into Mandarin. A second type is ***adding an accent, that is, converting Mandarin into a dialect***. For example, Mandarin is converted into Sichuanese.”), [0135] (“When the user selects “***Switch dialect***”, a level-2 category label is displayed in the interface. For example, the level-2 category label includes “***Remove accent***” and “***Add accent***.”), [0137] (“the first voice is input into a selected model (for example, ***a dialect conversion model for converting Mandarin into a Guangdong accent***).”).

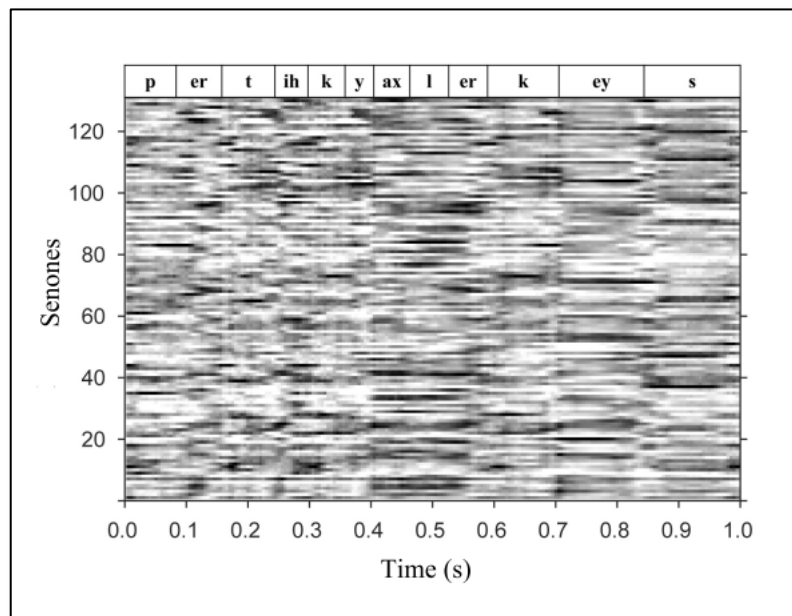
feature “describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and is equivalent to recognizing the phoneme.” EX1004-Fan, [0007]. A PHOSITA would have also understood that, because the input has the Sichuanese accent, the PPG feature produced from the first voice input speech reflects the underlying phonemes present in the first voice input speech, particularly with the pronunciations of the Sichuanese accent. For example, substituting [l] for the [n] pronounced in standard Mandarin. EX1039-Zhang, 3. Therefore, *Fan* teaches to *speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content.*

118. *Fan* teaches *apply[ing] the first machine-learning algorithm* to such *speech content* because *Fan* teaches its voice feature extraction model extracts PPGs from the first voice input. As explained in the next sub-section, these extracted PPGs teach the claimed *non-text linguistic representation[s]*.

(2) “*a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;*”

119. *Fan* teaches that dialect conversion is performed in a multi-step process. The first step requires extracting a PPG feature from the input voice. Namely, *Fan* teaches that “[t]he 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.*” EX1004-Fan, [0007], [0085]. The PPG describes the probability each frame of the input voice corresponds to a specific phoneme. EX1004-Fan, [0007]. A PHOSITA would have understood that as PPGs represent the posterior probabilities of each phonetic class in the form of a time-versus-class matrix, such as taught by *Sun*.



EX1029-Sun, Fig. 4; *see also* Section V.C.

120. *Fan's* PPG features are *non-text linguistic representation[s]* of the input speech, as claimed. A PPG “describes a probability of *each phoneme in a phoneme set* corresponding to a voice frame, and is equivalent to recognizing the phoneme.” EX1004-Fan, [0067], [0085]. The “PPG feature is used as a content information input of the first voice, and these voice units (phonemes) retain a language and *voice information of the voice.*” EX1004-Fan, [0067], [0086], [0150]. *Fan* also teaches “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.” EX1004-Fan, [0007]. Thus, the PPG feature is a *non-text linguistic representation*. A PHOSITA would have understood the PPG feature retains the content and voice information of the input voice in the accent of the speaker (for example, Sichuanese), which includes phonemes associated with said accent. Different accents in a language use different sets of phonemes, and that certain words are comprised of different phoneme for different accents. Thus, it would have been understood that the PPG feature represents the phonemes of the accent of the first voice for each word. For example, as shown in the table below, the word “reactions” comprises a different set of phonemes in South African English (SAEDICT) and Received Pronunciation (BEEP).

Table 3 Two aligned pronunciations of the word “reactions”.								
SAEDICT	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	<i>ax</i>	<i>n</i>	<i>s</i>
BEEP	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	–	<i>n</i>	<i>z</i>

EX1006-Loots, Table 3. Similarly, *Zhang* teaches that the Sichuanese accent and Standard Mandarin comprise differing sets of consonants, as shown in the table below.

Table2. Sichuan dialect consonants and Standard Mandarin consonants ¹								
Manner of Place of articulation	Plosive		Affricate		Fricative		Nasal	Lateral approximant
	Voiceless		Voiceless		Voiceless	Voiced	Voiced	Voiced
Bilabial	p	p ^h					m	
Labiodental					f			
Alveolar	t	t ^h	ts	ts ^h	s	(z)	n	<ɿ>
Retroflex			<ʈʂ>	<ʈʂ ^h >	<ʃ>	<ʒ>		
Alveolo-palatal			tʃ	tʃ ^h	ʃ		(ɲ)	
Velar	k	k ^h			x		(ŋ)	

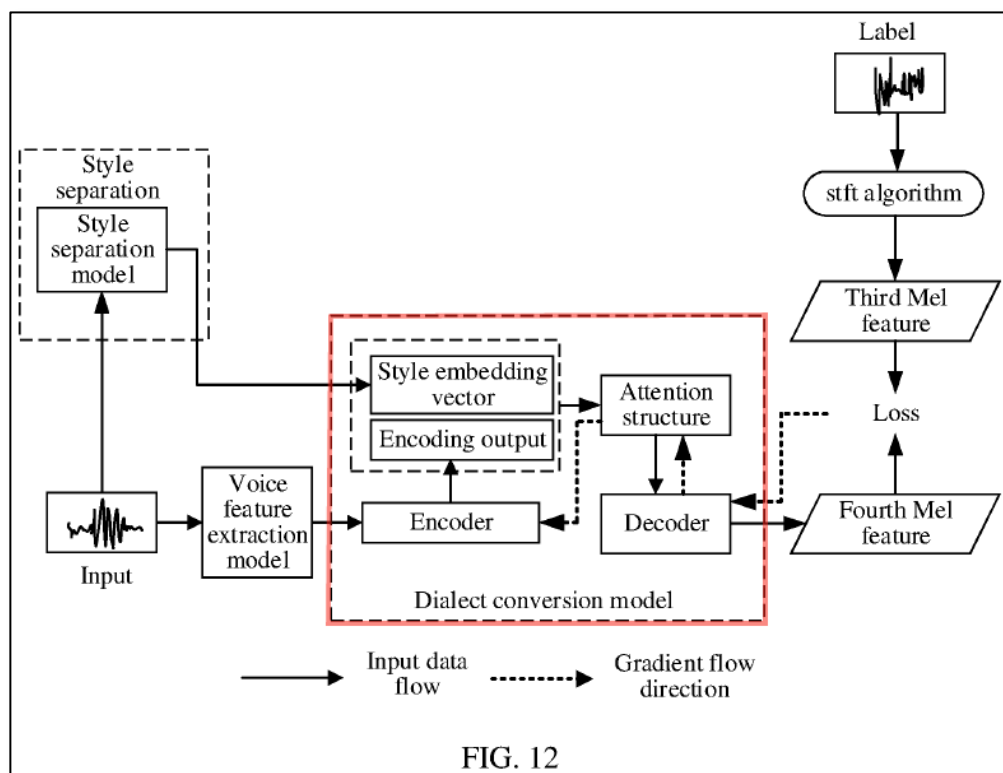
EX1039-Zhang, Table 2, 2 (“The symbols in angle brackets are the consonants which Standard Mandarin has, but do not exist in Sichuan dialect; the symbols in the parentheses exist in Sichuan dialect but not Standard Mandarin. The remaining consonants exist both in Sichuan dialect and Standard Mandarin.”). *Zhang* further teaches that certain words comprise different phonemes for the Sichuanese accent and Standard Mandarin. For example, the word ‘head’ in Standard Mandarin is *tou nao*, but in a Sichuanese accent ‘head’ is *tou lao*. Thus, *Fan*’s extracted PPG teaches *a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content.*

6. Claim 1[e]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent,

121. *Fan teaches based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content (e.g., Fan's extracted PPG feature representing the set of phonemes associated with a first accent, e.g., Sichuanese accent), synthesize, using a second machine-learning-algorithm (e.g., Fan's dialect conversion NN model) trained with (i) second audio data comprising the first accent (e.g., Fan's "input" audio data having a first accent, e.g., Sichuanese accent) and (ii) third audio data comprising a second accent (e.g., Fan's "label" audio data with a second accent, e.g., Mandarin accent), fourth audio data representative of the received speech content having the second accent (e.g., Fan's mel feature form consisting of the input first voice with a second accent, e.g., Mandarin accent), as claimed. I note here again that I have been instructed to construe fourth audio data representative of as "synthesized acoustic feature data generated as an output of a speech conversion process and used as input to a subsequent conversion process that produces output waveform audio."*

122. *Fan teaches a dialect conversion network which "includes a plurality of dialect conversion models, where the dialect conversion model is for performing*

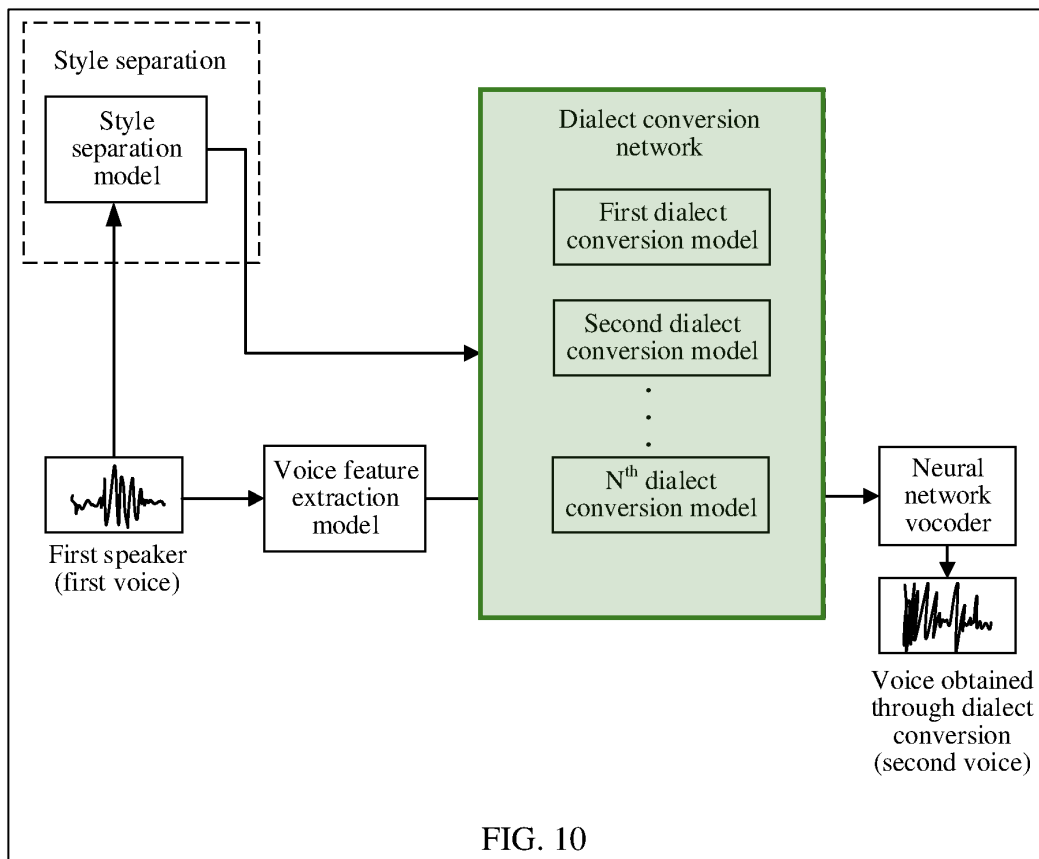
dialect conversion on an input first voice.” EX1004-Fan, [0119], Figs. 8-10, 12. Each “dialect conversion model *is a seq2seq neural network*” and is separate from the voice feature extraction model. EX1004-Fan, [0140], Fig. 8-10, 12. “The seq2seq neural network includes an encoder, a decoder, and an attention structure[,]” as shown in the figure below.



EX1004-Fan, Fig. 12 (The seq2seq neural network annotated in red). Thus, the dialect conversion model is a *second machine-learning-algorithm*.

123. Each of the dialect conversion models is “for a different dialect to be converted, and each dialect conversion has a corresponding model.” EX1004-Fan, [0130], Fig. 10. For instance, there are “at least two types of dialect conversion

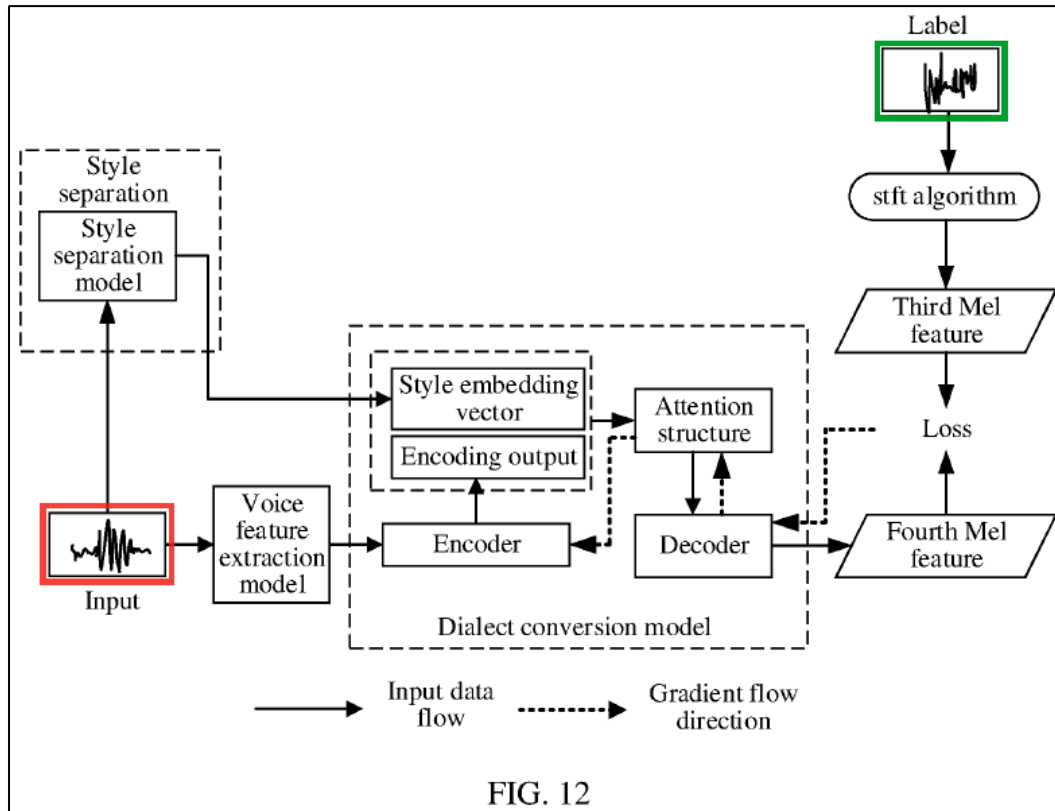
models” with one for “removing an accent” and one for “adding an accent, that is, converting Mandarin into a dialect.” EX1004-Fan, [0120], [0135]. As annotated in green below, the dialect conversion network that includes the dialect conversion models is separate from the voice feature extraction model (*first machine-learning-algorithm*).



EX1004-Fan, Fig. 10 (the dialect conversion network is annotated in green).

(1) *Second machine-learning-algorithm trained with*
(i) *second audio data comprising the first accent and* (ii)
third audio data comprising a second accent

124. *Fan* teaches the “**training** and updat[ing] processes of the dialect conversion model” (a **trained second machine-learning-algorithm**). EX1004-Fan, [0139]-[0141]. *Fan* teaches “each piece of **sample data** in a **training dataset** corresponding to the first dialect conversion model includes an input and a label.” EX1004-Fan, [0140], Fig. 12. In the provided example of training the dialect conversion model, “[t]he input is Sichuanese, and the label is Mandarin with the same content as Sichuanese[.]” EX1004-Fan, [0140], Fig. 12. A PHOSITA would have understood that the input is audio data comprising the first accent (Sichuanese) and the label is audio data comprising the second accent (Mandarin). As shown in *Fan*’s Figure 12 below, the Input (annotated in red) is in the form of an audio waveform, as is the Label (annotated in green). Further, the “Label” is processed using a short time Fourier transform, which is a mathematical function applied to a waveform to outputs a spectrum in the frequency domain (mel features).



EX1004-Fan, Fig. 12. Therefore, *Fan's* Sichuanese input teaches a *second audio data comprising the first accent* and *Fan's* Mandarin label teaches a *third audio data comprising a second accent*.

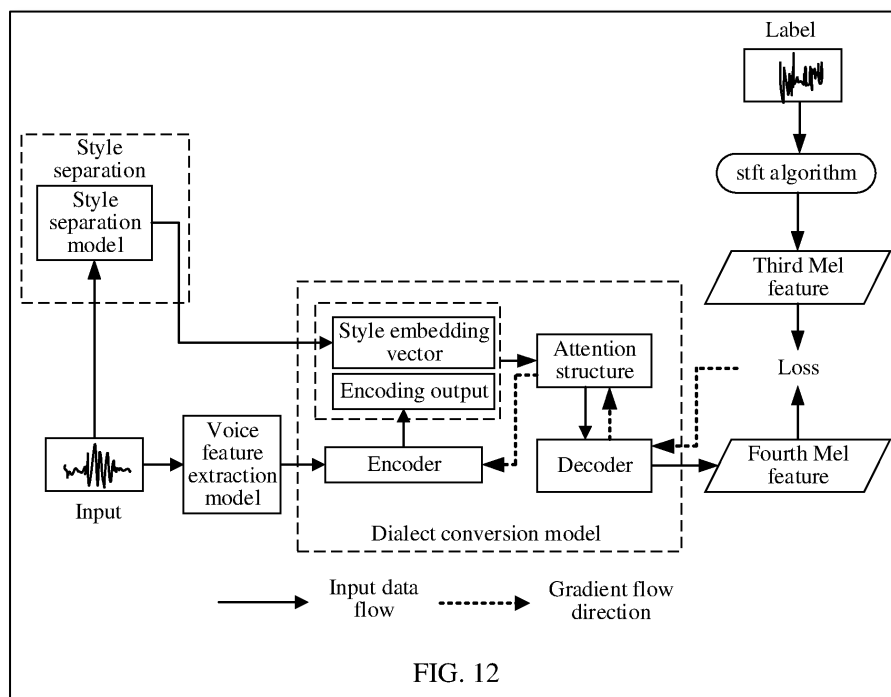
125. Each dialect conversion model is trained by using a Short-Time Fourier Transform algorithm to obtain a third mel⁷ feature of the label and a fourth mel

⁷ *Fan* discusses the third and fourth “melt” features. A PHOSITA would have recognized that “melt” is a typographic error throughout *Fan's* specification. *Fan* describes mel features as “a voice signal being stored as a waveform diagram feature is represented as feature transformation in time domain of the voice signal.” EX1004-Fan, [0053]. A PHOSITA would have understood a mel feature is a

feature of the input. A comparison between the third and fourth mel features produces “[a] loss value and a gradient[.]” EX1004-Fan, [0141]. “When the gradient flows through a corresponding structure (for example, an encoder, an attention

representation of audio data in the frequency domain in a scale which reflects the perception of human auditory system, like a mel spectrogram (and specifically a PHOSITA would have understood a mel feature as a mel-based spectrogram). *See* Section V.F.2. A mel-based spectrogram would be the expected result of applying the short time Fourier transform, which converts the time domain waveform into a frequency domain spectrum. In some paragraphs the specification properly uses the term mel feature. *See* EX1004-Fan, [0053] (“**Mel (mel) feature**: A voice signal being stored as a waveform diagram feature is represented as feature transformation in time domain of the voice signal... As a feature representation of the voice signal, **the mel feature** can clearly represent a feature of a voice resonance peak.”), [0097], [0099], Figs. 7 and 12. In other paragraphs the specification uses the term “melt” feature. *See* EX1004-Fan, [0116] (“The label obtains **a first melt feature** by using a short-time fourier transform (STFT) algorithm. After a voice is input into the style conversion network, the network outputs **a second melt feature**. ...”), [0126] (“...outputs **a melt feature** form of the second voice by using the dialect switching model. ...”), [0141] (“Therefore, a label obtains **a third melt feature** by using an STFT algorithm, and the network outputs **a fourth melt feature**. A loss value and a gradient are obtained by comparing **the third melt feature and the fourth melt feature**. ...”). As a melt feature is non-sensical in plain meaning and was not a term in the art, a PHOSITA would have understood “melt feature” to be properly read as “mel feature” throughout the specification.

structure, and a decoder), the parameter in the module is updated.” *Id.* This continues until the loss value decreases to a specific value and does not show a continued significant decrease trend. *Id.* This indicates the network has converged, and the training can stop. *Id.*



EX1004-Fan, Fig. 12, [0140]-[0141], [0040] (“FIG. 12 is a schematic diagram of training and updating a dialect conversion model according to an embodiment of this application”). Thus, the dialect conversion model (*second machine-learning algorithm*) is trained using voice data having both the first and second accents.

(2) based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm ... fourth audio data representative of the received speech content having the second accent

126. *Fan* teaches that its PPG feature (*derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content*) extracted from its voice feature extraction model is subsequently input into the trained dialect conversion model to output (*synthesize*) a mel feature which represents the input speech but now with a second dialect (*accent*):

the terminal inputs a PPG feature into a dialect switching model...and ***outputs a mel[] feature form of the second voice*** by using the dialect switching model. ... The first voice is a voice of a first dialect, and ***the second voice is a voice of a second dialect***.

EX1004-Fan, [0126], [0162]. This process is shown in Fig. 9.

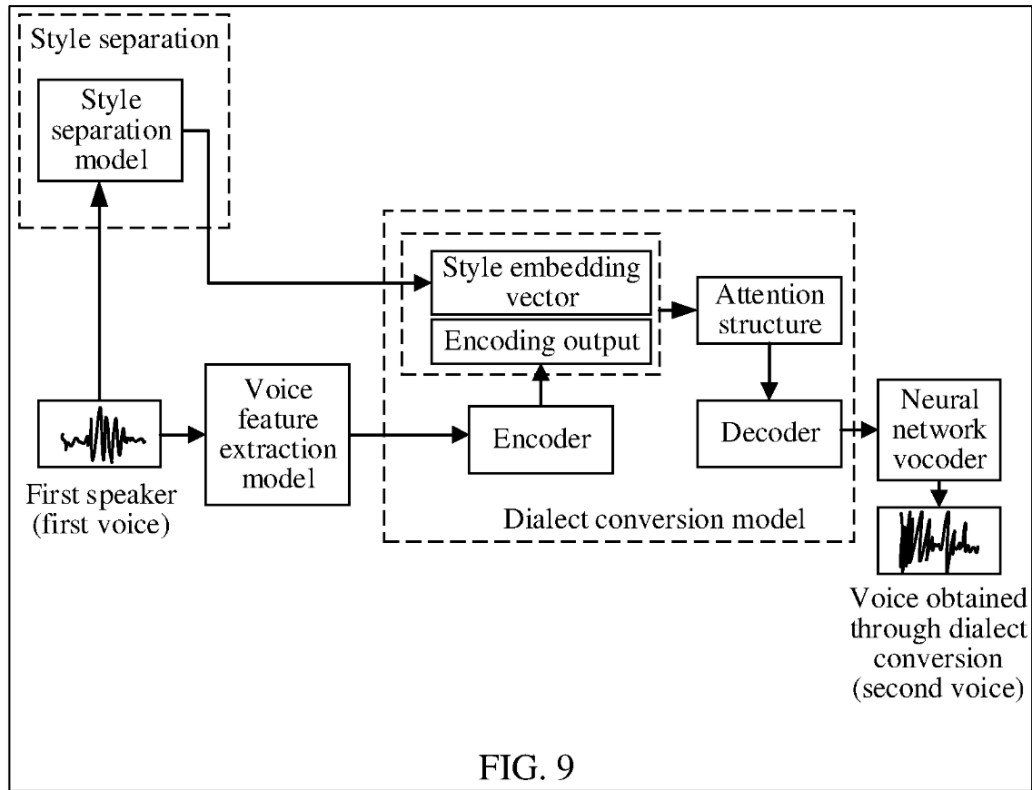


FIG. 9

EX1004-Fan, Fig. 9.

127. Further, *Fan* teaches that a style separation feature separates a style of the first voice to obtain a timbre feature vector of the first voice. EX1004-Fan, [0127] (“...the terminal inputs a [first] 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.”). The timbre feature and the PPG feature are input into the dialect conversion model, and a second voice is output that has both the same content (provided by the PPG feature) and speaking style (provided by the timbre feature) (*representative of the received speech content*) of the input voice but

having the second accent. EX1004-Fan, [0126], [0127] (“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.”). *Fan* teaches that the dialect conversion model outputs a mel feature (*fourth audio data representative of the received speech content having the second accent*), which is subsequently converted into a playable voice form by using a neural network vocoder. EX1004-Fan, [0126] (“The terminal...outputs a [mel] feature form of the second voice by using the dialect conversion model. Then, a [mel] feature is converted into a playable voice form by using a neural network vocoder.”). A PHOSITA would have understood that the dialect conversion model *synthesizes* the mel feature by processing the PPG feature and timbre feature with the trained dialect conversion model to output said mel feature. Accordingly, in my opinion, *Fan* also teaches the claimed *fourth audio data representative of* under Petitioner’s proposed construction of “synthesized acoustic feature data generated as an output of a speech conversion process and used as input to a subsequent conversion process that produces output waveform audio” because *Fan*’s dialect conversion model outputs

a mel feature which is then subsequently input into Fan's neural network vocoder (which produces waveform audio). EX1004-Fan, [0126], Fig. 9.

128. *Fan's* mel feature output by the dialect conversion model teaches audio data *representative of the received speech content having the second accent* because the mel feature is not itself playable audio data but rather a representation of audio data (particularly in the frequency domain). Indeed, the '550 Patent describes the "audio data that is synthesized in this way may take the form of a set of mel spectrograms." '550 Patent, 8:3-4. This understanding is further confirmed by the '550 Patent because: (1) the claimed *synthesis* comprises mapping a non-text representation (which is the subject of Claim 1[e][i] below); and (2) the '550 Patent explains subsequent processing (i.e., after the claimed *second machine-learning-algorithm*) is used "to convert the synthesized audio data into waveforms that are able to be heard" (which is the subject of Claim 1[f] below). '550 Patent, 8:36-37.

7. Claim 1[e][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

129. *Fan* in view of *Loots* teaches *wherein synthesizing the fourth audio data* (e.g., *Fan's* mel feature with the second dialect) *comprises mapping* (e.g., *Loots's*

decision tree mapping) *at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content (e.g., Loots's "focus" phoneme from the string of source pronunciation input phonemes) to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content (e.g., Loots's "current" phoneme from the string of target pronunciation output phonemes), wherein the first and second phonemes are different phonemes (e.g., Loots's "focus" and "current" phonemes are different), as claimed.*

130. Claim 1[e][i] requires that synthesizing fourth audio data comprises mapping the phonemes of the received speech to target phonemes that would exist in the received speech if it were pronounced in the second accent. For example, the '550 Patent recognizes that the phoneme for the digraph "th" found in Standard American English (SAE) is replaced in Indian English with a "d" or "t" sound. EX1001, 1:52-59. Thus, if translating SAE to Indian English, the "th" phoneme would be mapped to a "d" or "t" sound.

a) Fan's Teachings

131. As discussed in Claim 1[d], *Fan* teaches producing a PPG feature of the first voice (non-text linguistic representation of a first phoneme of the set of

phonemes associated with the first pronunciation of the received speech content). *Fan* also teaches that the PPG feature and a timbre feature are input into the dialect conversion model, and the dialect conversion model outputs a mel feature with the content and speaking style of the first voice, but with a second accent. *See* Claim 1[e]. *Fan* does not describe all the details regarding the specific mapping of phonemes between its input and output pronunciations associated with particular accents.

132. As discussed earlier, *Sonsaat* teaches collecting data from a number of accents and maps acoustic data from speakers in a number of languages to phonemes, aligned so they can be used for accent detection and conversion. In doing so, *Sonsaat* observes some common phoneme substitutions in which phonemes from a native language are used instead of the correct phoneme in English:

Table 2: Most frequent errors by native language; the top-5 error occurrences are listed in descending order			
L1	Substitutions	Deletions	Additions
Hindi	DH→D, Z→S, W→V EY→EH, TH→T	R, D, T ER, HH	R, AH, S, Y AA
Korean	DH→D, Z→S, IH→IY OW→AO, EH→AE	D, T, R HH, K	AX, IH, AH, S Y
Mandarin	Z→S, DH→D, IH→IY N→NG, V→F	D, T, R L, N	AH, AX, IH N, R
Spanish	Z→S, IH→IY, DH→D AE→AA, AH→AO	D, T, AH Z, IH	EH, AX, AH IH, IY
Arabic	P→B, OW→AO R→ERR, DH→Z, Z→S	T, R, D AH, IH	G, AH, IH AX, EH

EX1005-Sonsaat, Table 2. This chart describes common phoneme errors in English from non-native speakers having the listed native language.

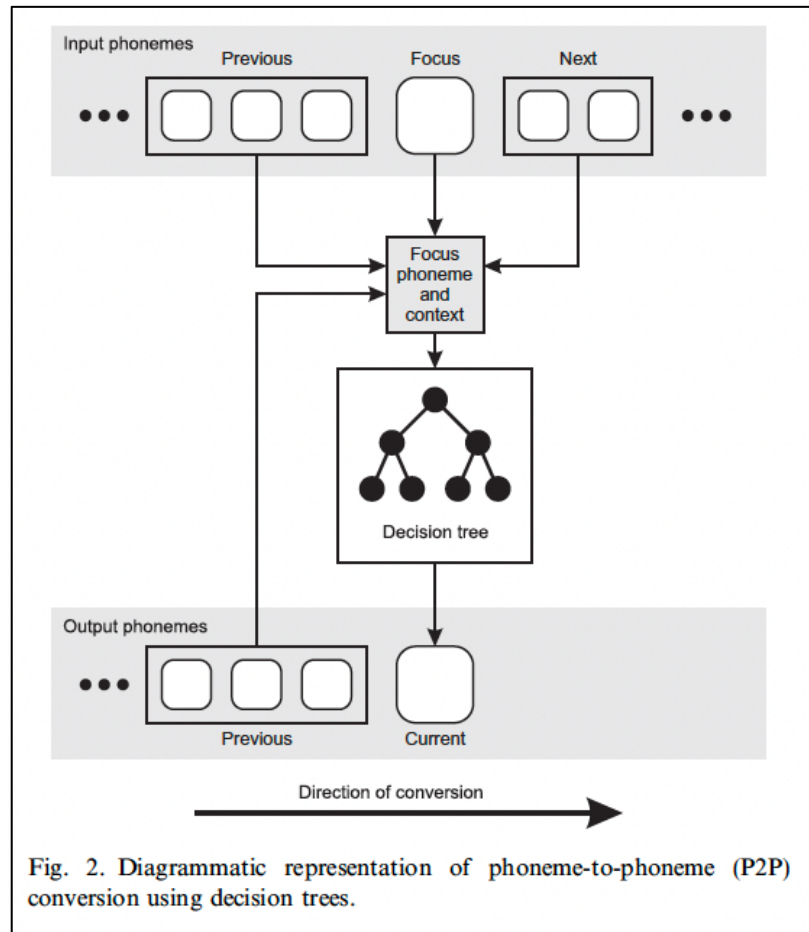
b) Loots's Teachings

Loots discloses phoneme-to-phoneme (“P2P”) conversion for words with different pronunciations in different accents. *Loots*, 1, Abstract. For example, *Loots* teaches phoneme differences in English words using four dictionaries covering three accents. EX1006-Loots, 2. Within the dictionaries “[t]here are a number of words in each dictionary with multiple pronunciations, either homographs (words with the same spelling but different pronunciations), or words with alternate pronunciations.” EX1006-Loots, 2. The “pronunciations of corresponding words in the three accents were aligned using a dynamic programming algorithm...the alignments will also be used to obtain the one-to-one correspondences between phonemes in different accents.” EX1006-Loots, 2. One example of an “alignment between SSAE and RP pronunciations of the word ‘reactions’” is shown below in Table 3 displaying that the “s” sound in the SAEDICT (South African English) dictionary maps to the “z” sound in the BEEP (Received Pronunciation, a British accent) dictionary:

Table 3 Two aligned pronunciations of the word “reactions”.								
SAEDICT	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	<i>ax</i>	<i>n</i>	<i>s</i>
BEEP	<i>r</i>	<i>ih</i>	<i>ae</i>	<i>k</i>	<i>sh</i>	–	<i>n</i>	<i>z</i>

EX1006-Loots, Table 3; *see* EX1006-Loots, 7 (discussing pronunciation differences for the word “spokesmen”).

133. To perform the conversion, *Loots* teaches “phoneme-to-phoneme (P2P) conversion using decision trees” to estimate pronunciation:



EX1006-Loots, 6, Fig. 2. Specifically, P2P conversion is “the conversion of one phoneme string (the source pronunciation) to another (the target pronunciation)[.]”

EX1006-Loots, 6.

134. As seen in *Loots*’s Fig. 2, a “focus” phoneme from the set of “[i]nput phonemes” associated with the string of source pronunciation phonemes (*a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content*) is mapped to a “current” phoneme which comes from the

set of “[o]utput phonemes” associated with the string of a target pronunciation phonemes (*a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content*). A PHOSITA would have appreciated that a “phoneme string (the source pronunciation)” of a word is made up of multiple phonemes. *Loots* teaches using a decision tree based algorithm to determine if a focus phoneme should be converted to a different phoneme in the output string. EX1006-Loots, 6, Fig. 2. *Loots* considers the “two phonemes to the left and two to the right of the focus phoneme” when deciding whether the focus phoneme should be converted. EX1006-Loots, 6. The decision tree based algorithm determines the target phoneme pronunciations based on the source phonemes to capture the pronunciation in the target accent.

135. Here, *Loots*’s current/output phonemes are an updated set of phonemes because they reflect the pronunciation of the speech in the target accent. EX1006-Loots, 6. As discussed, certain phonemes are pronounced differently between accents. As a result words containing those phonemes are pronounced differently (e.g., “reactions” in SAEDICT and BEEP or “tau nao” in Mandarin or Sichuanese). See Section IX.5.2. And *Loots*’s “focus” phoneme is different than *Loots*’s “current” phoneme in at least one case during accent conversion because accents have different sets of phonemes, and performing a P2P conversion would be unnecessary if each phoneme matched the proper phoneme in the target accent. A PHOSITA would have

understood that *Loots*'s decision tree would never be necessary unless at least one phoneme in the target accent varies from a phoneme in the input accent. For example, in converting a Sichuanese accent to Mandarin, [n] is replaced with [l]. See EX1039-Zhang, 3. In the *Fan-Loots* combination, the phonemes mapped (per *Loots*) is the PPG (phoneme posteriorgram) feature representing the phonetic information of each frame taught by *Fan*. A PHOSITA would have understood that the mapping of the PPG feature would be processed by the decision tree on a frame-by-frame basis, replacing the PPG for each frame that needs to be replaced with a PPG representing the target phoneme.

c) Motivation to Combine

136. A PHOSITA would have been motivated to incorporate *Loots*'s phoneme mapping functionality into *Fan*'s dialect conversion module with a reasonable expectation of success for several reasons.

137. First, although *Fan* discloses accent conversion including analysis of the original and target languages' phoneme structures to build PPGs on actual spoken training data, it does not disclose how to map phoneme translation if the PPGs comprise disparate phoneme sets. *Sansaat* discloses regular and predictable phoneme substitutions between phoneme sets in speech spoken by speakers with various accents. In light of these teachings, a PHOSITA would have looked to references that provide techniques for handling phoneme differences between

accents, such as *Loots*. *Loots* expressly discloses its P2P decision tree process improves mapping of accents, particularly for separate pronunciations between accents. EX1006-Loots, 8 (“it is clear that the pronunciation of a word in a new accent is in all cases substantially more accurately determined from its pronunciation in another accent by means of P2P conversion, than by means of G2P [grapheme-to-phoneme] conversion within the same accent.”), Abstract. Thus, a PHOSITA would have been motivated to implement *Loots*’s P2P decision tree into the dialect conversion model of *Fan* to improve the accent conversion.

138. Second, a PHOSITA would have understood that such a combination would have constituted the use of a known technique (e.g., P2P mapping to account for pronunciation differences, per *Loots*) to improve similar devices (e.g., machine-learning algorithms in voice conversion systems) in the same way (e.g., improved voice quality and system performance). A PHOSITA would have recognized that converting phonemes in an input accent to phonemes in the target accent would help with the accuracy of the output speech by ensuring the phonemes of the desired accent are present in the output speech when they differ from the phonemes of the input accent, and that P2P mapping as known in the art would have been a successful way to achieve accent and pronunciation conversion accuracy. It would be desired for the output voice with the converted accent to use the same phonemes as a speaker

of that accent to increase the accuracy of the converted speech when compared to natural speech of a speaker of the target accent.

139. A PHOSITA would have enjoyed a reasonable expectation of success in making such combination because the combination would have required only minor reprogramming of *Fan*'s software using the ordinary skill of a PHOSITA and hardware already present. *Fan* teaches that the dialect conversion model neural network includes functions such as a decoder, attention structure, and an encoder. EX1004-Fan, [0189], Fig. 9. *Fan* further teaches that "some or all functions of the processing apparatus may be implemented using software," where the processing apparatus is "[t]he processing module 1310" and configured to perform the dialect conversion. EX1004-Fan, [0173], [0148], [0150], [0162], [0165], [0167]. A PHOSITA would have understood how to incorporate additional features, such as a P2P decision tree, into the dialect conversion model because the model is an algorithm implemented in software that is easily modified to include additional algorithms. Modifying the algorithms used in the dialect conversion model would have been simple and a standard task for those with ordinary skill in the art.

8. Claim 1[f]: convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

140. *Fan* in view of *Loots* teaches *convert[ing] the synthesized fourth audio data (e.g., Fan's mel feature) into a synthesized version of the received speech content having the second accent (e.g., Fan's playable voice form of the input voice having a Mandarin accent), wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content (e.g., Fan's playable voice form contains Loots's output phonemes in the target accent), as claimed.*

141. The '550 Patent describes how “output speech generation engine 306 may convert the synthesized audio data into output speech, which may be a synthesized version of the received speech content 301 having the second accent.” EX1001, 5:45-52, 8:22-30 (describing how “the output speech generation engine 306 may take the form of a *vocoder* 306”).

142. *Fan* teaches that “the terminal inputs a PPG feature into a dialect switching model (a seq2seq neural network) based on the mode selection operation, and outputs *a mel[] feature form of the second voice* by using the dialect switching model. Then, a mel[] 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.*” EX1004-Fan, [0126], Fig. 9. *Fan* further teaches that “[t]he neural network vocoder is essentially a neural network, and is configured to convert an output feature of a voice processing neural network into a

highly natural voice.” EX1004-Fan, [0049]. A PHOSITA would have understood this disclosure describes the output mel feature being processed into playable audio. Thus, *Fan* teaches the mel feature (*synthesized fourth audio data*) is synthesized into the playable voice form having the second dialect (*synthesized version of the received speech content having the second accent*).

143. In the *Fan-Sonsaat-Loots* combination described above, *Fan*’s dialect conversion model outputs a mel feature which represents the input audio in a second dialect/accnt. This mel feature contains an updated set of target output phonemes from *Loots*’s P2P mapping process. Accordingly, *Fan*’s vocoder process which converts this mel feature into playable audio form comprises the updated set of target output phonemes.

B. Claim 2: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

144. *Fan* teaches wherein the program instructions are executable by the at least one processor to further cause the system (e.g., *Fan*’s program instructions executed by the processor) to apply, to the derived non-text linguistic representation (e.g., *Fan*’s PPG feature) of the received speech content having the first accent (e.g., *Fan*’s first voice input having a Sichuanese accent), a learned mapping (e.g., *Fan* applies the parameters learned by the dialect conversion NN model through training

to the PPG feature of the input voice) *between the second audio data comprising the first accent (e.g., Fan's "input" voice data) and the third audio data comprising the second accent (e.g., Fan's "label" voice data), as claimed.*

145. *Fan* teaches that the dialect conversion model is trained on two dialects, the input dialect and the label dialect, to determine parameters (*a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent*) that facilitate the accent conversion. See Claim 1[e]; EX1004-Fan, [0140] (“...a first dialect conversion model is converting Sichuanese into Mandarin, and each piece of sample data in a training dataset corresponding to the first dialect conversion model includes an input and a label. The input is Sichuanese, and the label is Mandarin with the same content as Sichuanese.”), [0141]. *Fan* further teaches that the training of the dialect conversion model updates the parameters of the model. EX1004-Fan, [0140] (“Different dialect conversion models correspond to different training datasets, and ***parameters at different layers are different.***”); EX1004-Fan, [0141] (“When the gradient flows through a corresponding structure (for example, an encoder, an attention structure, and a decoder), ***the parameter in the module is updated.***”). Although *Fan* uses the term “parameter,” a PHOSITA would have understood that internal values learned by a neural network as part of training are often referred to as weights in relevant literature. See Section V.A; see also EX1004-Fan, [0115] (“However, ***a weight***

parameter of each layer needs to be obtained through a large amount of data training.”)

146. In the *Fan-Loots* combination, the parameters reflect the desired phonetic changes (P2P conversion) taught by *Loots* as well as other linguistic differences learned through the training. A PHOSITA would have understood the parameters are responsible for the complete accent conversion, including all changes to the input speech present in the output speech in the second accent.

147. The PPG feature is input into the dialect conversion model (a seq2seq neural network) that outputs a mel feature of the input voice in a second accent based on the training. EX1004-Fan, [0126] (“...the terminal inputs a PPG feature into a dialect switching model (a seq2seq neural network) based on the mode selection operation, and outputs a mel[] feature form of the second voice by using the dialect switching model. ... The first voice is a voice of a first dialect, and the second voice is a voice of a second dialect”). A PHOSITA would have understood that the parameters determined in training would be applied to the PPG feature to produce the second (output) voice with a second accent. In the example of *Fan*, the parameters are applied to a PPG feature of an input voice with a Sichuanese dialect and produce a mel feature with a Mandarin dialect. Thus, the system of *Fan* applies the parameters to the PPG feature (*derived non-text linguistic representation of the received speech content having the first accent*) in the dialect conversion module.

C. Claim 3: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

148. *Fan teaches wherein the program instructions are executable by the at least one processor to further cause the system (e.g., Fan's program instructions executed by the processor) and a non-text linguistic representation (e.g., Fan's PPG feature) as claimed. Fan in view of Loots teaches to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content (e.g., Loots's decision tree maps the input phoneme of each frame of Fan's PPG feature associated with the source accent (Sichuanese) to an output phoneme associated with the target accent (Mandarin)) in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content (e.g., by mapping the input*

phoneme of each frame to the output frame, as taught by *Loots*, the PPG feature of *Fan* associated with phonemes representing the input accent is updated to a set of phonemes associated with the output accent), as claimed.

149. *Fan-Sonsaat-Loots* teaches using a decision tree incorporated into the dialect conversion model to map input phonemes of a PPG feature to phonemes used in the target accent. *See* Claim 1[e][i]. *Fan* further teaches that “[t]he phoneme posteriorgram is a posterior probability that **each voice frame** belongs to a group of predefined voice units (phoneme or triphoneme/assistant phoneme)[.]” EX1004-Fan, [0052]; *see also* Section V.C. A PHOSITA would have understood that because the PPG feature is a probability that each voice **frame** belongs to a particular phoneme, *Fan*’s processing (and the processing of the *Fan-Sonsaat-Loots* combination described above) performs the claimed mapping on a frame-by-frame basis. For example, each frame of each phoneme would be processed by the decision tree of *Loots* and updated to a corresponding frame of the proper phoneme representation in the PPG feature.

D. Claim 4: The system of claim 1, wherein the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent.

150. *Fan* in view of *Sonsaat* teaches *the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent* (e.g., *Sonsaat*’s training data comprises a plurality of speakers with a first accent).

151. *Fan* teaches that the dialect conversion model is trained with a training dataset that includes an input, voice data in a first accent (e.g., Sichuanese) and a label, voice data in a second accent (e.g., Mandarin). *See* Claim 1[e]. However, *Fan* alone does not expressly state that its training dataset comprises a plurality of speakers. It would have been advantageous to use multiple speakers for the first accent training data for the dialect conversion model for the same reasons it would be advantageous to do so in the feature extraction model. Namely, increasing the size of training data increases the accuracy of the dialect conversion network. Particularly, by using data of multiple speakers of a single accent, the model is trained to better understand the traits of an average speaker. This allows for better conversion of an any input voice in that accent.

152. A PHOSITA would have been motivated to use a dataset including multiple speakers having an accent as taught by *Sonsaat* as the input audio data to train the dialect conversion model of *Fan* to improve the accuracy and robustness of the model across speakers. A PHOSITA would have understood that every speaker has their own pronunciation and individualized accent quirks. *See* Section V.E. By using speech from multiple speakers with the same accent to train the dialect conversion model, individual accent quirks are minimized, improving the ability of the model to convert speech of any individual with that accent accurately. EX1035-Accentron, 1 (“The speech embeddings are extracted using an acoustic model trained

on a large corpus of native speech, so they are speaker-independent[.]”); *see also*, above Section V.E. (regarding idiolect, a dialect is an abstraction deriving from the analysis of a large number of individual speaker’s own unique linguistic system.). Further, a PHOSITA would have had a reasonable expectation of success, as this modification only requires gathering speech from an additional speaker and using that speech to train the dialect conversion model, both tasks a PHOSITA would have found to be simple. For example, *Sonsaat* teaches a corpus curation procedure that includes 10 speakers, with further plans to add 10 more speakers to the corpus. EX1005-Sonsaat, 2, 4. Furthermore, *Sonsaat*’s corpus curation procedure is an example of the simple process a PHOSITA would utilize to modify *Fan*. EX1005-Sonsaat, 2-3.

E. Claim 7

1. Claim 7[a]: The system of claim 1, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions are executable by the at least one processor to further cause the system to:

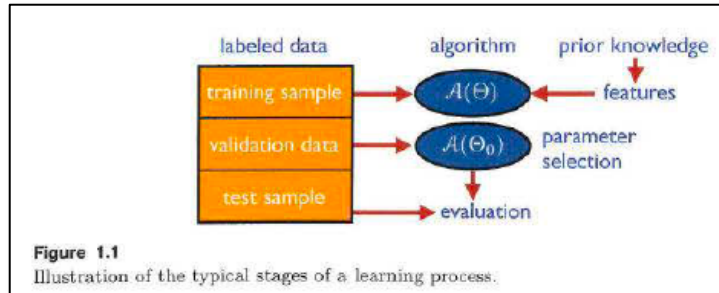
153. *Fan* in view of *Sonsaat* teaches *wherein the first machine-learning algorithm* (e.g., *Fan*’s voice feature extraction NN model) *comprises a non-text learned linguistic representation for the first accent* (e.g., *Fan*’s model is trained using *Sonsaat*’s corpus/method to extract a PPG feature for the first accent, Sichuanese, where such training results in learned parameters that become a part of

the voice feature extraction NN model) *and the program instructions are executable by the at least one processor* (e.g., *Fan's* program instructions executed by the processor), as claimed. *See* Claims 1[Pre]-1[d].

154. The '550 Patent discloses that the ASR engine “may be trained using captured speech content having a diversity of accents, which may enable the ASR engine 302 to develop a learned linguistic representation of not only the first accent, but also the second accent as well as additional accents.” EX1001, 6:41-46. Further, the learned linguistic representations “may be encoded as machine-readable data (e.g., a hidden representation)” and “may not be recognizable [] to a human.” EX1001, 6:62-67. The ASR engine “derive[s] a linguistic representation of the received speech content 301, based on the learned linguistic representation the ASR engine 302 has developed[.]” EX1001, 7:17-20. Thus, a PHOSITA would have understood a learned linguistic representation to be the learned parameters of a trained machine learning algorithm that permits the model to produce a linguistic representation. *See* Section V.A (discussing that machine learning algorithms are trained to determine the process by which given input data produces given output data).

155. The voice feature extraction model, trained in the manner suggested in Claim 1[c], comprises learned parameters to create a PPG feature for a Sichuanese accent and therefore comprises *a non-text learned linguistic representation for the*

first accent. For example, *Fan* recognizes the voice feature extraction model is “obtained by performing training in advance” and extracts a PPG feature of an input voice. EX1004-Fan, [0085], [0113]. The model is trained with voice from multiple speakers of the same accent as taught by *Sonsaat* (the model is *learned*) and the training allows it to generate a PPG feature from input speech of any accent the system is capable of converting. A PHOSITA would have understood “learning performance directly depends on the training sample EX1015-Mohri, 19. Labeled data is provided to the model (e.g., the voice feature extraction model) so that the model is trained to evaluate the output of the model against labeled validation data. EX1015-Mohri, 19. This is illustrated below:

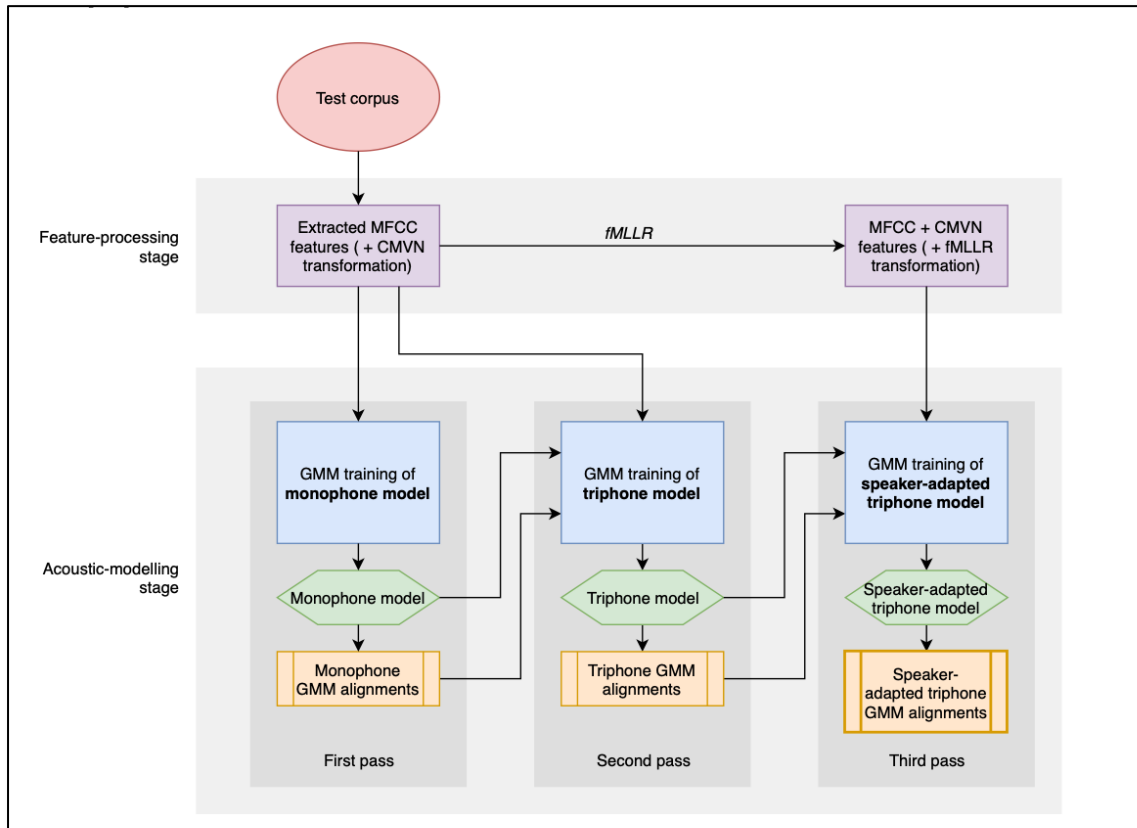


EX1015-Mohri, Fig. 1.1. For example, to train a model to produce PPG inputs to the NN are MFCC frames with target outputs of “senone labels obtained from force-alignment using an existing GMM-HMM speech recognizer.” EX1009-Osuna, 3. Thus, the learned parameters cause the voice feature extraction model, cause the model to create a PPG feature from an input voice.

2. Claim 7[b]: align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

156. *Sonsaat* teaches to align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content (e.g., *Sonsaat*'s Montreal forced-aligner aligns and classifies frames of input voice data according to phonemes and triphone sounds), as claimed. *Fan* teaches to train the first machine-learning algorithm (e.g., *Fan*'s training of the voice feature extraction NN model using input voice data aligned with *Sonsaat*'s forced-aligner), as claimed.

157. As explained above, the *Fan-Sonsaat* combination uses *Sonsaat*'s MFA to align and classify frames to train *Fan*'s feature extraction model. See Claim 1[c]. As seen in the figure below, the MFA process comprises and uses both a monophone model and a triphone model:



EX1013-Coles, Fig. 3.

158. *Sonsaat* shows that the output of the Montreal forced-aligner is a TextGrid with boundaries for words and phonemes; a PHOSITA would have understood that this includes the triphone alignments produced by the Montreal forced-aligner. A triphone is a series of three consecutive phonemes, and thus the series of phonemes output by the Montreal forced-aligner would be understood to classify both the phoneme alignment and the triphone alignment of the captured speech. Thus, *Sonsaat* teaches to align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content via *Sonsaat's* use of an MFA.

159. *Sonsaat* teaches aligning and classifying input speech data by words and phonemes using a Montreal forced-aligner. *See* Claim 1[c]; EX1005-Sonsaat, 3, Fig. 1. A PHOSITA would have understood that the Montreal forced-aligner also aligns and classifies frames of input speech by triphones. The Montreal forced-aligner aligns and classifies speech in three phases. First, the speech is aligned by phonemes (also called monophones). EX1013-Coles, (“It thus builds a monophone acoustic model for the data and then produces a first alignment.”); EX1014-McAuliffe, 1 (“To train a model, monophone GMMs are first iteratively trained and used to generate a basic alignment.”). Second, the speech is aligned by triphones. EX1013-Coles, (“The second pass takes into account the surrounding phones on either side of the phone in question, these three constituting a triphone, when building its HMMs. ...The resulting triphone acoustic model is used to produce a second alignment.”); EX1014-McAuliffe, 1 (“Triphone GMMs are then trained to take surrounding phonetic context into account, along with clustering of triphones to combat sparsity.”). Last, the speech is aligned by speaker-adapted triphones. EX1013-Coles, (“The third and las[t] pass also uses triphones, but on newly-computed features, transformed using Feature-Space Maximum Likelihood Linear Regression (fMLLR) [32].”); EX1014-McAuliffe, 1 (“The triphone models are used to generate alignments, which are then used for learning acoustic feature transforms on a per-speaker basis, in order to make the models more applicable to speakers in

other datasets [19]. ...This speaker-adapted triphone acoustic model is then used to produce the final alignment.”).

3. Claim 7[c]: determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

160. *Fan teaches determin[ing], for each of a second plurality of frames in the received speech content (e.g., Fan’s frames of the first input voice), a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent (e.g., Fan’s PPG feature determines the monophone and triphone sound for each frame based on the learned parameters of the trained voice feature extraction NN model).*

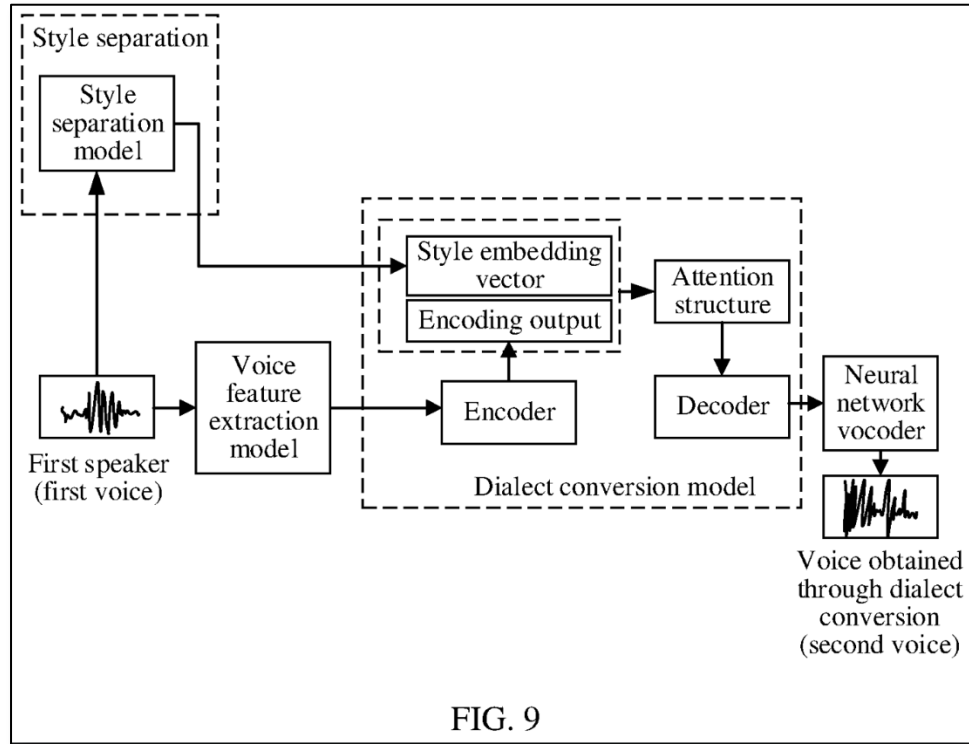
161. *Fan teaches that the PPG feature extracted by voice feature extraction model (comprising the non-text learned linguistic representation for the first accent) represents the probability that each frame of the input voice belongs to a particular “phoneme or triphoneme[.]” EX1004-Fan, [0007], [0052]. A PHOSITA would have understood that phonemes and monophones both refer to the same concept, a single phoneme. See Section V.B. Further, triphonemes and triphones similarly refer to the same concept, a sequence of three neighboring phonemes. See Section V.B. Further, a PHOSITA would have understood that a PPG feature that represents the probability each frame belongs to a particular triphone includes information of the*

particular phoneme the frame belongs to. *See* Section V.C; EX1009-Osuna 3 (describing that concatenating each speech frame with neighboring right and left frames during computation includes context). Thus, *Fan* teaches that both the monophone and triphones for each speech frame of input speech are determined by the voice feature extraction model in producing the PPG feature.

F. Claim 10: The system of claim 1, wherein the received speech content having the first accent further comprises a set of prosodic features, the program instructions are executable by the at least one processor to further cause the system to synthesize the fourth audio data representative of the received speech content having the second accent and the set of prosodic features, and the synthesized version of the received speech content having the second accent has the set of prosodic features.

162. *Fan* teaches wherein the received speech content having the first accent further comprises a set of prosodic features (e.g., *Fan*'s first input voice has a timbre including a prosody) the program instructions are executable by the at least one processor to further cause the system (e.g., *Fan*'s program instructions executed by the processor) to synthesize the fourth audio data representative of the received speech content having the second accent (e.g., *Fan*'s mel feature has a Mandarin accent) and the set of prosodic features (e.g., *Fan*'s timbre of the first voice including a prosody), and the synthesized version of the received speech content having the second accent has the set of prosodic features (e.g., *Fan*'s playable voice form contains the Mandarin accent and the timbre of the input voice including a prosody), as claimed.

163. *Fan* teaches that the style separation model that obtains a timbre feature that is input into the dialect conversion model retains the content and style of the input voice through accent conversion. EX1004-Fan, [0127] (“Then, the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model... . 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.”); *see* Claim 1[e]-[e][i]. This is illustrated in Fig. 9, which shows the first voice being input into the style separation model, which outputs a timbre feature vector to the dialect conversion model. *See* EX1004-Fan, [0127]; *see also* EX1004-Fan, [0132].



EX1004-Fan, Fig. 9.

164. 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.” EX1004-Fan, [0050]. Further, the timbre feature includes a plurality of sub-features including “a *prosody*, an accent, a speaking speed, and the like.” EX1004-Fan, [0088]. And a PHOSITA would have understood that prosody, as taught by *Fan*, includes multiple prosodic features such as intonation, stress, rhythm, and loudness. Similar to *Fan*, *Schmandt* teaches that intonation, stress, pitch, phrasing, and duration are all aspects of prosody. EX1025-Schmandt, 335-326; *see also* Section V.D. Furthermore, *Schmandt* teaches that prosody “refers to the spoken style of discourse independent of lexical content,” and “encompasses the majority of information lost when comparing an utterance to its transcription, and much of the richness of speech is conveyed by exactly this nonlexical means of expression.” EX1025-Schmandt, 325-326. Accordingly, A PHOSITA would understand that prosody encompasses elements of spoken language that are not encoded by lexical content, such as vocabulary. *See* Section V.D.

165. The input voice (*received speech content*) has timbre including a prosody (*set of prosodic features*) that is retained in the mel feature (*fourth audio data representative of the received speech content having the second accent*) output by the dialect conversion model and subsequently remains in the playable voice

feature output by the neural network vocoder (*synthesized version of the received speech content having the second accent*). EX1004-Fan, [0126]-[0127]. Claim 11

G. Claim 11

1. 11[Pre]: A non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device, cause the computing device to:

166. To the extent the preamble is limiting, *Fan* teaches a *non-transitory computer-readable medium provisioned with program instructions that, when executed by at least one processor of a computing device* (e.g., *Fan*'s memory 1520, processor 1580, program instructions, and terminal). See Claim 1[Pre]-[b].

167. *Fan* teaches the terminal “may be a mobile phone, a tablet computer, a notebook computer, a smartwatch, or the like.” EX1004-Fan, [0191]. A PHOSITA would have understood each of these to be a commonly known *computing device*. Further, each of these is a *computing device* as disclosed by the '550 Patent. EX1001, 3:8-14 (“As shown in FIG. 1, the computing device may include one or more processors 102, data storage 104, a communication interface 106, one or more input/output (I/O) interfaces 108, all of which may be communicatively linked by a communication link 110 that may take the form of a system bus, among other possibilities.”).

2. 11[a]: train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the

captured speech content corresponding to respective ones of the plurality of different speakers;

168. See Claim 1[c].

3. *11[b]: apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;*

169. See Claim 1[d].

4. *11[c]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent,*

170. See Claim 1[e].

5. *11[c][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and*

171. See Claim 1[e][i].

6. *11[d]: convert the synthesized fourth audio data into a synthesized version of the received speech content having the second*

accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

172. See Claim 1[f].

H. Claim 12: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to apply, to the derived non-text linguistic representation of the received speech content having the first accent, a learned mapping between the second audio data comprising the first accent and the third audio data comprising the second accent.

173. See Claims 2, 11.

I. Claim 13: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to map each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

174. See Claims 3, 11.

J. Claim 16

1. Claim 16[a]: The non-transitory computer-readable medium of claim 11, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the program instructions, when executed by the at least one processor, further cause the computing device to:

175. See Claims 11 [Pre], [b].

2. Claim 16[b]: align and classify each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and

176. See Claims 7[b], 11.

3. Claim 16[c]: determine, for each of a second plurality of frames in the received speech content, a respective (i) monophone and (ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

177. See Claims 7[c], 11.

K. Claim 19

1. 19[pre]: A method comprising:

178. To the extent the preamble is limiting, *Fan* teaches or suggests a *method* (e.g., *Fan*'s voice conversion method), as claimed.

179. *Fan*'s voice conversion method includes dialect conversion (i.e., “adding an accent or removing an accent” from human speech). EX1004-Fan, Abstract, (“A voice conversion method.... ...a dialect conversion mode, for adding an accent to or removing an accent from the first voice”), [0002], [0003], [0005] (“...an embodiment of this application provides a voice conversion method. ...For example, the method may include:...a dialect conversion mode...;”), [0006] (“...the dialect conversion mode, for adding an accent or removing an accent;”), [0144] (“The foregoing describes a voice conversion method.”), Figs. 10-12.

2. 19[a]: training a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of

different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

180. See Claim 1[c].

3. 19[b]: applying the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

181. See Claim 1[d].

4. 19[c]: based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesizing, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent,

182. See Claim 1[e].

5. 19[c][i]: wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

183. See Claim 1[e][i].

6. *19(d): converting the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.*

184. See Claim 1[f].

L. Claim 21: The method of claim 19, further comprising mapping each frame in the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to a corresponding frame in the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content in order to map the non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content to the non-text linguistic representation of the updated set of phonemes associated with the second pronunciation of the received speech content.

185. See Claims 3, 19.

M. Claim 22

1. *Claim 22[a]: The method of claim 19, wherein the first machine-learning algorithm comprises a non-text learned linguistic representation for the first accent and the method further comprises:*

186. See Claim 19[b].

2. *Claim 22[b]: aligning and classifying each of the first plurality of frames according to monophone and triphone sounds of the captured speech content to train the first machine-learning algorithm; and*

187. See Claims 7[b], 19.

3. *Claim 22[c]: determining, for each of a second plurality of frames in the received speech content, a respective (i) monophone and*

(ii) triphone sound detected in the frame based on the non-text learned linguistic representation for the first accent.

188. See Claims 7[c], 19.

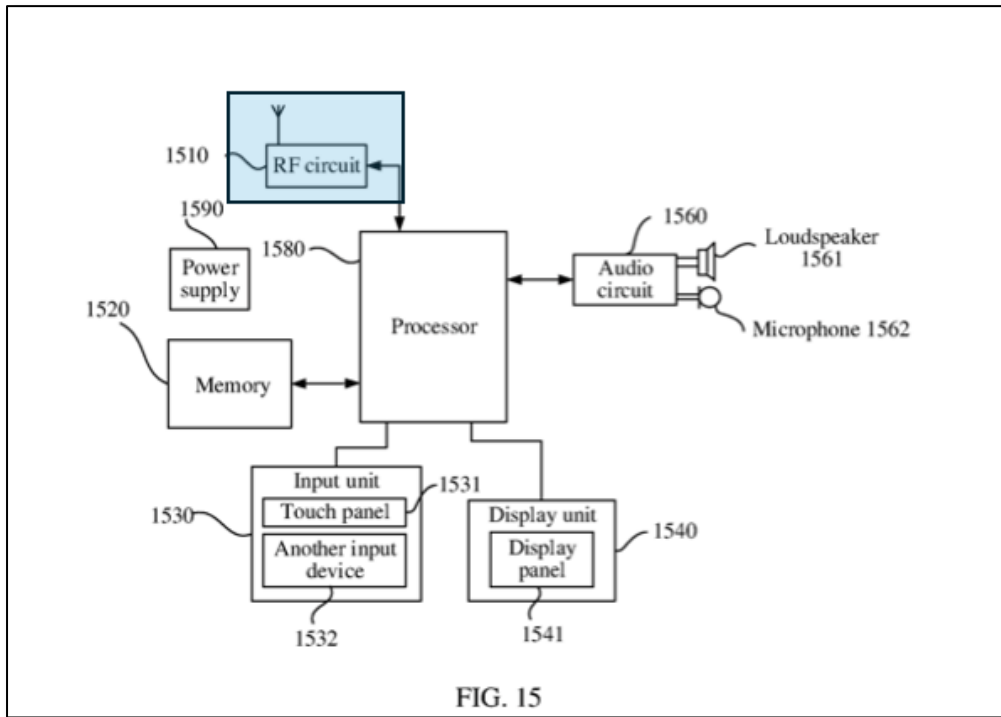
X. OPINIONS REGARDING GROUND II: *FAN*, *SONSAAT*, *LOOTS*, AND *FEINAUER*

A. Claim 8: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to transmit the synthesized version of the received speech content having the second accent to a second computing device.

189. *Fan* in view of *Feinauer* teaches wherein the program instructions are executable by the at least one processor to further cause the system (e.g., *Fan*'s program instructions executed by the processor) to transmit the synthesized version of the received speech content having the second accent to a second computing device (e.g., *Fan*'s synthesizing a first input voice and *Feinauer*'s sending modified speech content from a first to a second device), as claimed.

a) *Fan*'s Teachings

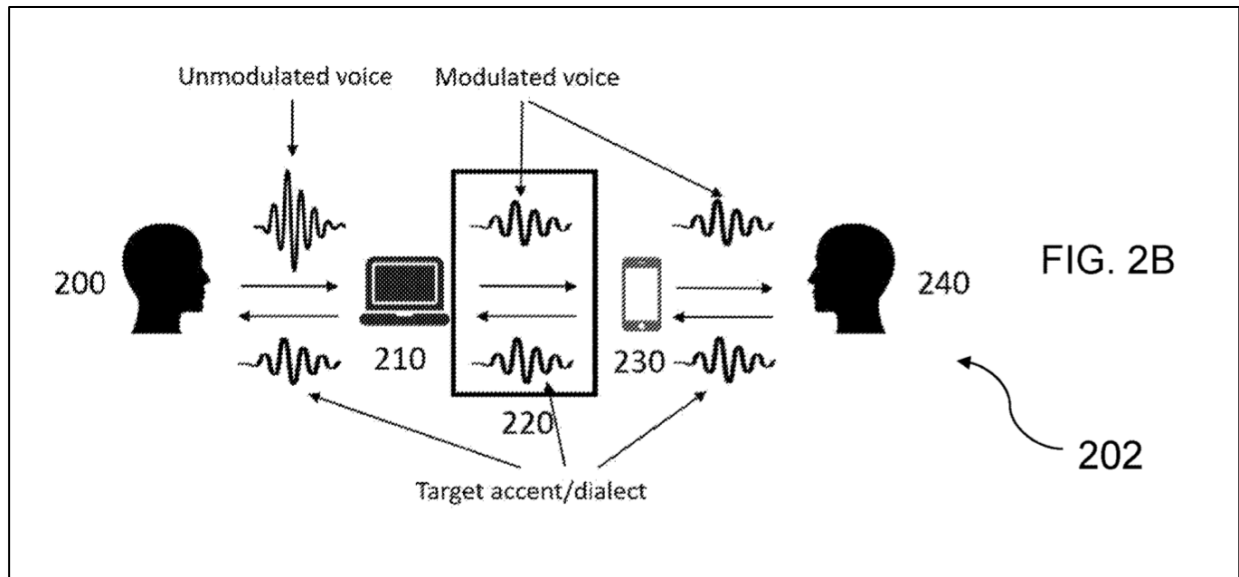
190. *Fan* teaches a system comprising a terminal with program instructions that are executable by the at least one processor. See Claim 1[Pre]. In one embodiment, the terminal is a mobile phone. EX1004-Fan, [0045]. The mobile phone includes a radio frequency circuit "configured to receive and send a signal in an information receiving and sending process or a call process. EX1004-Fan, [0193]-[0194].



EX1004-Fan, Fig. 15. Annotated to highlight RF circuit.

b) *Feinauer's* Teachings

191. *Feinauer* teaches an accent and dialect modifier for converting the accent of a speaker and transmitting the modified waveform via a communication network to another device. EX1008-Feinauer, [0046]. *Feinauer's* accent and dialect converter is a software program installed on a communication device that modifies a user's voice into a particular desired accent or dialect. EX1008-Feinauer, [0041]. Once the voice is converted, the "accent and dialect modifier 318 can transmit the modified data packet or packets in real-time, in a continuous stream, at a constant time interval or at a variable time interval." EX1008-Feinauer, [0056].



EX1008-Feinauer, Fig. 2B.

192. *Feinauer's* communication device 230 is a computing device 210 that includes an “audio transmitter 312” and “audio receiver 314” for sending and receiving voice and/or audio signals via communication network 220. EX1008-Feinauer, [0047], Fig. 3. Communication devices 210 and 230 are smartphones. EX1008-Feinauer, [0043], Fig. 3. *Feinauer* teaches that a first device modifies the accent of speech content before transmitting the modified speech content to the second device where it is provided to a user (i.e. *transmit the synthesized version of the received speech content having the second accent to a second computing device*). EX1008-Feinauer, [0046], Fig. 3.

c) Motivation to Combine

193. *Fan* teaches outputting a *synthesized version of the received speech content having the second accent*. *Fan* does not expressly provide all of the details of transmitting the output speech to a second computing device. But, *Feinauer* teaches *to transmit the synthesized version of the received speech content having the second accent to a second computing device*.

194. A PHOSITA would have been motivated to incorporate *Feinauer*'s transmission functionality into *Fan*'s terminal with a reasonable expectation of success for several reasons.

195. First, *Feinauer* expressly discloses converting an accent on a device, such as a cellular phone, and transmitting the modulated voice to a second device. EX1008-*Feinauer*, [0043], [0046], Fig. 2B. *Fan* teaches that its terminal may be a mobile phone. EX1004-*Fan*, [0045]. Because the accent conversion of *Fan* is performed on a mobile phone, a PHOSITA would have understood it would have been advantageous to use the capabilities of the mobile phone to transmit the output audio to other devices in the manner taught by *Feinauer*. A PHOSITA would have understood that a device such as the mobile phone of *Fan* would have similar capacities of the device of *Feinauer* and includes the necessary components for data transmission, such as a RF circuit. Using transmission capabilities to share data with other devices is one of the primary uses of mobile phones. Thus, a PHOSITA would

have been motivated to use the transmission capabilities of the mobile phone taught by *Feinauer* to transmit the modified speech of *Fan*.

196. Second, a PHOSITA would have understood that this modification would have constituted combining prior art elements (*Fan*'s mobile phone and accent conversion technology, and *Feinauer*'s audio transmission performed by a mobile phone) according to known methods (*Feinauer*'s signal transmission) to yield a predictable result (providing a user with capabilities to transmit modified speech). Transmitting voice data, both in real time and as stored audio files, was already widespread at the time of the invention. Cell phones were widely capable of sending audio files and using mobile VoIP as of the priority date of the '550 Patent. In fact, mobile VOIP has been understood for use in real-time and non real-time applications since at least 2002. *See* EX1040-Ayorkor, 1-2. A PHOSITA would have understood that mobile phones had the ability to send real-time and recorded audio to other devices through a range of networking technologies, such as mobile VOIP and cellular technologies (e.g., 3G, 4G, and 5G).

197. A PHOSITA would have had a reasonable expectation of success in making the combination. *Fan*'s RF circuitry is already capable of "send[ing] a signal in an information receiving and sending process or a call process." EX1004-Fan, [0194], Fig. 3. It was well known in the art that an RF circuit included both a receiver and a transmitter. A PHOSITA would have appreciated that *Fan*'s RF circuitry is

able to send and receive signals, therefore indicating *Fan's* RF circuitry includes both a receiver and a transmitter. Receivers and transmitters are a requirement for sending and receiving data wirelessly. *Fan's* discloses that “[t]he RF circuit 1510 may be configured to ***receive and send a signal*** in an information receiving and sending process or call process.” EX1004-*Fan*, [0194]. Thus, a PHOSITA would have understood *Fan's* RF circuit to include at least one receiver and at least one transmitter. A PHOSITA would have appreciated that *Feinauer* similarly teaches both an audio transmitter and audio receiver “for sending and receiving voice and/or audio signals via network 220, such as from communication device 230.” EX1008-*Feinauer*, [0047]; Fig. 3. Therefore, a PHOSITA would have appreciated that modifying *Fan* to transmit the synthesized version of the received speech content having the second accent to a second computing device (e.g., mobile phone) would not have required adding any additional hardware into *Fan's* mobile phone, as a mobile phone already contains all components necessary for wireless transmission. Any required changes would be minor changes to the software to allow the modified voice data to be sent to another device. These modifications merely allow the synthesized speech to be sent through the components of the mobile phone and received at another similar device.

B. Claim 9

1. Claim 9[a]: *The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to:*

198. *See Claim 1[Pre].*

2. Claim 9[b]: *receive, in real time, continuous speech content having the first accent; and continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.*

199. *Fan in view of Feinauer teaches to receive, in real time, continuous speech content having the first accent; and (e.g., Feinauer's processing continuous incoming speech content in real time, the speech content having a first accent) continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent (e.g., Fan's mel feature is converted into a playable voice form in the continuous manner of Feinauer) between 50-700 ms after receiving the continuous speech content having the first accent (e.g., Feinauer's entire voice conversion process occurs continuously in under 100 ms), as claimed.*

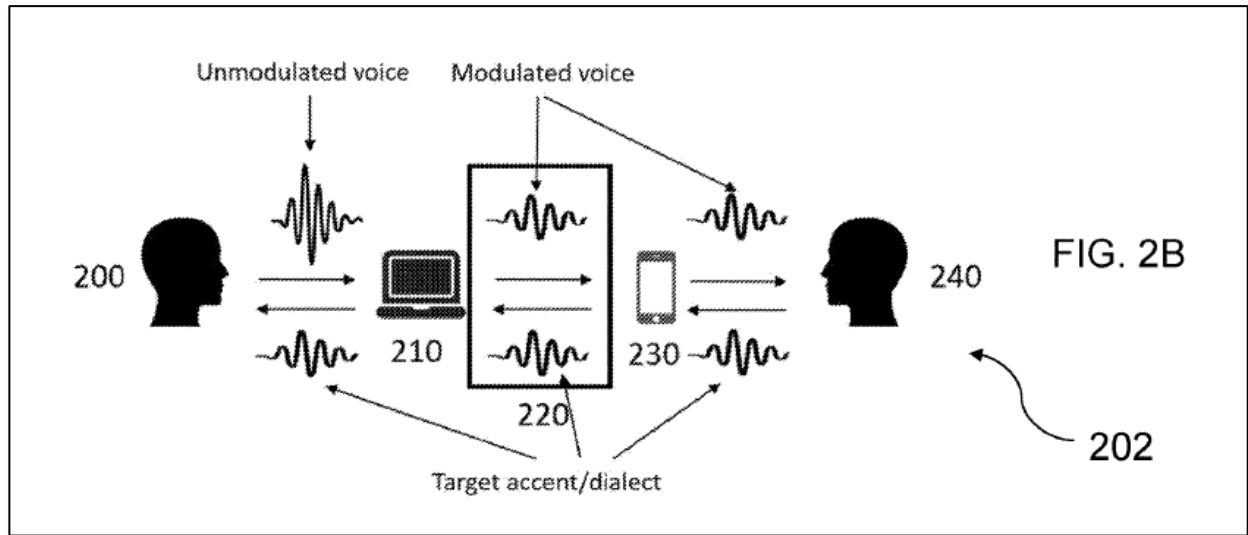
a) Fan's Teachings

200. *Fan teaches a first voice with a first (Sichuanese) accent. EX1004-Fan, [0129]-[0130]. Therefore, Fan teaches to receive speech content having the first accent.*

201. As discussed in Claims 1[e]-[e][i], *Fan* teaches converting a mel feature (*the synthesized fourth audio data*) into a synthesized version of speech content having the second accent with the use of a vocoder. *Fan* does not provide detail about performing the accent conversion in real-time, as claimed.

b) *Feinauer's* Teachings

202. *Feinauer* teaches *real-time* conversion of an accent in a speech audio signal that receives an unmodulated voice signal, processes it, and selects “a target dialect and accent to use to modify [the] voice communication based on a context.” EX1008-Feinauer, [0001]-[0003], [0046], [0049]. *Feinauer* uses one or more neural network machine learning models to perform the accent modification. EX1008-Feinauer, [0036]-[0037], [0039], [0098]. *Feinauer's* accent conversion method and system 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 (*continuous speech content*). EX1008-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.”



EX1008-Feinauer, [0035]-[0036], [0041], Fig. 2B. *Feinauer* explains its process may be performed within a “lag duration...that sounds like ***real-time*** to a user (***e.g. 100ms***).” EX1008-Feinauer, [0086]. *Feinauer* teaches “accent and dialect modifier 318 can receive...data packets...representing or containing audio of conversation(s), such as audio fragments.” EX1008-Feinauer, [0053]. “The accent and dialect modifier 318 can receive, retrieve, accept, or intercept one data packet and process the data packet before processing the next data packet.” EX1008-Feinauer, [0056]. As a result, “[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.” EX1008-Feinauer, [0057]. Thus, *Feinauer* teaches an accent conversion model that receives data packets of audio content with a first accent during an audio call (*receive, in real*

time, continuous speech content having the first accent). After modifying the dialect of the speech in the data packets, the accent and dialect modifier 318 “store[s] the processed data packet in memory or a buffer for transmission.” EX1008-Feinauer, [0056]. Subsequently, the “accent and dialect modifier 318 can transmit the modified data packet or packets in real-time, in a ***continuous*** stream, at a constant time interval or at a variable time interval.” EX1008-Feinauer, [0056].

c) Motivation to Combine

203. *Fan* teaches accent conversion. But *Fan* does not expressly provide all of the details of performing accent conversion in real-time. *Feinauer* teaches performing an accent conversion in real-time. The *Fan-Feinauer* combination performs the process of *Fan* on data packets of voice speech to perform accent conversion in real-time as taught by *Feinauer*. Namely, in the proposed combination, the dialect conversion model of *Fan* is modified to continuously receive and modify voice data packets to allow for real-time dialect conversion. Additionally, the neural network vocoder of *Fan* continuously converts the mel features representing voice data packets produced by the dialect conversion model into a playable voice form (*continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent*). See Claim 1[f]. The playable voice form is stored in memory or a buffer as taught by *Feinauer* until being transmitted. EX1008-Feinauer, [0056]. Finally, the entire *Fan-Feinauer* process

occurs in under 100 ms as taught by *Feinauer* (*between 50-700 ms after receiving the continuous speech content having the first accent*). EX1008-Feinauer, [0086].

204. A PHOSITA would have been motivated to incorporate *Feinauer's* real-time functionality into *Fan's* terminal with a reasonable expectation of success for several reasons.

205. First, *Feinauer* expressly discloses real-time accent conversion on a neural network accent conversion system. EX1008-Feinauer, [0064]-[0067], [0086], Fig. 3. Additionally, *Feinauer* is implemented on a phone, and allows a user to translate speech during a phone call. *Fan* is a similar neural network conversion system implemented on a phone, and a PHOSITA would have been motivated in light of *Feinauer's* teachings to perform *Fan's* accent conversion in real-time. Performing real-time accent conversion would allow the invention of *Fan* to take advantage of the capabilities of a mobile phone, particularly allowing a user to have a phone call where the accent of a user is converted so it can be easier understood by other parties. See EX1008-Feinauer, [0003] (“For example, in a phone conversation, comprehending a person at the other end of a call who has a difficult to understand accent and/or dialect can pose a challenge. The present disclosure is directed to identifying an optimal or desired dialect and accent to use to achieve a desired or target goal in different scenarios and modifying vocal characteristics of

audio communication of a user in real-time to more closely resemble the identified optimal dialect and/or accent.”).

206. Second, such a modification constitutes the use of a known technique (performing accent conversion in real-time) to improve similar devices (mobile device with accent conversion functionality) in the same way by allowing a user to convert an accent in real-time, such as during a phone call. For many useful applications of accent conversion functionality, such as during a phone call or teleconference, real-time accent conversion is necessary. Lag from converting an entire sentence structure of speech delays the output of audio data, and this delay can hinder communication between parties. For example, if the output speech is delayed enough to produce a pause, the other party may begin to speak, causing overlap in speech and preventing effective communication between the parties. Receiving audio packets continuously and performing real-time processing, as taught by *Feinauer*, was a common technique in the art. EX1008-Feinaur, [0053], [0056]. For example, *Ayorkor* teaches that VoIP applications transmit real-time voice data over IP networks that are packet-based when using the H.323 protocol stack. EX1040-Ayorkor, 2. Thus, a PHOSITA would have been motivated to modify *Fan*'s accent conversion process to output modified speech within 100ms, as taught by *Feinauer*, to allow for effective real-time communication. A low latency, such as 100ms, would make the output speech viable for real-time applications, such as in

phone or conference calls. As discussed, *Feinauer* recognizes that processing large volumes of speech, such as a full sentence, risks latency that prevents effective real-time communication. EX1008-Feinauer, [0038]. Further, the '550 Patent similarly recognizes the necessity of processing intervals of speech content for real-time functionality:

Further, the accent-conversion application 203 may process incoming speech content as it is received, making it capable of handling both extended periods of speech as well as frequent stops and starts that may be associated with some conversations. For example, in some embodiments, the accent-conversion application 203 ***may process incoming speech content every 160 ms***. In other embodiments, the accent-conversion application 203 may process the incoming speech content more frequently (*e.g., every 80 ms*) or less frequently (*e.g., every 300 ms*).

EX1001, 5:31-40. '550 Patent claims this allows the system to operate “at speeds that enable real-time communications, ***having a latency as low as 50-700 ms (e.g., 200 ms)*** from the time the input speech received by the accent-conversion application 203 to the time the converted speech content is provided to the communication application[.]” EX1001, 5:24-30. As such, a PHOSITA would have been motivated to process the audio data of *Fan* in packets.

207. Last, the combination would have had a reasonable expectation of success because the combination only requires simple changes to *Fan*'s software

that would have been well within the ordinary skill of a PHOSITA. As discussed in Claim 8, *Fan* and *Feinauer* teach similar hardware and software (mobile phones equipped with voice conversion software). Further, the accent conversion of *Fan* and *Feinauer* each utilize neural networks to convert input voice into a second accent. *Fan* uses a voice feature extraction model, a dialect conversion model, and a neural network vocoder, while *Feinauer* uses an accent detection model, an accent translation model, and an optimal target accent model. EX1004-Fan, [0113], [0126]-[0127], Fig. 9; EX1008-Feinauer, [0036]. Additionally, a PHOSITA would have understood that voice data is commonly processed in discrete packets, as taught by *Feinauer*, and that the process of *Fan* would be applied to smaller data packets to achieve real-time accent conversion.

208. Nothing in *Fan* would have suggested to a PHOSITA that the input speech must be processed as a whole. Further, the dialect conversion model of *Fan* is a seq2seq neural network. EX1004-Fan, [0140]. A seq2seq neural network compresses an input sequence into a vector using an encoder and then converts that vector into an output sequence. EX1004-Fan, [0047]. *Fan* teaches that the PPG feature is created on a frame-by-frame basis. EX1004-Fan, [0067] (“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.”). Because the dialect conversion model is based around receiving and processing the PPG feature, the

entire system taught by *Fan* is linked to the time to perform that translation of a PPG feature to a mel feature in the second accent. Thus, the dialect conversion model is technically capable of processing a sequence of one or more frames represented in the PPG feature at a time. A PHOSITA would have understood that by limiting the data in the income sequence provided to the dialect conversion model, data is output in smaller sequences, decreasing the system lag. As described by *Feinauer*, this allows the model to convert the accent of audio on a per sound or per word basis within a delay period that is perceived as real-time. *See* EX1008-Feinauer, [0057]; *see generally* EX1041-Umeda; *see generally* EX1042-UmedaVowels.

209. Any additional necessary changes, such as providing input and target accent selection at the beginning of the call, would have also been understood to be simple to a PHOSITA. For example, while *Fan* teaches selecting the target and input accents for a recorded audio file, *Feinauer* teaches making these selections at the start of a call. *Compare* EX1004-Fan, [0121]-[0124], Figs. 8, 11A-11B; *with* EX1008-Feinauer, [0064] (“In some implementations, the user 200 can bypass using a detection model 332 by selecting the dialect or accent of the call participant 240 to whom the user is speaking.”). The technical changes to combine *Fan* and *Feinauer*, such as changing the timing of the accent selection, are simple programming changes easily made by a PHOSITA. Thus, a PHOSITA would have expected success in

modifying *Fan*, consistent with *Feinauer's* teachings, to perform accent conversion in real-time on voice data.

C. Claim 17: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to transmit the synthesized version of the received speech content having the second accent to a second computing device.

210. See Claims 8, 11.

D. Claim 18

1. Claim 18[a]: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to:

211. See Claim 11[Pre].

2. Claim 18[b]: receive, in real time, continuous speech content having the first accent; and

212. See Claims 9[b], 11.

3. Claim 18[c] continuously convert the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.

213. See Claims 9[c], 11.

E. Claim 20

1. Claim 20[a]: The method of claim 19, further comprising:

214. See Claim 19[Pre].

2. *Claim 20[b]: receiving, in real time, continuous speech content having the first accent; and*

215. See Claims 9[b], 19.

3. *Claim 20[c] continuously converting the synthesized fourth audio data into a synthesized version of the continuous speech content having the second accent between 50-700 ms after receiving the continuous speech content having the first accent.*

216. See Claims 9[c], 19.

XI. OPINIONS REGARDING GROUND III: *FAN*, *SONSAAT*, *LOOTS*, AND *OSUNA*

A. Claim 5: The system of claim 1, wherein the third audio data comprising the second accent corresponds to a single speaker having the second accent.

217. *Fan* in view of *Osuna* teaches the third audio data comprising the second accent corresponds to a single speaker having the second accent (e.g., *Osuna*'s using a single speaker with a target accent to train *Fan*'s dialect conversion NN model), as claimed.

a) *Fan*'s Teachings

218. *Fan* teaches that the dialect conversion model is trained with a training dataset that consists of an input (speech in a source accent) and a label (speech in a target accent), each of which is voice data. See Claim 1[e]. However, *Fan* does not specify if the label voice data is from a single or plurality of speakers with a second accent.

b) *Osuna's Teachings*

219. *Osuna* teaches “accent conversion” using a neural network to perform non-native to native accent conversion. EX1009-*Osuna*, 1. Specifically, the neural network is trained on speech data from a single non-native speaker and a single native speaker to perform accent conversion on speech from the non-native speaker. EX1009-*Osuna*, 7. Similar to *Fan*, *Osuna* utilizes PPGs to perform the accent conversion. EX1009-*Osuna*, 1-4.

c) Motivation to Combine

220. *Fan* teaches a trained second machine-learning algorithm. But *Fan* does not expressly provide that the speech data used to train this algorithm is from a single speaker with a second accent. However, *Osuna* does teach *the third audio data comprising the second accent corresponds to a single speaker having the second accent*.

221. A PHOSITA would have been motivated to incorporate *Osuna's* training voice data into *Fan's* training of its accent conversion module with a reasonable expectation of success for several reasons.

222. First, a PHOSITA would have been motivated to make this modification as *Fan* expressly discloses that the dialect conversion models are trained on a training dataset of voice data, but does not describe the low-level implementation details of the makeup of this training dataset. Therefore, a

PHOSITA would have been motivated to look to references such as *Osuna* for potential sets of training data. EX1004-Fan, [0140]. A PHOSITA would have understood that sets of training data and methods for creating them were well known and would have looked for references outlining these sets and methods for the specific model functionality desired.

223. Second, a PHOSITA would have understood that a training set must consist of voice data from either a single speaker or a plurality of speakers with the target accent. Each of these training sets is capable of being used to train an accent conversion model depending on the desired functionality of the model. For example, voice data from a plurality of speakers is used when a generalized output representing an average speaker of an accent is desired. Alternatively, voice data from a particular speaker is used where the exact accent of said speaker is desired. For example, a user may wish to imitate the accent of a celebrity, or a particular speaker with a unique accent. Another benefit of using a training data that has speech of a single user is that it reduces the difficulty in acquiring training data. It is less time-consuming and easier to obtain voice data from a single speaker of an accent compared to a plurality, reducing the difficulty in training the model. A PHOSITA would have understood that finding a group of individuals with a single accent and obtaining hours of recordings from each individual is costly and time-consuming. A single speaker with an exemplary accent would require less total recording time and

thus be less expensive and take less time. Thus, using training data of a single speaker to train the accent conversion model constitutes the use of a known technique to improve a similar device (machine-learning algorithms in voice conversion systems) in the same way.

224. Finally, a PHOSITA would have had a reasonable expectation of success in making the modification, as this change does not require any hardware or software changes. It merely requires inputting training voice data from a single speaker instead of a plurality of speakers as the label, a task a PHOSITA would have found to be commonplace and simple. As discussed, a PHOSITA would have understood how to implement different data sets in training to achieve different model outputs.

B. Claim 14: The non-transitory computer-readable medium of claim 11, wherein one or more of the second audio data comprising the first accent corresponds to a plurality of speakers having the first accent or the third audio data comprising the second accent corresponds to a single speaker having the second accent.

225. *See* Claims 4-5, 11.

XII. OPINIONS REGARDING GROUND IV: *FAN*, *SONSAAT*, *LOOTS*, AND *DALCE*

A. Claim 6: The system of claim 1, wherein the program instructions are executable by the at least one processor to further cause the system to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

226. *Fan* in view of *Dalce* teaches wherein the program instructions are executable by the at least one processor to further cause the system (e.g., *Fan*'s program instructions executed by the processor) to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent (e.g., *Dalce*'s selection of a speakers dialect and a target dialect), as claimed.

a) *Fan*'s Teachings

227. *Fan* teaches that “[a] list of the plurality of modes may be displayed in a display interface of the terminal.” EX1004-Fan, [0058]. “A user selects a target conversion mode in the plurality of modes by performing a tap operation based on a requirement.” *Id.* A user selects “Remove accent” or “Add accent” as well as which accent the user would like to add (e.g., “Shanghai accent”, “Wuhan accent”, “Guangdong accent”, “Sichuan accent”, and “Fujian accent”). EX1004-Fan, [0124], Figs. 11A-B. A terminal displaying these options is shown in Fig. 11A.

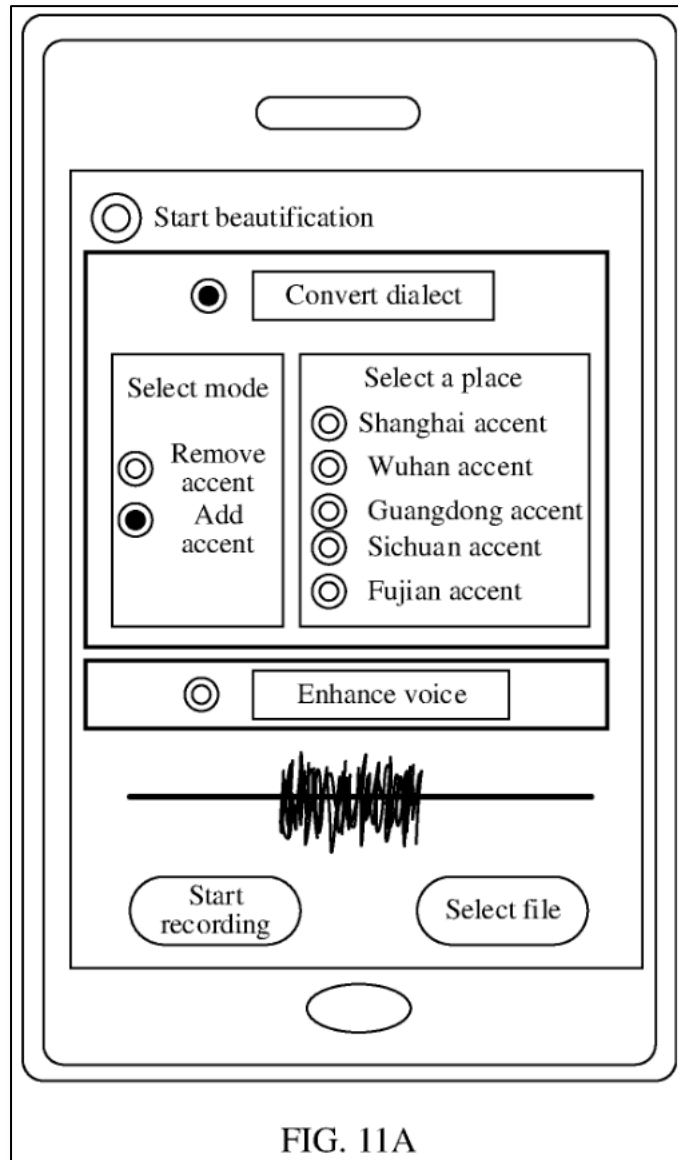


FIG. 11A

EX1004-Fan, Fig. 11A. *Fan* discloses that in its disclosure, Mandarin is the base accent, as in, when accent removal is selected, the accent of the input voice is converted into Mandarin. EX1004-Fan, [0120] (“A first type is removing an accent, that is, converting a dialect into Mandarin. For example, Sichuanese is converted into Mandarin.”). When selecting the “Add accent” option and the target accent (the

options shown under “Select a place”) the user is choosing to convert the input accent (the base accent Mandarin) to the output accent (for example, Sichuanese). EX1004-Fan [0120] (“A second type is adding an accent, that is, converting Mandarin into a dialect. For example, Mandarin is converted into Sichuanese.”), Figs. 11A-B. A PHOSITA would have understood that these options provide the user with the ability to specify both the input and output accents. Further, a PHOSITA would have further understood that since *Fan* does not disclose any accent/dialect recognition software or components, the user selection must specify both the input and output accents so that the correct dialect conversion model is selected for the accent conversion process. EX1004-Fan, [0120]. For example, the Thus, *Fan* teaches selecting “Add accent” to specify the input accent is a Mandarin accent (*receive a first user input indicating a selection of the first accent*) and making a second selection from the “select a place” menu to choose the output accent (*a second user input indicating a selection of the second accent*).

b) *Dalce’s Teachings*

228. To the extent it is argued *Fan* does not teach both a user selection of the first accent and a user selection of the second accent, *Dalce* teaches these elements. *Dalce* teaches a system for dialect conversion between a speaker dialect and a target dialect. EX1007-Dalce, [0013], [0015]-[0016]. *Dalce* teaches that the system allows “a user to manually select a ***speaker dialect***.” EX1007-Dalce, [0013]; *see id.*, Fig.

2A. The system also permits “selection of a target dialect.” EX1007-Dalce, [0015]. “The speaker’s words are translated by a translation model” into words in the target dialect. EX1007-Dalce, [0016]. Thus, *Dalce* teaches *to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.*

c) Motivation to Combine

229. *Fan* teaches selecting to add or remove an accent, and which accent to add or remove. *Fan* does not disclose directly selecting the first accent and second accent. However, *Dalce* does teach *the computing device to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.* The *Fan-Dalce* combination utilizes the dialect conversion of *Fan* with the user selections of *Dalce*.

230. A PHOSITA would have been motivated to incorporate *Dalce*’s selection of speaker and target dialect into *Fan*’s dialect conversion system with a reasonable expectation of success for several reasons.

231. A PHOSITA would have been motivated to make this change to improve the user experience by improving the user interface. *Fan* discloses its selections allow the terminal to select the corresponding dialect conversion model. EX1004-Fan, [0124]. However, *Fan*’s system assumes the user knows the standard base accent (e.g., Mandarin), which may not always be the case. In an alternate

example, a speaker of British English and a speaker of American English would both be likely to assume their accent is the standard accent of the language. As a result, it may be confusing for a user of *Fan* to determine whether their accent needs to be removed or if the target accent should be added, which could lead to the selection of the incorrect dialect conversion model. A PHOSITA would have understood that this system is more prone to error than providing a user the option to select the input accent and the output accent. Further, *Fan* discloses that in an optional solution two dialect conversion models are jointly used to perform conversion between two non-mandarin dialects. EX1004-Fan, [0120] (“Certainly, in an optional solution, the two dialect conversion models may also be jointly used, to perform conversion between the two dialects. For example, Sichuanese is converted into Cantonese. Sichuanese can be converted into Mandarin first, and then Mandarin is converted into Cantonese.”). A PHOSITA would have understood the need for providing a user the option to select the input and target dialect in this scenario, which is not provided by exemplary interface. The exemplary interface currently does not provide options to simultaneously remove and add an accent. *See* EX1004-Fan, Fig. 11A-B. Thus, a PHOSITA would be motivated by *Dalce* to provide an interface that allows a user to directly select the first and second accent.

232. Further, a PHOSITA would have understood that this modification is nothing more than combining prior art elements (*Fan*’s accent conversion and

Dalce's user selection of the input and target dialect) to yield predictable results of improving the usability of the accent conversion system. For example, if a user with a non-Mandarin accent wants to convert their accent to a different non-Mandarin accent, the limited options require removing a first operation to remove their accent, and a second operation to add the desired accent. *See* EX1004-Fan, [0120], [0124]. It would be an improvement if the user simply identifies the input accent and the target accent to allow the system to select the appropriate dialect conversion model. As specified above, if the system has models to convert between two non-Mandarin dialects as *Fan* envisions, a user interface that allows a user to select those dialects would provide the ability for a user to easily use that functionality, improving usability of the system.

233. Finally, a PHOSITA would have had a reasonable expectation of success in implementing the selections of *Dalce* into *Fan*, since all that is required is simple programming changes to the user interface of *Fan* that are well within the skill level of a PHOSITA. A user interface that provides options to a user and ties user selections to program functionality was a simple programming change for a PHOSITA to make. The accent conversion functionality of *Fan* is not affected by this change because the system is already able to convert between any accents it has a dialect conversion model for.

B. Claim 15: The non-transitory computer-readable medium of claim 11, wherein the program instructions, when executed by the at least one processor, further cause the computing device to receive a first user input indicating a selection of the first accent and a second user input indicating a selection of the second accent.

234. *See* Claims 6, 11.

XIII. CONCLUSION

235. 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 27, 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 (AUIC) 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.