

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.
Petitioner

v.

Sanas.ai Inc.
Patent Owner

Post Grant Review Case No. PGR2026-00033
U.S. Patent No. 12,417,756

DECLARATION OF CHRISTOPHER SCHMANDT

TABLE OF CONTENTS

I.	INTRODUCTION.....	14
A.	Educational Background and Professional Experience	14
II.	METHODOLOGY AND MATERIALS CONSIDERED	18
III.	OVERVIEW AND LEGAL STANDARDS	22
A.	Person of Ordinary Skill in the Art.....	22
B.	Obviousness	23
C.	Analogous Art	29
D.	Claim Construction.....	29
IV.	LEVEL OF A PERSON OF ORDINARY SKILL.....	30
V.	OVERVIEW OF TECHNOLOGY	31
A.	History of Speech Processing	31
B.	Vocal Features.....	32
C.	Encoding	34
D.	Feature Extraction	35
E.	Audio Interface.....	36
F.	Machine Learning Model.....	38
VI.	OVERVIEW OF THE '756 PATENT.....	39
A.	Description of the '756 Patent.....	39
B.	Field of Endeavor of the '756 Patent.....	40
C.	Problem Purportedly Solved by the Inventors of the '756 Patent.....	40

D.	Prosecution History of the '756 Patent	41
VII.	OVERVIEW OF THE PRIOR ART	42
A.	Overview of <i>Nguyen</i>	43
B.	Overview of <i>Dirac</i>	45
C.	Overview of <i>Narayanan</i>	46
D.	Overview of <i>Kamruzzaman</i>	48
E.	Overview of <i>Kee</i>	49
F.	Overview of <i>Wang</i>	50
VIII.	SUMMARY OF UNPATENTABILITY	51
IX.	OPINIONS REGARDING GROUND I: <i>NGUYEN, KEE, DIRAC, AND NARAYANAN</i>	53
A.	Claim 1	53
1.	<i>Claim 1[Pre]: “A speech processing system, comprising an audio interface coupled to a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to:”</i>	<i>53</i>
2.	<i>Claim 1[a]: “apply one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user;”</i>	<i>62</i>
3.	<i>Claim 1[b]: “analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns;”</i>	<i>73</i>

4.	<i>Claim 1[c]: “synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent; and”</i>	86
5.	<i>Claim 1[d]: “provide to the audio interface output audio data for output via the audio output device, wherein the output audio data is generated based on the modified version of the second input speech.”</i>	91
B.	Claim 2	92
1.	<i>Claim 2: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.”</i>	92
C.	Claim 3	94
1.	<i>Claim 3: “The speech processing system of claim 1, wherein the accent features comprise one or more pitch contours, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”</i>	94
D.	Claim 6	96
1.	<i>Claim 6: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”</i>	96

X.	OPINIONS REGARDING GROUND II: NGUYEN, KEE, DIRAC, NARAYANAN, AND KAMRUZZAMAN.....	99
A.	Claim 4	99
1.	<i>Claim 4: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.”</i>	<i>99</i>
XI.	OPINIONS REGARDING GROUND III: NGUYEN, KEE, DIRAC, NARAYANAN, AND WANG.....	107
A.	Claim 5	107
1.	<i>Claim 5: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.”</i>	<i>107</i>
XII.	OPINIONS REGARDING GROUND IV: NGUYEN, DIRAC, NARAYANAN, AND KAMRUZZAMAN.....	114
A.	Claim 7	114
1.	<i>Claim 7[Pre]: “A method for real-time accent mimicking, the method implemented by a speech processing system and comprising:”</i>	<i>114</i>
2.	<i>Claim 7[a]: “applying one or more trained machine learning models to first input audio data to extract accent features of first input speech associated with a first accent of a first user;” ..</i>	<i>116</i>
3.	<i>Claim 7[b]: “analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated</i>	

	<i>characteristics comprise a unique fingerprint of the voice of the second user;”</i>	116
4.	<i>Claim 7[c]: “synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and”</i>	117
5.	<i>Claim 7[d]: “providing output audio data generated based on the modified version of the second input speech.”</i>	117
B.	Claim 8	117
1.	<i>Claim 8: “The method of claim 7, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.”</i>	117
C.	Claim 9	117
1.	<i>Claim 9: “The method of claim 7, further comprising extracting from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.”</i>	117
D.	Claim 10	117
1.	<i>Claim 10: “The method of claim 7, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech. The intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”</i>	117
E.	Claim 11	118
1.	<i>Claim 11: “The method of claim 7, further comprising applying a mel frequency cepstral coefficient (MFCC) analysis to extract the unique fingerprint of the voice of the second user.”</i>	118
F.	Claim 13	118

1. *Claim 13: “The method of claim 7, further comprising receiving the second input audio data from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”*.118

XIII. OPINIONS REGARDING GROUND V: NGUYEN, DIRAC, NARAYANAN, AND WANG.....118

A. Claim 14 118

1. *Claim 14[Pre]: “A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:”* 118
2. *Claim 14[a]: “apply one or more trained machine learning models to first input audio data obtained via a microphone to extract accent features of first input speech associated with a first accent of a first user;”* 119
3. *Claim 14[b]: “analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise speaker-specific voice characteristics;”* 119
4. *Claim 14[c]: “synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and”* 119
5. *Claim 14[d]: “provide for output via an audio output device output audio data generated based on the modified version of the second input speech.”* 119

B. Claim 15 120

1. *Claim 15: “The non-transitory computer-readable medium of claim 14, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first user.”* 120

C.	Claim 16	120
1.	<i>Claim 16: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.”</i>	<i>120</i>
D.	Claim 17	120
1.	<i>Claim 17: “The non-transitory computer-readable medium of claim 14, wherein the accent features comprise one or more pitch contour, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”</i>	<i>120</i>
E.	Claim 19	120
1.	<i>Claim 19: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor to apply a speaker identity encoding technique to encode the speaker-specific voice characteristics.”</i>	<i>120</i>
F.	Claim 20	121
1.	<i>Claim 20: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”</i>	<i>121</i>
XIV.	OPINIONS REGARDING GROUND VI: NGUYEN, DIRAC, NARAYANAN, KAMRUZZAMAN, AND WANG	121
A.	Claim 12	121

1. *Claim 12: “The method of claim 7, further comprising applying a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.” 121*

B. Claim 18 122

1. *Claim 18: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.” 122*

XV. CONCLUSION.....122

CLAIM LISTING

Claim Designation	Claim Language
Claim 1(Pre)	A speech processing system, comprising an audio interface coupled to a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to:
Claim 1(a)	apply one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user;
Claim 1(b)	analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns;
Claim 1(c)	synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent; and
Claim 1(d)	provide to the audio interface output audio data for output via the audio output device, wherein the output audio data is generated based on the modified version of the second input speech.
Claim 2	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.
Claim 3	The speech processing system of claim 1, wherein the accent features comprise one or more pitch contours, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech,

	the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.
Claim 4	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.
Claim 5	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.
Claim 6	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.
Claim 7(Pre)	A method for real-time accent mimicking, the method implemented by a speech processing system and comprising:
Claim 7(a)	applying one or more trained machine learning models to first input audio data to extract accent features of first input speech associated with a first accent of a first user;
Claim 7(b)	analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise a unique fingerprint of the voice of the second user;
Claim 7(c)	synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and
Claim 7(d)	providing output audio data generated based on the modified version of the second input speech.

Claim 8	The method of claim 7, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.
Claim 9	The method of claim 7, further comprising extracting from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.
Claim 10	The method of claim 7, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.
Claim 11	The method of claim 7, further comprising applying a mel frequency cepstral coefficient (MFCC) analysis to extract the unique fingerprint of the voice of the second user.
Claim 12	The method of claim 7, further comprising applying a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.
Claim 13	The method of claim 7, further comprising receiving the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.
Claim 14(Pre)	A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:
Claim 14(a)	apply one or more trained machine learning models to first input audio data obtained via a microphone to extract accent features of first input speech associated with a first accent of a first user;
Claim 14(b)	analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise speaker-specific voice characteristics;
Claim 14(c)	synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and

Claim 14(d)	provide for output via an audio output device output audio data generated based on the modified version of the second input speech.
Claim 15	The non-transitory computer-readable medium of claim 14, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.
Claim 16	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.
Claim 17	The non-transitory computer-readable medium of claim 14, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.
Claim 18	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.
Claim 19	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a speaker identity encoding technique to encode the speaker-specific voice characteristics.
Claim 20	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.

I, Cristopher Schmandt, declare as follows:

I. INTRODUCTION

1. I have been retained by counsel for Petitioners as a technical expert in the above-captioned case. Specifically, I have been asked to render certain opinions regarding the PGR petition with respect to U.S. Patent No. 10,417,756 (the “’756 Patent”). I understand that the Challenged Claims are 1-20, and my opinions herein are limited to those claims. A true and correct copy of my Curriculum Vitae, which provides further details about my background and experience, is appended to 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 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. 12,417,756 (“’756 Patent”) (Ex. 1001) including the specification and claims, and the ’756 Patent’s prosecution history (“’756 File History”) (Ex. 1002). In developing my

opinions relating to the '756 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. 12,417,756 (“’756 Patent”)
Exhibit 1002	Prosecution History of U.S. Patent No. 12,417,756 (“’756 Prosecution History”)
Exhibit 1003	Declaration of Christopher Schmandt
Exhibit 1004	U.S. Patent Application Publication No. 2024/0098218 to Nguyen et al. (“ <i>Nguyen</i> ”)
Exhibit 1005	U.S. Patent No. 10,163,451 to Dirac et al. (“ <i>Dirac</i> ”)
Exhibit 1006	U.S. Patent Application Publication No. 2022/0417659 to Narayanan et al. (“ <i>Narayanan</i> ”)
Exhibit 1007	U.S. Patent Application Publication No. 2024/0274122 to Wang et al. (“ <i>Wang</i> ”)
Exhibit 1008	B.H. Juang and Lawrence Rabiner, <i>Automatic Speech Recognition – A Brief History of the Technology Development</i> , 2004 (“ <i>Juang</i> ”)
Exhibit 1009	Li Deng, <i>Deep learning: from speech recognition to language and multimodal processing</i> , APSIPA Transactions on Signal and Information Processing, Vol. 5, E. 1, 2016 (“ <i>Deng</i> ”)
Exhibit 1010	Vlado Delic, Zoran Peric, Milan Secujski, Niksa Jakovljevic, Jelena Nikolic, Dragisa Miskovic, Nikola Simic, Sinisa Suzic, and Tijana Delic, <i>Speech Technology Progress Based on New Machine Learning Paradigm</i> , Computational Intelligence and Neuroscience, 2019 (“ <i>Delic</i> ”)
Exhibit 1011	S.M. Kamruzzaman, A.N.M. Rezaul Karim, Saiful Islam, and Emdadul Haque, <i>Speaker Identification using MFCC-Domain Support Vector Machine</i> , International Journal of Electrical and Power Engineering, 2007 (“ <i>Kamruzzaman</i> ”)
Exhibit 1012	U.S. Patent Application Publication No. 2005/0066209 to Kee et al. (“ <i>Kee</i> ”)
Exhibit 1013	C. O’Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallaghan.com/research/papers/Pitch.pdf) September 30, 2015. (“ <i>O’Callaghan</i> ”)
Exhibit 1014	Berrak Sisman, Mingyang Zhang and Haizhou Li, <i>Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum</i>

	<i>and Prosody Conversion</i> , IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27, (2019) (“ <i>Sisman</i> ”)
Exhibit 1015	Ivan Chow and Steven Brown, <i>A Musical Approach to Speech Melody</i> , Frontiers in Psychology, Vol. 9, 2018 (“ <i>Chow</i> ”)
Exhibit 1016	Johan ‘t Hart, Rene Collier, and Antonie Cohen, <i>A Perceptual Study of Intonation</i> , Press Syndicate of the University of Cambridge, ISBN-0521366437, 1990 (“ <i>Hart</i> ”)
Exhibit 1017	Elise Piazza, Marius Iordan, and Casy Lew-Williams, <i>Mothers Consistently Alter their Unique Vocal Fingerprints when Communicating with Infants</i> , Current Biology, Vol 27, Issue 20, 2017 (“ <i>Piazza</i> ”)
Exhibit 1018	U.S. Patent No. 10,614,826 to Huffman et al. (“ <i>Huffman</i> ”)
Exhibit 1019	U.S. Patent Application Publication No. 2024/0304200 to Kashet et al. (“ <i>Kashet</i> ”)
Exhibit 1020	Constantinos Xydeas, <i>Differential Encoding Techniques Applied to Speech Signals</i> , 1978, (“ <i>Xydeas</i> ”)
Exhibit 1021	Jaison Joshy and Koj Sambyo, <i>A Comparison and Contrast of the Various Feature Extraction Techniques in Speaker Recognition</i> , International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 9, No. 11, 2016 (“ <i>Joshy</i> ”)
Exhibit 1022	Gaurav Giroti, Tejas Nakhate, Mahesh Laddha, and Prof. Manoj Sarve, <i>Person Identification through Voice using MFCC and Multi-class SVM</i> , International Research Journal of Engineering and Technology, Vo. 5, Issue 5, 2018 (“ <i>Giroti</i> ”)
Exhibit 1023	Cathy van Eck, <i>Between Air and Electricity</i> , Bloomsburry Academic, ISBN-978151327609 PDF, 2017 (“ <i>Eck</i> ”)
Exhibit 1024	Ashok Ambardar, <i>Analog and Digital Signal Processing</i> , Second Edition, Brooks/Cole Publishing Company, ISBN-9780534954093, 1999 (“ <i>Ambardar</i> ”)
Exhibit 1025	Mehrayar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, <i>Foundations of Machine Learning second edition</i> , The MIT Press, 2018 (“ <i>Mohri</i> ”)
Exhibit 1026	Su-Hyun Han, Ko Woon Kim, Sang Yun Kim and Young Chul Youn, <i>Artificial Neural Network: Understanding the Basic Concepts without Mathematics</i> , Dement Neurocogn Disord, 2018 Sep;17(3):83-89 (“ <i>Han</i> ”)
Exhibit 1027	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 1028	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 1029	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 1030	U.S. Patent No. 8,564,544 to Jobs et al. (“Jobs”)
Exhibit 1031	Qingchen Zhang, Laurence Yang, Zhikui Chen, and Peng Li, <i>A Survey on Deep Learning for Big Data</i> , Information Fusion, Vol. 42, 2018 (“Zhang”)
Exhibit 1032	Josef Michalek and Jan Vanek, <i>An Open-Source GPU-Accelerated Feature Extraction Tool</i> , 12 th International Conference on Signal Processing, 2014 (“Michalek”)
Exhibit 1033	Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schluter, Shuo-yiin Chang, and Tara Sainath, <i>Deep Learning for Audio Signal Processing</i> , IEEE Journal of Selected Topics in Signal Processing, Vol. 13, No. 2, 2019 (“Purwins”)
Exhibit 1034	U.S. Patent Application Publication No. 2021/0082402 to Sun et al. (“Sun”)
Exhibit 1035	John Holmes and Whendy Holmes, <i>Speech Synthesis and Recognition</i> , Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003 (“Holmes”)
Exhibit 1036	Declaration of June Munford

14. I have considered these materials from the viewpoint of a PHOSITA as of the priority date of the ’756 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the ’756 Patent is August 1, 2024. I note that my opinions provided in this Declaration are made from the perspective of a PHOSITA as of this priority date of the ’756 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 ’756 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 a *post grant* review (“PGR”) proceeding, patent claims may be deemed unpatentable if it is shown that they were anticipated or rendered obvious in view of the prior art or if it is shown that the claims are lacking written description support in the specification, the claims are not enabled by the specification, or if the claims are found to be indefinite. I understand that the prior art in a PGR is limited to patents or printed publications that predate the priority date of the patent at issue.

A. Person of Ordinary Skill in the Art

16. I understand that the ’756 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 PGR 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 a PGR 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 PGR 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 PGR will be construed according to their plain and ordinary meaning as understood in light of the claim language, the patent’s description, and the prosecution history viewed from the perspective of a PHOSITA. I further understand that where a patent defines claim

language, the definition in the patent controls, even if there are other definitions that might be understood by those working in the art. I have been instructed to construe “characteristics specific to a natural voice of the second user” as “speaker-specific attributes that are globally representative of the second user’s voice and that are used to maintain the second user’s voice throughout synthesis.”

IV. LEVEL OF A PERSON OF ORDINARY SKILL

27. Based on my review and analysis of the ’756 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a PHOSITA in the field of the ’756 Patent at the time of the earliest possible priority date (August 1, 2024) 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 August 1, 2024, 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. A person with less relevant practical experience but with additional education may also qualify as a PHOSITA in the field of the ’756 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 August 1, 2024.

V. OVERVIEW OF TECHNOLOGY

28. I was asked to briefly summarize the background of the technology from the standpoint of a PHOSITA prior to August 1, 2025.

A. History of Speech Processing

29. Speech is “by far the most widely used and natural means of communication between people.” EX1035-Holmes, 14. “However, this deceptively simple means of exchanging information is, in fact, extremely complicated. EX1035-Holmes, 14. Speech contains both content and vocal characteristics; respectively, what is said and how it is said. EX1004-Nguyen, [0013] (both content and characteristics are a part of the input speech). Vocal characteristics include traits such as pitch, intonation, and timbre. EX1006-Narayanan, [0043]. Characteristics vary based on the accent, the individual, and the expressiveness used. EX1007-Wang, [0023], [0029].

30. A machine capable of processing and producing human like speech has been a desire for engineers and scientists for decades. As early as “the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis.” EX1008-Juang, 1. Major advances in the field of statistical modeling of speech in the 1980s led to the widespread use of automatic speech recognition systems, which allowed machines to understand speech content and respond. EX1008-Juang, 1.

31. Speech processing once again evolved with the additional computing power that came along with access to modern machine learning models. EX1009-Deng, 1. With massive amounts of data available to companies for training, machine learning speech processing models became more accurate with fewer errors than any previously used techniques. EX1009-Deng, 4.

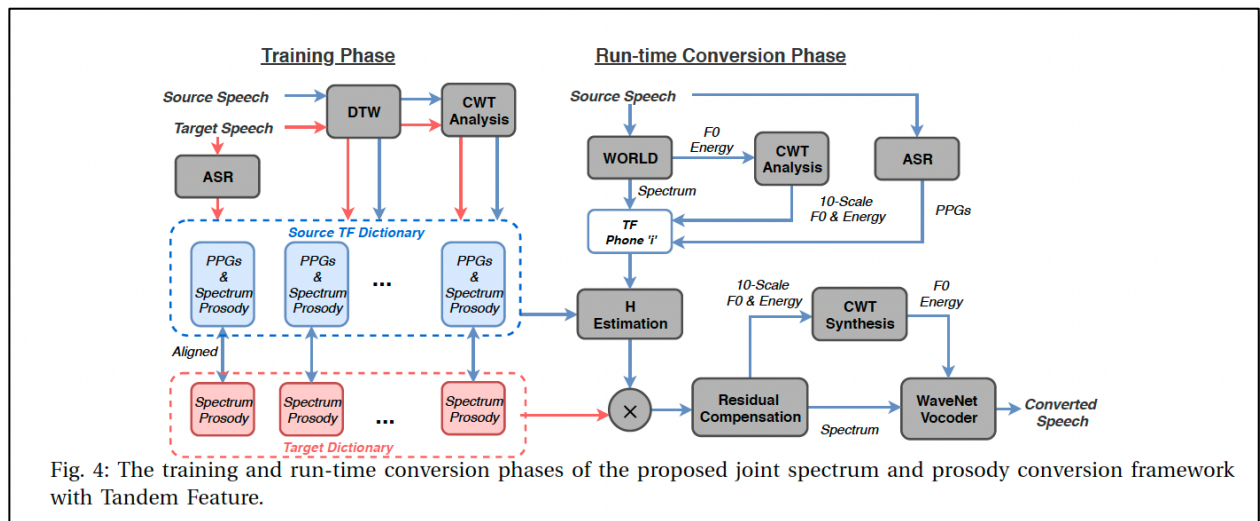
B. Vocal Features

32. As discussed above, vocal characteristics are traits, such as pitch, intonation, and timbre. EX1006-Narayanan, [0043]. These characteristics are often used as features that are extracted and input into models for further processing. EX1006-Narayanan, [0018], [0075] (various vocal characteristics are extracted and used to generate audio sounding like the speaker); EX1011-Kamruzzaman, 1 (MFCCs represent vocal characteristics and are used as inputs to a neural network).

a. Pitch or Fundamental Frequency (F0)

33. *Pitch* to O’Callaghan (“*O’Callaghan*”) teaches pitch “of a sound is identical with its fundamental frequency; that is, the pitch of a periodic sound is the greatest whole-number frequency by which the frequency of each of its sinusoidal components is divisible without remainder.” EX1013-O’Callaghan, 4. In other words, “the fundamental frequency of a complex periodic tone determines its perceived pitch.” EX1013-O’Callaghan, 4. As another example, EX1014, *Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and*

Prosody Conversion to Sisman et al. (“Sisman”) teaches “[a]t run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour.” EX1014-Sisman, 8.



EX1014-Sisman, Fig. 4. Pitch or fundamental frequency is “a crucial prosodic feature in speech.” EX1014-Sisman, 4. *Speech Synthesis and Recognition* to Holmes et al. (“Holmes”) teaches that pitch may change by “the pattern of stressed syllables...that are influenced by the phonetic detail of an utterance.” EX1035-Holmes, 123-124.

b. Intonation and Melody

34. “‘Speech melody’ and ‘intonation’ refer strictly to the pitch changes associated with speech communication.” EX1015-Chow, 2. Specifically, regarding melody, *Chow* further teaches that “‘speech melody’ refers to the pitch trajectory associated with utterances of any length.” EX1015-Chow, 2. *Chow* further states that

“‘speech melody’ refers to all pitch variations associated with speech communication.” EX1015-Chow, 2. Whereas, intonation is “the ensemble of pitch variations in speech.” EX1016-Hart, 11. Both intonation and melody are vocal characteristics used in speech processing. EX1005-Dirac, 8:59-61 (using melody as a feature for accent detection); EX1006-Narayanan, [0075] (using intonation as a characteristic in a voice profile).

c. Timbre

35. Timbre “is a spectral fingerprint that helps us instantly identify and classify sound sources, such as individual people.” EX1017-Piazza, 1. Specifically, timbre is “the tone color or unique quality of a sound.” EX1017-Piazza, 1. “Timbre enables us to discriminate, recognize, and enjoy a rich variety of sounds.” EX1017-Piazza, 4. MFCCs “provide a strong model for perceptual timbre space as a whole.” EX1017-Piazza, 4. Timbre can be used to ensure that generated audio sounds like a specific person. EX1004-Nguyen, [0073], EX1018-Huffman, 2:26-38.

C. Encoding

36. Encoding is converting data into a different format, potentially a representation of the data. EX1019-Kashet, [0095] (encoders learn a representation for a set of data); EX1020-Xydeas, 16 (the output of a microphone “is encoded into a digital form prior to transmission”). Encoding techniques are techniques used to convert data from one format to another. EX1020-Xydeas, 21 (direct waveform

encoding techniques convert waveform into digital format). Encoding is often done to convert data to a format usable by computers or machine learning models.

D. Feature Extraction

37. “Feature Extraction refers to the process of extracting useful features from speech which can be used for comparing and discriminating one sample from another.” EX1021-Joshy, 2. Various techniques are used for feature extraction, such as MFCCs. EX1021-Joshy, 2.

a. MFCC

38. “MFCC is the most commonly used feature extraction technique in Speaker Recognition and Verification.” EX1021-Joshy, 2. Generally, the steps for completing an MFCC analysis are: 1) Framing, 2) Windowing or Hamming, 3) Applying the Fourier Transform, 4) Applying mel-filter banks, and 5) Applying the Discrete Cosine Transformation. EX1021-Joshy, 2-4; EX1022-Giroti, 2. “Framing is the process of segmenting the digitally converted speech signal received as input into small frames.” EX1021-Joshy, 3. Hamming decreases the amplitude of the audio as it reaches the frame’s border “to avoid any discontinuities.” EX1022-Giroti, 2. Next, the Fast Fourier transform is applied. EX1021-Joshy, 3. This converts the data to the frequency domain from the time domain. EX1021-Joshy, 3. Next is the mel filter banks or band-pass filters. EX1021-Joshy, 3; EX1022-Giroti, 2. The filter banks are triangular and are applied to the output of the Fourier transform. EX1021-

Joshy, 3. The filter banks are placed logarithmically, or equally placed on the mel frequency scale, to be tailored to the spectrums humans hear with greater accuracy. EX1021-Joshy, 3; EX1022-Giroti, 2. The output of applying the filter banks is a set of numbers based on the number of filter banks used. EX1022-Giroti, 2. Finally, the cepstrum is determined by taking the log of the power spectrum and applying the Discrete Cosine transform. EX1021-Joshy, 3-4. The output of this analysis is a set of coefficients that represent the audio qualities of the audio data. EX1021-Joshy, 4; EX1022-Giroti, 2, EX1011-Kamruzzaman, 2. “The set of coefficients is called an acoustic vector.” EX1011-Kamruzzaman, 2.

39. MFCCs are used in many different ways in speech processing. One way they can be used is to represent the timbre of a voice. EX1017-Piazza, 4. MFCCs can also be used as a unique fingerprint of a voice. EX1017-Piazza, 1 (because they represent the timbre of a voice). Further, MFCCs can be used to represent the vocal characteristics of a speaker or utilized as inputs to neural networks for further processing. EX1011-Kamruzzaman, 1-2.

E. Audio Interface

40. It was well known that a computer cannot process data in analog form. EX1024-Ambardar, 16; EX1020-Xydeas, 18. Because computers only accept digital data, if analog audio data is to be received and processed by a computer, it must first be converted to digital form. EX1024-Ambardar, 16; EX1020-Xydeas, 18. This

conversion is performed by an analog-to-digital converter. EX1024-Ambardar, 16. The analog-to-digital converter digitizes sampled audio received from a microphone. EX1024-Ambardar, 16. Sampling takes the continuous data stream in the form of an electrical audio signal and samples it at intervals, making the data discrete. EX1024-Ambardar, 15-16.

41. Microphones receive audio waves and convert the audio wave pressure variances into an electrical signal. EX1023-Eck, 72. The electrical signal can then be converted to digital data using the sampling process discussed above.

42. In order for a speaker or other audio output device to play audio processed or generated by a computer, it must first be converted from digital form to analog form. EX1024-Ambardar, 15-16. To accomplish this, the sampling process is reversed, and the electrical current is sent to a speaker or other output device. A speaker is the opposite of the microphone as it converts an electrical signal into pressure waves by vibrating a diaphragm. EX1023-Eck, 72.

43. An audio interface contains the software and/or hardware that allows the computing device to receive audio for processing and output processed or generated audio. The components that make up an audio interface are an analog-to-digital converter, a digital-to-analog converter, and any other software or hardware required by the computing device to use digital audio data. EX1012-Kee, [0027]. The use of audio interfaces has been well known in the art for a long time because

the components that make up the audio interface, as taught by the '756 Patent, are required for a computing device to receive, process, and output sound. For example, the computing device with an audio interface taught by *Kee*, which contains the components mentioned above. EX1012-Kee, [0027].

F. Machine Learning Model

44. 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. EX1025-Mohri, 14. Input and output data is provided, and the machine determines the process by which the given input produces the given output data. EX1026-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.” EX1025-Mohri, 15.

45. 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.*

46. An artificial neural network is a machine learning algorithm inspired by the connections of human neurons. EX1026-Han, 1. A neural network seeks to recognize the underlying relationship with a set of data with a series of algorithms to simulate the working of a human brain. EX1028-Siddique, 1. Typical applications of neural networks include voice recognition and speech synthesis. EX1029-Mijwel, 3. “Neural Networks are capable of learning and can be trained to produce desired results.” EX1028-Siddique, 2. Specifically, each connection between nodes in the neural network has a weight that controls the signal between those nodes. *Id.* In training, the weights are altered so that the model provides more accurate results. EX1028-Siddique, 2; EX1026-Han, 3-4.

VI. OVERVIEW OF THE '756 PATENT

47. I have reviewed the '756 Patent (Ex. 1001). Descriptions of the '756 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the '756 Patent are intended only as a summary of the '756 Patent and/or intended as being pertinent to the Challenged Claims of the '756 Patent. My opinions provided herein are not intended as a discussion of my entire understanding of the '756 Patent or any other claims in any related patent to the '756 Patent.

A. Description of the '756 Patent

48. The '756 Patent “relates to audio analysis and, more particularly, to methods and systems for real-time accent mimicking.” EX1001, 1:10-12. The claims

of the '756 Patent are directed to a speech processing system, method, and medium for instructions that processes speech. EX1001, 9:43-12:37. The system first extracts from a first input audio, from a first user, accent features using machine learning model(s). EX1001, 9:48-51. Next, the system analyzes a second input audio, from a second user, to generate characteristics of the second user's voice. EX1001, 9:52-60. The system then synthesizes a modified version of the second input audio "based on" the previously extracted accent features and generated characteristics. EX1001, 9:60-64. Finally, the system provides output audio data that is "based on" the modified version of the second input audio. EX1001, 10:1-4.

B. Field of Endeavor of the '756 Patent

49. The '756 Patent describes systems and methods "for real-time accent mimicking." EX1001, Abstract, Title. The '756 Patent discloses systems such as "a speech processing system 100 that aids speakers with accents in adopting the listeners' accents, thereby enhancing communication." EX1001, 2:14-16. After reviewing the '756 Patent, it is my opinion that a PHOSITA would have understood that the field of endeavor of the '756 Patent includes speech audio processing methods, devices, and systems.

C. Problem Purportedly Solved by the Inventors of the '756 Patent

50. I have been asked to analyze the '756 Patent and determine the problem that the inventors were trying to solve. The '756 Patent purportedly solves issues

with real-time accent mimicking. EX1001, 2:10-12, *see also* Title, Abstract, 1:16-26, 2:6-9.

51. Another problem the '756 Patent purportedly solves is preserving the voice of the speaker while mimicking the accent of the user. EX1001, Abstract, 4:63-67, 5:10-15, 6:36-40, Fig. 4. A further problem the '756 Patent addresses is generating a speaker's unique vocal traits from received audio so that audio can be generated with the voice of the speaker but with the accent of the user. EX1001, Abstract, 3:43-64, 3:65-4:7, 4:59-67, Fig. 4. The '756 Patent notes that existing solutions struggle "to adapt to the nuances and variability of real-time conversations." EX1001, 1:50-51.

D. Prosecution History of the '756 Patent

52. I have reviewed the prosecution history for the '756 Patent. The '756 Patent claims priority to Provisional Application No. 63/678,180 filed August 1, 2024. Applicant received a non-final rejection on March 13, 2025, in which the claims were rejected under 35 U.S.C. 102 and 103 as anticipated and obvious in view of the prior art. EX1002, 81-86. In response, the Applicant amended the claims to narrow the scope of "generated characteristics." EX1002, 89-100. The Examiner found the claims patentable in view of these amendments:

Prior art of record discloses a similar method, but fails to teach the claims in combination with analyzing obtained second input

audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns and synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.

EX1002, 132.

VII. OVERVIEW OF THE PRIOR ART

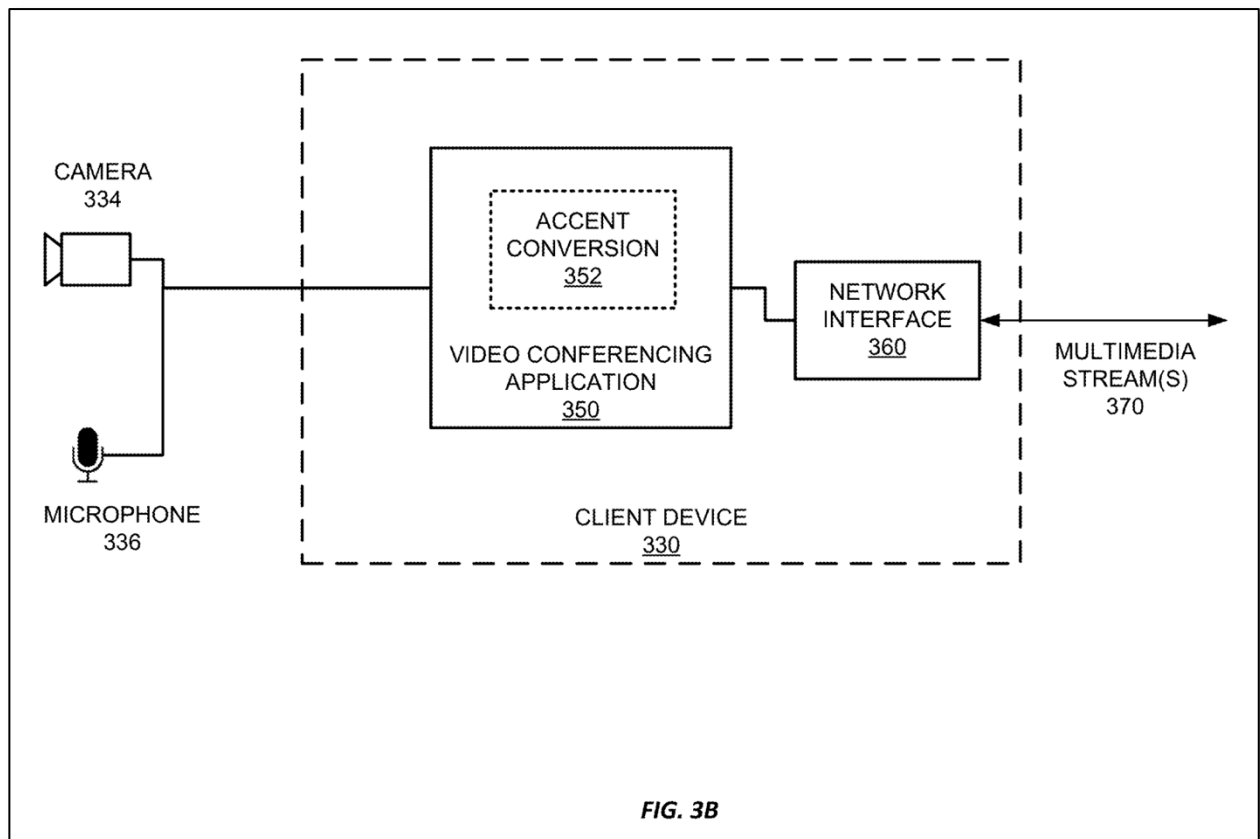
53. The prior art is analogous to the claimed invention of the '756 Patent, as discussed with greater detail for each prior art reference included. In general, the prior art is from the same field of endeavor as the '756 Patent, namely speech audio processing. EX1001, Abstract, 1:10-12, 2:14-22. Moreover, a PHOSITA would have found the prior art to be of the same field of endeavor and thus analogous to the '756 Patent.

54. The prior art is reasonably pertinent to at least one problem facing the '756 Patent inventors, namely at least one of: retaining the voice characteristics of a speaker while processing speech audio, performing real-time accent conversion,

extracting vocal features that are unique to a speaker, and inputting and outputting audio data, as described further below with respect to each denoted prior art reference.

A. Overview of *Nguyen*

55. U.S. Patent Application Publication No. 2024/0098218 (“*Nguyen*”) was filed on January 30, 2023, and published on March 21, 2024. *Nguyen* teaches a system for accent conversion for virtual conferences. EX1004-*Nguyen*, Title, Abstract.



EX1004-Nguyen, Fig. 3B. In one embodiment, the system performs accent detection on the participants in a virtual conference. EX1004-Nguyen, [0095]. When audio streams with different accents are detected, a user is prompted and may select to perform accent conversion. EX1004-Nguyen, [0095]. When selected, the client device performs accent conversion using the AC model corresponding to the detected accent. EX1004-Nguyen, [0114]-[0115]. The AC model receives “speech with a source accent and convert[s] it to speech having a target accent, while retaining the identity characteristics of the original speaker.” EX1004-Nguyen, [0072].

56. A PHOSITA would have identified *Nguyen* as being within the same field of endeavor as the '756 Patent because both *Nguyen* and the '756 Patent relate to speech audio processing. *Compare* EX1004-Nguyen, Abstract, [0075], [0098], [0117] with EX1001, Abstract, 1:10-12, 2:14-22. *Nguyen* discloses a system that performs accent conversion using one or more accent conversion (“AC”) models, where the AC models “receive audio streams from the virtual conference and each generates an output audio stream having the corresponding target accent.” EX1004-Nguyen, [0070].

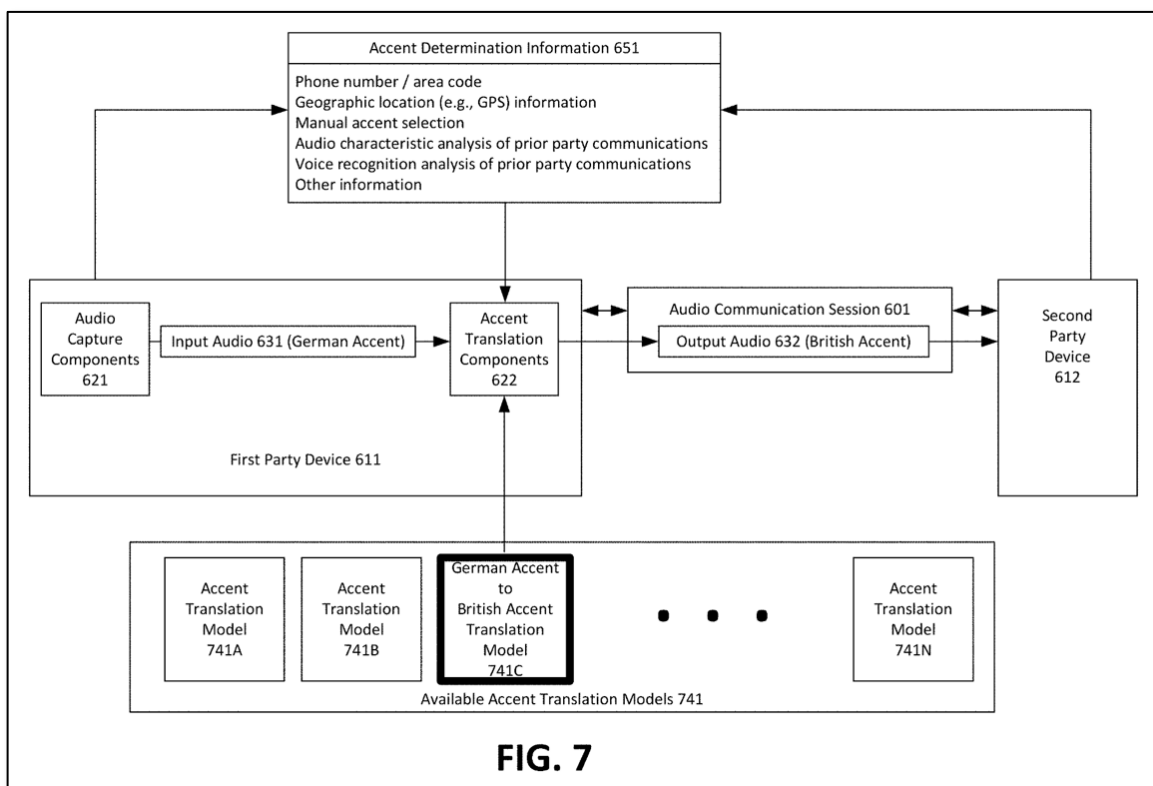
57. A PHOSITA would have found *Nguyen* reasonably pertinent to the problem faced by the inventors of the '756 Patent because both *Nguyen* and the '756 Patent describe improvements in performing real-time accent conversion. *Compare*

EX1001, Title, 1:10-12, 1:35-44, 1:49-55, *with* E. X1004-Nguyen, [0012]-[0013], [0099].

B. Overview of *Dirac*

58. U.S. Patent No. 10,163,451 (“*Dirac*”) was filed on December 21, 2016, and issued on December 25, 2018. *Dirac* teaches “techniques of accent translation.”

EX1005-Dirac, Abstract.



EX1005-Dirac, Fig. 7. In this example, the system receives input audio 631 spoken in a German accent and selects the German accent to British accent translation model from a collection of accent translation models. EX1005-Dirac, 10:13-27. The model

is used to translate the accent of the input audio 631 to generate output audio 632 with a second accent. *Id.*

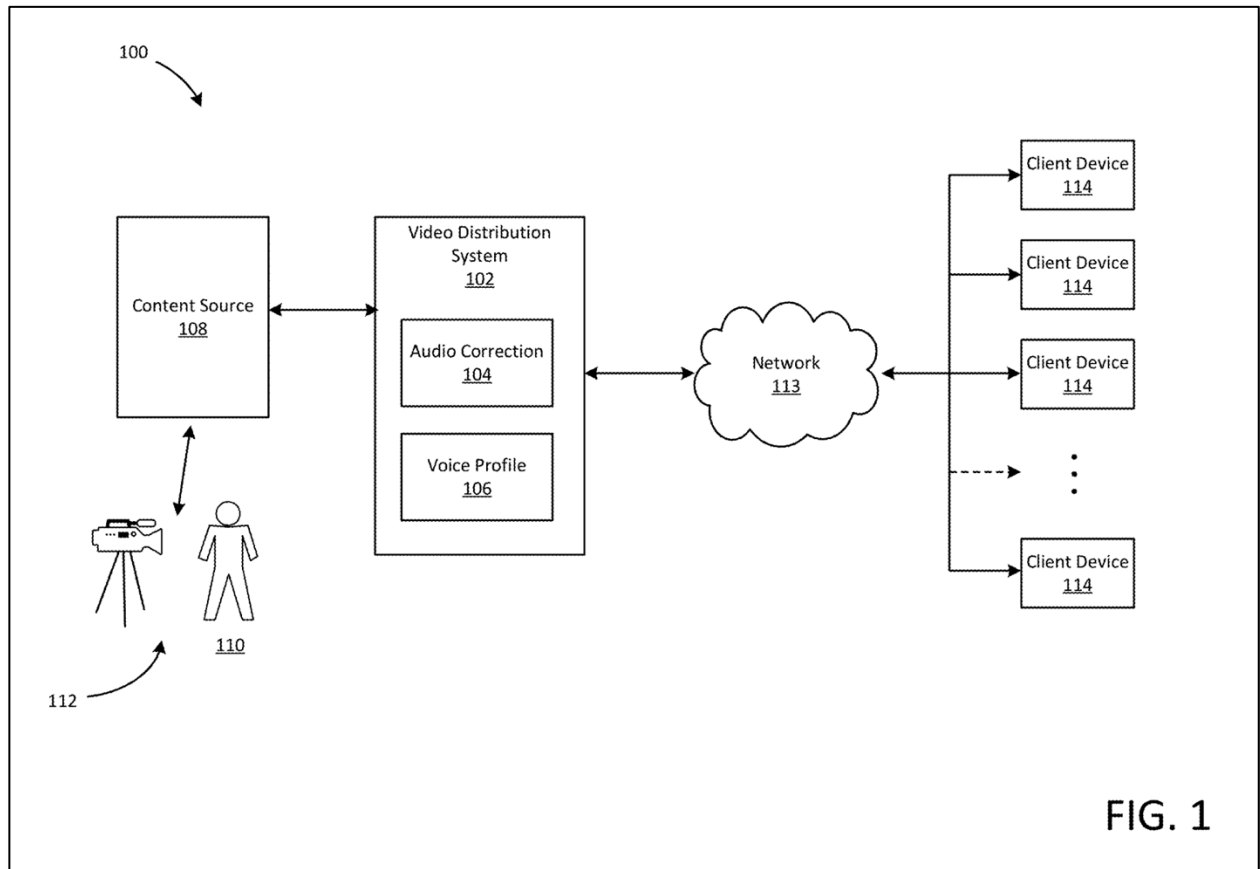
59. A PHOSITA would have identified *Dirac* as being within the same field of endeavor as the '756 Patent because both *Dirac* and the '756 Patent relate to speech audio processing. *Compare* EX1001, Abstract, 1:10-12, 2:14-22 with EX1005-Dirac, Abstract, 1:5-20, 2:19-22, 3:12-17. *Dirac* provides “techniques for accent translation.” EX1005-Dirac, Abstract, 1:66.

60. A PHOSITA would have found *Dirac* reasonably pertinent to the problem faced by the inventors of the '756 Patent because both *Dirac* and the '756 Patent describe the difficulties posed by real-time accent conversion. *Compare* EX1001, Title, 1:10-12, 1:35-44, 1:49-55, with EX1005-Dirac, 3:7-9, 7:61-8:15. For example, *Dirac* teaches “in many cases, even when individuals speak the same language, they may have difficulty understanding one another, for example due to various different accents that may be employed by different speakers.” EX1005-Dirac, 1:9-12. *Dirac* identifies a system addressing such problems via accent conversion, particularly by providing “techniques for accent translation.” EX1005-Dirac, 1:66.

C. Overview of *Narayanan*

61. U.S. Patent Application Publication No. 2022/0417659 (“*Narayanan*”) was filed on June 23, 2021, and published on December 29, 2022. *Narayanan*

teaches “systems, methods, and devices relating to audio correction.” EX1006-Narayanan, Abstract.



EX1006-Narayanan, Fig. 1. *Narayanan* teaches “[s]ystems, methods, and devices relating to audio correction.” EX1006-Narayanan, Abstract. *Narayanan* discloses using a voice profile as part of the audio correction:

For example, the voice profile module 106 may determine the voice profile based on the spoken audio content in the content portion that indicates the incorrect word (as opposed to background audio content in the content portion). The voice profile may be determined based on audio, vocal, and/or linguistic characteristics of the spoken audio content, such as

audio spectral patterns, vocal frequency, vocal pitch, speaking speed, intonation, loudness, amplitude, speech patterns, speech cadence, the number and/or duration of utterances, the number and/or duration of breaks between utterances.

EX1006-Narayanan, [0023].

62. A PHOSITA would have identified *Narayanan* as being within the same field of endeavor as the '756 Patent because both *Narayanan* and the '756 Patent relate to speech audio processing. *Compare* EX1001, Abstract, 1:10-12, 2:14-22 *with* EX1006-Narayanan, Title, Abstract, [0004], [0017]. *Narayanan* describes audio correction” of words spoken in content. EX1006-Narayanan, [0004].

63. A PHOSITA would have found *Narayanan* reasonably pertinent to the problem faced by the inventors of the '756 Patent because both *Narayanan* and the '756 Patent describe the difficulties of retaining the voice characteristics of a speaker while processing speech audio. *Compare* EX1001, Abstract, 1:45-55, 5:11-15, 6:32-40, 7:44-48, *with* EX1006-Narayanan, Abstract, [0018], [0023]-[0024], [0035], [0075].

D. Overview of *Kamruzzaman*

64. “*Speaker Identification using MFCC-Domain Support Vector Machine*” by Kamruzzaman et al. (“*Kamruzzaman*”) was publicly available at least as early as September 20, 2007. EX1036-Munford, ¶¶62-66 (declaration of June Munford establishing public availability of *Kamruzzaman*). *Kamruzzaman* teaches

the use of mel-frequency cepstrum coefficients (MFCCs) for speaker identification. EX1011-Kamruzzaman, Abstract. *Kamruzzaman* discloses the use of mel-frequency cepstrum coefficients (MFCCs) for speaker identification. EX1011-Kamruzzaman, Abstract. The cepstrum coefficients represent speaker characteristics. EX1011-Kamruzzaman, Abstract. *Kamruzzaman* discloses that the “set of [cepstrum] coefficients is called an acoustic vector.” EX1011-Kamruzzaman, 2. “These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.” EX1011-Kamruzzaman, 2.

65. A PHOSITA would have identified *Kamruzzaman* as being within the same field of endeavor as the ’756 Patent because both *Kamruzzaman* and the ’756 Patent related to speech audio processing. *Compare* EX1001, Abstract, 1:10-12, 2:14-22 *with* EX1011-Kamruzzaman, Title, Abstract, 2-3.

66. A PHOSITA would have found *Kamruzzaman* reasonably pertinent to the problem faced by the inventors of the ’756 Patent because both *Kamruzzaman* and the ’756 Patent describe extracting vocal features that are unique to a speaker. *Compare* EX1001, 4:19-29, 5:1-5, 7:58-67, 8:34-47 *with* EX1011-Kamruzzaman, Abstract, 1-2, 5.

E. Overview of *Kee*

67. U.S. Patent Application Publication No. 2005/0066209 (“*Kee*”) was filed on June 17, 2004, and published on March 24, 2005. *Kee* teaches “a computer

system.” EX1012-Kee, Abstract. *Kee* further teaches that the computer system has an audio interface with audio processing capabilities. EX1012-Kee, [0027]. The system contains an audio interface as well as a digital-to-analog converter and an analog-to-digital converter. EX1012-Kee, [0002].

68. A PHOSITA would have identified *Kee* as being within the same field of endeavor as the ’756 Patent because both *Kee* and the ’756 Patent relate to speech audio processing. *Compare* EX1001, Abstract, 1:10-12, 2:14-22 *with* EX1012-Kee, [0027].

69. A PHOSITA would have found *Kee* reasonably pertinent to the problem faced by the inventors of the ’756 Patent because both *Kee* and the ’756 Patent describe a system inputting and outputting audio data. *Compare* EX1001, 3:4-7 *with* EX1012-Kee, [0027].

F. Overview of *Wang*

70. U.S. Patent Application No. 2024/0274122 (“*Wang*”) was filed on March 30, 2023, and published August 15, 2024. *Wang* teaches a speech translation system. EX1007-Wang, Abstract. *Wang* teaches a speech translation system. EX1007-Wang, Abstract. *Wang* discloses using a speaker encoder as part of the speech translation system:

The expressive speech translation system 100 may include a *speaker encoder 130* for generating *speaker embedding data*

135. The speaker embedding data 135 represents *voice identity characteristics* that differ from speaker to speaker such as *fundamental frequency (e.g., pitch), timbre, cadence*, etc. In some implementations, the expressive speech translation system 100 may use the speaker encoder 130 to generate speaker embedding data 135b using the input audio data; thus, generating a speaker embedding that represents a speaker of the source speech.

EX1007-Wang, [0030].

71. A PHOSITA would have identified *Wang* as being within the same field of endeavor as the '756 Patent because both *Wang* and the '756 Patent related to speech audio processing. Compare EX1001, Abstract, 1:10-12, 2:14-22 with EX1007-Wang, Abstract, [0028], [0030], [0047]. *Wang* describes processing source speech. EX1007-Wang, Abstract.

72. A PHOSITA would have found *Wang* reasonably pertinent to the problem faced by the inventors of the '756 Patent because both *Wang* and the '756 Patent describe the difficulties retaining the voice characteristics of a speaker while processing speech audio. Compare EX1001, Abstract, 1:45-55, 5:11-15, 6:32-40, 7:44-48, with EX1007-Wang, Abstract, [0030], [0036], [0047], [0049].

VIII. SUMMARY OF UNPATENTABILITY

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

Proposed Grounds of Unpatentability	Exhibits
<u>Ground 1:</u> <i>Nguyen</i> in view of <i>Kee, Dirac, and Narayanan</i> renders obvious Claims 1-3, and 6 of the '756 Patent.	1004, 1005, 1006, 1012
<u>Ground 2:</u> <i>Nguyen</i> in view of <i>Kee, Dirac, Narayanan, and Kamruzzaman</i> renders obvious Claims 4 of the '756 Patent.	1004, 1005, 1006, 1011, 1012
<u>Ground 3 :</u> <i>Nguyen</i> in view of <i>Kee, Dirac, Narayanan, and Wang</i> renders obvious Claims 5 of the '756 Patent.	1004, 1005, 1006, 1012, 1007
<u>Ground 4:</u> <i>Nguyen</i> in view of <i>Dirac, Narayanan, and Kamruzzaman</i> renders obvious Claims 7-11 and 13 of the '756 Patent.	1004, 1005, 1006, 1011
<u>Ground 5:</u> <i>Nguyen</i> in view of <i>Dirac, Narayanan, and Wang</i> renders obvious Claims 14-17 and 19-20 of the '756 Patent.	1004, 1005, 1006, 1007
<u>Ground 6:</u> <i>Nguyen</i> in view of <i>Dirac, Narayanan, Kamruzzaman, and Wang</i> renders obvious Claims 12 and 18 of the '756 Patent.	1004, 1005, 1006, 1011, 1007

IX. OPINIONS REGARDING GROUND I: NGUYEN, KEE, DIRAC, AND NARAYANAN

A. Claim 1

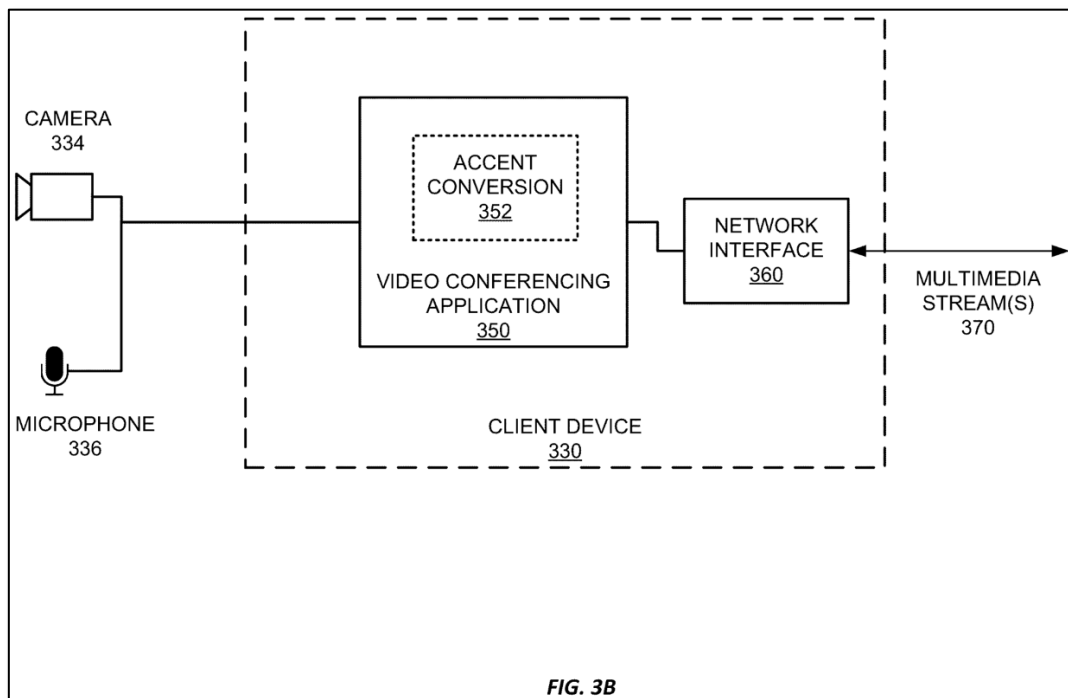
- 1. Claim 1[Pre]: “A speech processing system, comprising an audio interface coupled to a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to:”**

74. In my opinion, *Nguyen* in view of *Kee* renders obvious a speech processing system (e.g., *Nguyen*’s accent conversion system), comprising an audio interface (e.g., *Kee*’s audio interface) coupled to a microphone (e.g., *Nguyen*’s microphone) and an audio output device (e.g., *Nguyen*’s speaker or headphones), memory (e.g., *Nguyen*’s memory) having instructions stored thereon (e.g., *Nguyen*’s processor executable instructions), and one or more processors (e.g., *Nguyen*’s processor) coupled to the memory (e.g., *Nguyen*’s memory) and the audio interface (e.g., *Kee*’s audio interface) and configured to execute the instructions, as claimed.

a. Nguyen’s Teachings

75. *Nguyen* discloses a “system 300 for accent conversion for virtual conferences.” EX1004-*Nguyen*, [0063], Title, Abstract. In the system, a video conferencing application located on client devices is used to perform accent conversion. EX1004-*Nguyen*, [0072]. The client devices are “able to interact with each other to conduct the meeting, such as by viewing video feeds and hearing audio feeds from other participants, and by capturing and transmitting video and audio of

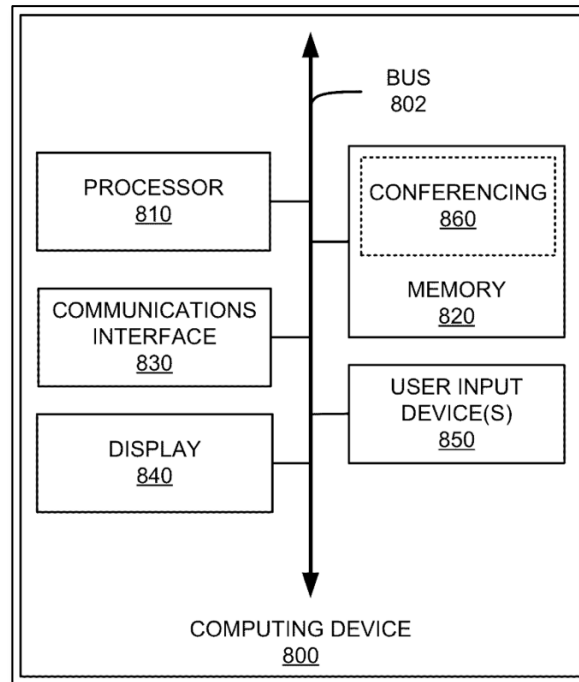
themselves.” EX1004-Nguyen, [0064]. The client devices also utilize a microphone for audio input and a speaker or headphones for audio output. EX1004-Nguyen, [0072], [0012], [0115]. Fig. 3B shows a client device on which “Accent Conversion” (AC) is performed by way of the video conferencing application:



EX1004-Nguyen, Fig. 3B. *Nguyen* discloses that the video conferencing application 350 “includes accent conversion functionality 352 that includes one or more trained AC models”. EX1004-Nguyen, [0071]-[0072]. In this system, the video conferencing application 350 receives audio data from a microphone 336. EX1004-Nguyen, [0072]. The video conferencing application also receives audio data from other participant devices participating in the conference. EX1004-Nguyen, [0072].

76. The system “may perform accent detection on audio streams within the virtual conference.” EX1004-Nguyen, [0108]. Accent detection is used to determine if “a suitable AC model is available to perform accent conversion.” EX1004-Nguyen, [0108].

77. Fig. 8 discloses an exemplary device for performing accent conversion during a virtual conference. EX1004-Nguyen, [0117]. The device contains one or more input devices, including a microphone, as well as “a processor 810 which is in communication with the memory 820 and other components of the computing device 800 using one or more communications buses.” EX1004-Nguyen, [0117]. Further, “the processor 810 is configured to execute processor-executable instructions stored in the memory 820 to perform one or more methods for accent conversion for virtual conferences according to different examples, such as part or all of the example methods 600, 700.” *Id.* Thus, *Nguyen* teaches a *memory having instructions stored thereon, and one or more processors coupled to the memory ... and configured to execute the instructions*. The processor may comprise a microprocessor, a digital signal processor, and other similar processors. EX1004-Nguyen, [0120].



EX1004-Nguyen, Fig. 8.

78. Therefore, Nguyen discloses a *speech processing system, comprising ... a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to*. A PHOSITA would have understood that *Nguyen* contains the functionality of an audio interface, otherwise audio processing could not have occurred in *Nguyen*'s system. However, Nguyen does not expressly disclose an audio interface.

b. Kee's Teachings

79. The '756 patent describes an audio interface as sound processing hardware and/or software with analog-to-digital and digital-to-analog converters. EX1001, 3:4-7.

80. *Kee* teaches *an audio interface*, as claimed. Specifically, *Kee* teaches:

An audio interface 162 that is connected to an internal **microphone** 164 and a pair of **speakers** 166a, b. The audio interface 162 includes a **digital-to-analog converter** having a pair of outputs that are coupled to the speakers 166a, b. The audio interface 162 also includes a sampler producing analog samples of a signal from the microphone 164, and an **analog-to-digital converter**, which digitizes the analog samples and passes the digital sample data to the peripheral bus 140.

EX1012-Kee, [0027].

81. The audio interface taught by *Kee* contains all of the components of the audio interface taught by the '756 Patent. For example, the '756 provides that the audio interface comprises hardware or software that includes a digital-to-analog converter and an analog-to-digital converter. EX1001, 3:4-7. In the same way, *Kee* discloses an audio interface that includes a digital-to-analog converter and an analog-to-digital converter. The *Nguyen-Kee* combination teaches *an audio interface coupled to a microphone and an audio output device and one or more processors coupled to the memory and the audio interface*. *Nguyen's* "processor 810 [] is in communication with the memory 820 and **other components of the**

computing device 800 using one or more communications buses,” where such “other components” in the *Nguyen-Kee* combination include *Nguyen*’s microphone, *Nugyen*’s speaker/headphone, and *Kee*’s audio interface. EX1004-*Nguyen*, [0072], [0012], [0115], [0117], Fig. 8; EX1012-*Kee*, [0027].

c. Motivation to Combine

82. *Nguyen* teaches receiving input audio and outputting processed audio. But, *Nguyen* does not expressly recite the use of an audio interface that has analog-to-digital and digital-to-analog capabilities. However, *Kee* teaches a device with an audio interface that has analog-to-digital and digital-to-analog capabilities.

83. A PHOSITA would have been motivated to combine *Kee*’s use of an audio interface with *Nguyen*’s accent conversion system with a reasonable expectation of success for several reasons.

84. First, *Nguyen* explicitly contemplates receiving input audio for processing, but does not provide details on how this functionality is implemented. *Nguyen* simply states that a client device is a computing or communications device with audio or video capability. EX1004-*Nguyen*, [0027]. *Nguyen* teaches that during the virtual conference the client device receives audio but, *Nguyen* does not provide the software means by which the audio is received and converted from analog to digital form. See EX1004-*Nguyen*, [0071]-[0072]. However, *Nguyen* does teach that processing is performed on the input audio stream. EX1004-*Nguyen*, [0072]. A

PHOSITA would have known that for a computing device to process audio data, the audio data first must be converted to digital form. When audio signals are converted into digital signals, “they can be processed by the same logic circuits used in digital computers.” EX1024-Ambardar, 16. After signals are converted to digital form, computers can perform complex signal processing. EX1020-Xydeas, 18. This is why audio data processing on computers is done in digital form. *Kee* provides a detailed description of an audio interface with analog-to-digital conversion capabilities. EX1012-Kee, [0027]. Thus, A PHOSITA would have been motivated to look to a reference such as *Kee* as a guide for implementing an audio interface to receive input audio and convert that audio to digital form because *Kee* expressly teaches an audio interface for capturing analog audio by a microphone, and converting analog audio into digital form for use in a computing device. EX1012-Kee, [0027].

85. Further, a PHOSITA would have understood that using an audio interface would have provided *Nguyen’s* system with not only the means to receive audio data and digitize it, but also the means to convert processed or generated audio data back into analog form so that the processed audio is able to be transmitted to the output device. A PHOSITA would have understood that processed or generated audio is converted from digital to analog form for output to the user via a speaker or headphones. Digital data must be converted to analog form before it can be heard. EX1024-Ambardar, 15-16.

86. Moreover, the use of an audio interface was well-known in the art long before the '756 Patent. For example, *Jobs*, published in 2013, teaches the use of an audio interface with a speaker and a microphone. EX1030-Jobs, 11:18-19. Similar to the proposed combination, electrical signals (analog data) are converted to digital form for processing. EX1030-Jobs, 11:24-28. Digital data is converted to electrical signals (analog data) for transmission to a speaker. EX1030-Jobs, 11:18-24.

87. Second, such a combination would have constituted combining prior art elements (e.g., *Kee's* audio interface and *Nguyen's* accent conversion system) according to known methods to yield predictable results (e.g., enabling *Nguyen's* accent conversion system to convert analog microphone signals to digital signals for subsequent processing). The result of this combination, an accent conversion system capable of analog-to-digital and digital-to-analog conversion, facilitates the functionality described by *Nguyen*. A PHOSITA would have understood that computer audio processing is performed on digital data, but that audio is received and output in analog. EX1020-Xydeas, 18; EX1024-Ambardar, 15-16, (stating that digital audio data “must be processed and converted to analog form before you can hear the music.”), 16 (“The digital processing of analog signals requires that we use an analog-to-digital converter (ADC) for sampling the analog signal prior to processing and a digital-to-analog converter (DAC) to convert the processed digital signal back to analog form.”). Thus, *Kee's* audio interface effectuates *Nguyen's*

accent conversion system by providing a means to convert audio data between analog and digital formats.

88. Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because both *Nguyen* and *Kee* teach similar electronic devices that use processors, and the implementation of audio interfaces (and particularly analog-to-digital and digital-to-analog converters) was well known to PHOSITAs. EX1004-*Nguyen*, [0071]-[0072], Fig. 8; EX1012-*Kee*, Title, [0002]. Thus, a PHOSITA would have understood that *Nguyen*'s system is capable of incorporating *Kee*'s audio interface. *Nguyen* teaches a client device that "may be any suitable computing or communications device." EX1004-*Nguyen*, [0027]. *Nguyen* further teaches that desktop or laptop computers may be such computing devices. EX1004-*Nguyen*, [0027]. And like *Nguyen*, *Kee* also teaches a computer system. EX1012-*Kee*, [0016]. In my opinion, a PHOSITA would have understood that the audio interface of *Kee* could have been implemented in the system of *Nguyen* because both references teach computing devices with microphones and speakers that intake, process, and output audio. EX1004-*Nguyen*, [0072], [0115]; EX1012-*Kee*, [0027]. A PHOSITA would have understood that only minor modifications to the hardware and software would have been required to implement the audio interface, as taught by *Kee*, into the accent conversion system, as taught by *Nguyen*. Because both *Nguyen* and *Kee* teach computer devices, a PHOSITA would have

understood that *Nguyen's* client device is able to utilize components of *Kee's* computer system with minimal changes to *Nguyen's* system. A PHOSITA would have had the requisite tools and experience necessary to make this combination without undue experimentation.

89. Therefore, in my opinion, *Nguyen* in view of *Kee* renders obvious Claim 1[Pre].

2. Claim 1[a]: “apply one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user;”

90. In my opinion *Nguyen* in view of *Kee* and *Dirac* renders obvious *apply[ing] one or more trained machine learning models (e.g., Dirac's accent translation system artificial neural networks) to first input audio data (e.g., Nguyen's audio stream of a first participant) obtained via the microphone (e.g., Nguyen's microphone) and the audio interface (e.g., Kee's audio interface) to extract accent features (e.g., Dirac's technique for obtaining audio characteristics of an accent) of first input speech associated with a first accent of a first user (e.g., Nguyen's audio received from its first participant in a target accent), as claimed.*

a. Nguyen's Teachings

91. *Nguyen* teaches performing accent conversion (“AC”) on communications between at least two users. EX1004-*Nguyen*, [0016], [0068]-

[0070], [0072]. As established above in Claim 1[Pre], the *Nguyen-Kee* combination system renders obvious the use of an audio interface. *Nguyen* teaches that “the video conferencing application 350 receives audio and video data from a microphone 336 and a camera 334, respectively, connected to the client device 330.” EX1004-*Nguyen*, [0071]. A PHOSITA would have understood that the microphone and audio interface, which convert the signals from analog to digital, would have worked together to allow the device and video conferencing application to receive audio data. In the combination, the audio interface of *Kee* receives analog audio data and converts it to digital form before transmitting it to the accent conversion functionality for further processing. *See* Claim 1[Pre].

92. *Nguyen*’s accent conversion system modifies a received audio stream of a second participant with a source accent into output speech with a target accent matching a first participant’s (a *first user*) speech. EX1004-*Nguyen*, [0107], [0112]. *Nguyen*’s system may detect when participants have different accents by performing accent detection on the speech of each participant:

the *client device 330* or the virtual conference provider 310 *may perform accent detection* or may analyze user profiles for participants in a virtual conference *to identify accents associated with the different participants*. If different accents among the participants are detected, the virtual conference provider 310 may provide an indication to one or more of the client devices to suggest accent conversion.

EX1004-Nguyen, [0095], Fig. 5D. Thus, *Nguyen* discloses receiving speech of the first participant with a target accent (*first input audio data* from *first user* with a *first accent*) and a second participant with a source accent, where the participants have different accents.

93. The *Nguyen-Kee* combination teaches the system receiving a *first input audio data obtained via the microphone and the audio interface*, where the *first input speech* is associated with a *first accent of a first user*. See Claim 1[Pre].

94. *Nguyen* does not disclose a specific process used by the system to identify the accent(s) of the user based on the input audio data received. Although *Nguyen* teaches using machine-learning (“ML”) models to perform the accent conversion, *Nguyen* does not specifically teach that an ML model is used to detect an accent, or which accent characteristics are used to identify an accent. EX1004-Nguyen, [0013]-[0014]. However, *Dirac* does.

b. *Dirac*’s Teachings

95. *Dirac* teaches an accent translation system for translating a party’s speech in a first accent to a second accent. Specifically, *Dirac* teaches that the “accent translation system may determine that the first party speaks with the first accent[.]” EX1005-Dirac, 3:12-14. “[T]he accent translation system may also determine that the second party speaks with the second accent” and “may then translate the first party’s speech from the first accent to the second accent, for

example by applying a respective accent translation model for translating speech from the first accent to the second accent.” EX1005-Dirac, 3:26-35.

96. *Dirac* teaches that the accent translation system includes neural networks for parsing and analyzing data:

[T]he accent translation system 130 includes ***one or more artificial neural networks***, such as a long short-term memory (LSTM) architecture or other systems that may ***efficiently parse and analyze large amounts of input data***. In one embodiment, the accent translation system 130 may continually and/or repeatedly receive new audio samples and use these new audio samples to update and refine the accent translation models, for example by employing machine learning or other associated techniques.

EX1005-Dirac, 7:52-60. Thus, a PHOSITA would have understood that functions of the accent translation system that require analyzing incoming data is done by one or more neural networks. *Dirac*’s machine learning models are used to accomplish specific tasks such as parsing and analyzing input data. EX1005-Dirac, 7:52-60. Because the amount of data contained in input audio is often large, parsing through and analyzing the input audio data would have been accomplished by one or more machine learning models. EX1031-Zhang, 1-2. Further, a PHOSITA would have understood that the neural networks of *Dirac* are trained. Machine learning models are trained to accomplish tasks, and an untrained model

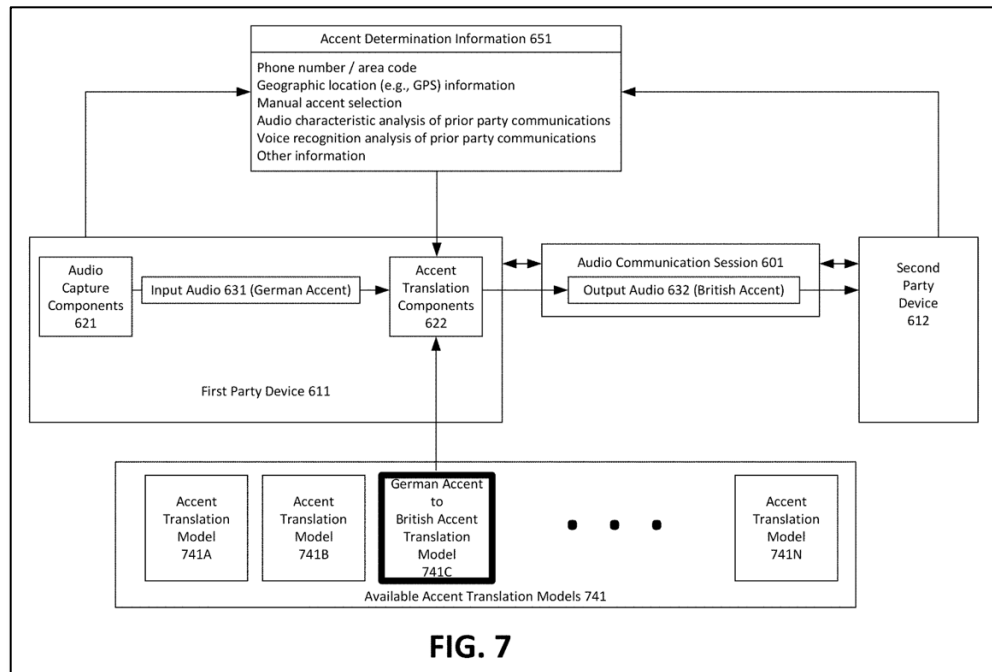
would not be as effective in accomplishing the task the model has been given to accomplish. EX1009-Deng, 4 (trained models are the most accurate and “the more training data are used, the better the accuracy”).

97. *Dirac* teaches *extract[ing] accent features* because *Dirac* obtains audio characteristics from one party’s speech to analyze and detect which accent is spoken. Specifically, *Dirac* discloses that “an audio characteristic analysis of audio communications by the first party may be used to ***determine the first party’s accent.***” EX1005-Dirac, 8:53-55. To determine the accent, “audio communications from the first party may be analyzed to obtain audio characteristics (e.g., pitch, tone, stress, melody, etc.) of the first party’s accent.” EX1005-Dirac, 8:59-61. Audio characteristics analysis techniques, such as fast Fourier transform (FFT) and mel-frequency cepstral coefficients (MFCC), can be used to obtain the audio characteristics. EX1005-Dirac, 8:62-64.

98. Next, “the audio characteristics of the first party’s accent may then be compared to audio characteristics of various accent sample sets ... to determine an accent set and an accent to which the first party’s accent substantially correlates.” EX1005-Dirac, 8:64-9:2. *Dirac* further teaches that the translation model is determined “by one or more artificial neural networks.” EX1005-Dirac, 22:9-11. Thus, a PHOSITA would have understood that the one or more neural networks of *Dirac*’s accent translation system would have been responsible for analyzing the

input audio communications to obtain the audio characteristics of the first party's accent and determining which sample accent it correlates with. Analyzing input audio data to extract features is a computationally intensive activity, and the amount of data contained in input audio is large. EX1032-Michalek, 1. Tasks requiring parsing through and analyzing large amounts of input data are often accomplished by machine learning models due to the success they have shown when dealing with large amounts of data. EX1031-Zhang, 1-2; EX1005-Dirac, 7:52-60, 8:58-67 (describing analysis of audio to determine a party's accent). Thus, Dirac teaches *apply[ing] one or more trained machine learning models to first input audio data ... to extract accent features of first input speech associated with a first accent of a first user*, where the *first input speech associated with the first accent of a first user* is the speech from *Nguyen's* exemplary first user with a target accent.

99. Fig. 7 provides an example of the accent translation system taught by *Dirac*. EX1005-Dirac, 1:48-50.



EX1005-Dirac, Fig. 7. In this example, input audio 631 spoken in a German accent is detected by the accent translation system, which selects the German accent to British accent translation model for translating the input audio 631 to output audio 632. EX1005-Dirac, 10:13-27.

c. Motivation to Combine

100. *Nguyen* teaches use of machine learning models (such as a neural network) and teaches performing accent detection. But *Nguyen* does not expressly provide that accent detection is performed by a machine learning model, nor does it expressly provide to extract accent features of the first input speech associated with a first accent of a first user. However, *Dirac* teaches accent detection performed by a machine learning model, utilizing accent feature extraction.

101. A PHOSITA would have been motivated to incorporate *Dirac*'s accent detection and feature extraction process into the *Nguyen-Kee*'s accent conversion system with a reasonable expectation of success for several reasons. In addition to the reasons provided below, both *Nguyen* and *Dirac* teach accent conversion methods. EX1004-Nguyen, Title; EX1005-Dirac, Abstract.

102. First, *Nguyen* explicitly contemplates performing accent detection on an audio stream but does not provide details on how this functionality is implemented. *Nguyen* teaches "the virtual conference provider may automatically detect accents and if it detects participants with different accents than the user of the client device 330, the virtual conference provider may select any participants who have a different accent." EX1004-Nguyen, [0090].

103. Because *Nguyen* does not provide the method by which the virtual conference provider automatically detects an accent, a PHOSITA would have been motivated to look to other references, such as *Dirac*, for detailed processes to do so. *Nguyen*'s accent conversion requires determining the accent of a participant to apply the correct accent conversion model; accordingly, a PHOSITA would have understood that an effective method of determining the participant's accent would have been required. Because the accent detection is used to select the accent conversion model, inaccurate accent detection results in inaccurate output audio from the accent conversion model.

104. *Dirac* provides a detailed description of accent detection on the speech of a participant, which includes extracting accent audio characteristics to accurately determine the accent of a speaker for accent conversion. EX1005-*Dirac*, 8:53-9:2, 9:32-38. Thus, a PHOSITA would have been motivated to look to *Dirac* as a guide for implementing accent detection in *Nguyen's* device. *Dirac* provides one such framework for accent detection that a PHOSITA would have looked to for achieving the functionality described in *Nguyen*.

105. Further, as established above, *Dirac* teaches the use of a machine learning model to analyze the input audio data. EX1005-*Dirac*, 7:52-56. A PHOSITA would have appreciated that machine learning models are advantageously used for accent detection because machine learning models are able to make accurate predictions, such as the accent of speakers, from large amounts of input data, such as an input audio stream during live conferences. Machine learning models excel when used in these contexts. EX1033-Purwins, 2. A PHOSITA would have understood that training and utilizing a machine learning model would have improved the accent detection process by creating a robust and accurate prediction model. Machine learning models are more accurate than traditional models used in speech processing. EX1009-Deng, 4; EX1033-Purwins, 2. Thus, utilizing a machine learning model to perform accent detection would improve the system.

106. Third, such a combination would have constituted applying a known technique (e.g., *Dirac*'s accent recognition) to a known device (e.g., *Nguyen*'s accent conversion system) ready for improvement to yield predictable results (e.g., improved, robust, and accurate accent recognition for subsequent accent conversion). Additionally, accent detection using a neural network was well known in the art before the '756 Patent. For example, *Sun*, which was published in 2021, teaches a system that detects an accent. EX1034-Sun, [0037], [0070], [0089].

107. Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because *Nguyen* and *Dirac* both teach the use of accent detection in an accent conversion process. *Nguyen* teaches that the "virtual conference provider may automatically detect accents." EX1004-Nguyen, [0090]. *Dirac* teaches "an audio characteristic analysis of audio communications by the first party may be used to determine the first party's accent." EX1005-Dirac, 8:53-55. Further, both teach the use of neural networks in their respective accent conversion systems. *Nguyen* teaches the use of both a voice conversion model and an accent conversion model. EX1004-Nguyen, [0014]. *Dirac* teaches the use of "one or more artificial neural networks." EX1005-Dirac, 7:53. Thus, a PHOSITA would have understood that *Nguyen*'s system is capable of incorporating *Dirac*'s neural network to implement accent characteristic extraction and accent detection. A PHOSITA would have understood that the combination would have only required minor

modifications to the software of *Nguyen*. *Nguyen*'s system already teaches the use of two separate machine learning models, and adding a third for accent detection would have been within the skill level of a PHOSITA. EX1004-Nguyen, [0014].

108. In the *Nguyen-Kee-Dirac* combination, *Nguyen*'s system receives an audio stream through the microphone and *Kee*'s audio interface during a virtual conference. The accent conversion functionality of the client device, as taught by *Nguyen*, receives audio in the form of speech from the microphone, which is then digitized by the client device's audio interface, as taught by *Kee*. The accent conversion functionality of the client device receives audio in the form of speech from a second participant via a network. The accent features of the first participant (a *first user*) and the second participant are extracted utilizing the neural network(s) of *Dirac*. The neural network(s) compare these characteristics to sample sets of accents accessible by the model to determine which accent correlates with that of the first participant and which correlates with the second participant's accent, as taught by *Dirac*. This allows the system of *Nguyen* to select the relevant AC model to perform a conversion from the accent of the second participant to that of the first participant.

109. In summary, *Nguyen-Kee-Dirac* teaches applying one or more neural networks to the speech of a first participant having a target accent to extract the accent features of the first participant's accent to determine the accent. Thus,

Nguyen-Kee-Dirac teaches appl[ing] one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user.

110. Therefore, in my opinion, *Nguyen* in view of *Kee* and *Dirac* renders obvious Claim 1[a].

3. ***Claim 1[b]: “analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns;”***

111. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *analyz[ing] obtained second input audio data associated with second input speech associated with a second accent of a second user (e.g., Nguyen’s received audio from its second participant with a source accent) to generate characteristics specific to a natural voice of the second user (e.g., Narayanan’s generating a voice profile), wherein the generated characteristics correspond to vocal traits that are distinct to the second user (e.g., Narayanan’s voice profile contains characteristics of the speaker’s specific speech) and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation*

patterns (e.g., *Narayanan's* voice characteristics, including intonation), as claimed.

I note here again that I have been instructed to construe *characteristics specific to a natural voice of the second user* as “speaker-specific attributes that are globally representative of the second user’s voice and that are used to maintain the second user’s voice throughout synthesis.”

112. The ’756 Patent teaches that the system generates characteristics of the second user’s voice. EX1001:7:44-48. However, the specification of the ’756 Patent uses “extract” and “generate” interchangeably to describe this process. *Compare* EX1001:7:44-48, 8:56-64; *with* EX1001: 5:66-6:1, 6:19-23.

a. *Nguyen's* Teachings

113. *Nguyen* teaches that after the accents of the first and second participants have been identified, the first participant at a client device is provided an indication that accent conversion for the second participant’s audio stream is available, so that the second participant’s accent is converted to that of the first participant. EX1004-*Nguyen*, [0095] (describing that “the client device 330 itself could make such a determination [(accent mismatch)] and provide the suggestion to the user” that accent conversion is available). If the first participant accepts, the client software 350 in the first participant’s client device performs accent conversion on the second participant’s speech. EX1004-*Nguyen*, [0095] (“If the user accepts the suggestion

544, ... the client software 350 may enable its accent conversion functionality 352 to perform accent conversion.”).

114. *Nguyen* generally teaches that the speech input into the accent conversion system comprises of two parts besides content, the accent and the identity characteristics. *See* EX1004-*Nguyen*, [0013], [0016], [0072]-[0073]. A PHOSITA would have understood that the input speech contains the traits that make up both the accent and the identity characteristics of the user. Otherwise, it would not be possible to extract accent and the identity characteristics from the input speech, nor would it be possible to convert the accent while retaining the identity characteristics by merely adjusting parts of the audio stream. *Nguyen* teaches that its system retains identity characteristics through accent conversion but doesn’t expressly teach that the identity characteristics are *generate[d]*, as claimed. EX1004-*Nguyen*, [0072]. *Nguyen* discusses that identity characteristics are those “such as timbre and pitch[.]” EX1004-*Nguyen*, [0073]; EX1001, 4:19-22 (the claimed *characteristics* include “prosodic features (e.g., pitch and timbre)”). A PHOSITA would have understood that identity characteristics identify a unique speaker and are therefore specific to a natural voice of a user and distinct to the user. Identification of an individual using these characteristics would not be possible if the characteristics were not specific and distinct to the speaker’s voice. For example, timbre, which can be represented from data extracted from speech audio, can be used for the recognition of individual

speakers. EX1017-Piazza, 7. Further, timbre and pitch are prosodic features. EX1001, 4:19-22.

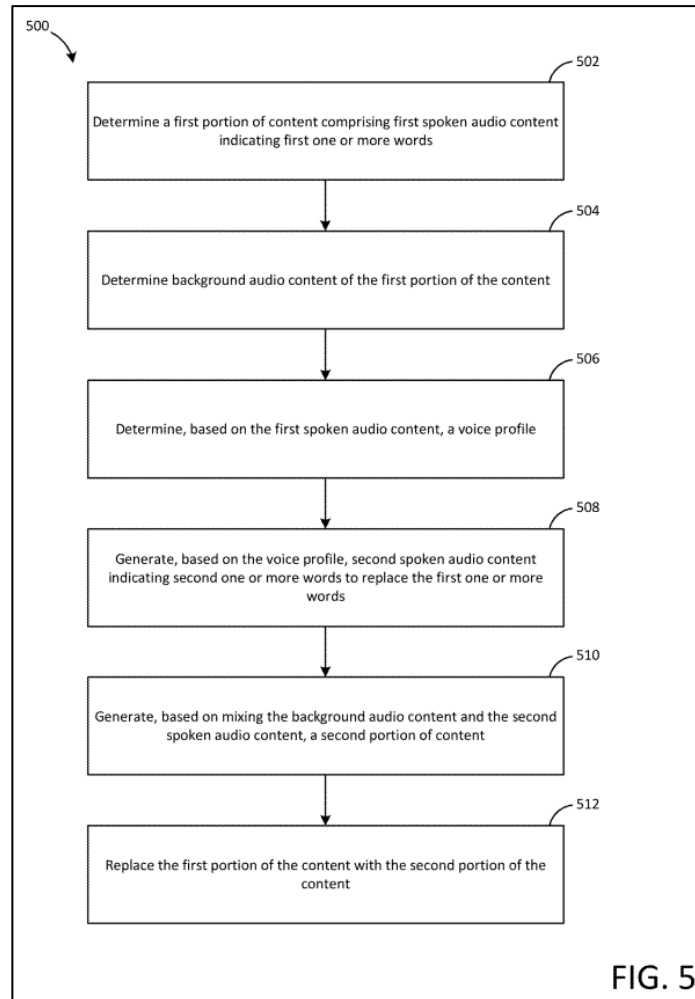
115. In *Nguyen's* accent conversion process, the incoming audio stream from the second participant has speech with a source accent that is converted by the AC model to a target accent “while *retaining the identity characteristics of the original speaker.*” EX1004-Nguyen, [0072]; *see id.* [0013], [0016], [0073], [0107], [0114]. Therefore, if the second participant with a second accent speaks during a video conference, the audio stream of their speech (*obtained second input audio data associated with second input speech*) is modified by the AC model to have the accent of the first participant, while retaining the identity characteristics of the second participant. However, *Nguyen* does not describe in detail generating the identity characteristics of the speaker for retention in the accent conversion process.

b. Narayanan's Teachings

116. *Narayanan* teaches “systems, methods, and devices relating to audio correction.” EX1006-Narayanan, Abstract. *Narayanan* discloses a method and system for real-time audio correction for mispronounced or incorrect words. EX1006-Narayanan, [0001], [0004]. This method includes the creation of a voice profile in order to retain the characteristics of the speaker through the audio correction process. *Narayanan* explains its “*voice profile may be determined a*

single time and applied” across various portions of speech. EX1006-Narayanan, [0035], Fig. 2.

117. Fig. 5 shows a method for audio correction. EX1005, [0070]. In the method, a first spoken audio content is received. EX1006-Narayanan, [0071]. Subsequently, “*a voice profile is determined* based on the first spoken audio content.” EX1006-Narayanan, [0075]. *Narayanan* discloses that “a *machine learning algorithm* (e.g., a machine learning model or other machine learning techniques) may be used to determine (e.g., update) a voice profile (e.g., the voice profile associated with the speaker in the first spoken audio content).” EX1006-Narayanan, [0079]. *Narayanan*’s system uses the voice profile to generate corrected output audio content where incorrect words are replaced with correct words, where the corrected words “resemble the speaker’s speech with respect to audio, vocal, and/or linguistic characteristics.” EX1006-Narayanan, [0083]-[0084].



EX1006-Narayanan, Fig. 5.

118. *Narayanan's* voice profile is representative of the speaker's specific speech and includes characteristics of their speech:

The voice profile may be *generally representative of the speaker's speech* such that the voice profile may be used to generate spoken audio content. *The voice profile may indicate various characteristics of the speaker's speech, such as audio, vocal, and/or linguistics characteristics.* Example audio, vocal, and/or linguistic characteristics may include audio spectral

patterns, vocal frequency, *vocal pitch*, speaking speed, *intonation*, loudness, amplitude, speech patterns, speech cadence, the number and/or duration of utterances, the number and/or duration of breaks between utterances, etc. Accordingly, the voice profile may be determined based on similar characteristics of the spoken audio content and/or the first one or more words spoken by the speaker.

EX1006-Narayanan, [0075]. The '756 Patent teaches that intonation patterns at least include the rise and fall of pitch. EX1001, 4:11-12. This is consistent with how a PHOSITA would have understood intonation. EX1016-Hart, 2. Therefore, a PHOSITA would have understood intonation to be intonation patterns.

119. A PHOSITA would have understood that the characteristics included in the voice profile, such as intonation and vocal pitch, are *specific to the natural voice of the [] user*. The '756 Patent discusses that intonation patterns and pitch (a prosodic feature) are vocal traits that are distinct to and specific to the natural voice of the user. EX1001, 8:2-7, 9:54-60.

120. Finally, a PHOSITA would have understood that the characteristics contained within the voice profile are *generated* from the spoken audio data. *Narayanan* teaches “the voice profile module 106 may determine the voice profile based on the spoken audio content.” EX1006-Narayanan, [0023]. Thus, *Narayanan* discloses *analyz[ing] obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics*

specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns. And *Narayanan's* disclosures also teach *characteristics specific to a natural voice of the second user* under Petitioner's proposed construction of "speaker-specific attributes that are globally representative of the second user's voice and that are used to maintain the second user's voice throughout synthesis" because *Narayanan* teaches that its voice profile may be generally representative of the speaker's voice and may be determined only a single time. EX1006-Narayanan, Fig. 3, [0035], [0075]. A PHOSITA would have understood that generating the voice profile a single time means that the voice profile characteristics are not frame-specific but rather globally representative of the speaker's voice that can be used throughout the speech process. *Id.* *Narayanan* does disclose that its voice profile may be updated, but, a PHOSITA would have understood that these updates may refine a persistent profile rather than replacing it with frame-level extraction. *Id.* Indeed, *Narayanan's* teachings of a single determined voice profile that is used throughout the speech modification process, and its global nature, is in line with how speaker attributes are stored in such systems. EX1007-Wang, [0036] (describing storing embedding data and reusing it for future synthesis, including for known identities).

c. Motivation to Combine

121. *Nguyen* teaches analyzing input audio data to retain identity characteristics in speech that has its accent converted. But *Nguyen* does not expressly explain the generation of voice characteristics. However, *Narayanan* does teach the generation of a voice profile containing characteristics unique to the speaker.

122. A PHOSITA would have been motivated to incorporate *Narayanan's* voice profile generation and functionality into *Nguyen-Kee-Dirac's* accent conversion system with a reasonable expectation of success, for several reasons. In addition to the reasons outlined below, both *Nguyen* and *Narayanan* teach audio processing in real time. EX1004-*Nguyen*, [0052]; EX1006-*Narayanan*, [0030].

123. First, *Nguyen* specifically states that to alleviate communication issues, its system converts one accent to another accent while retaining the identity characteristics of the original speaker. EX1004-*Nguyen*, [0013]. However, *Nguyen* does not provide a comprehensive list of its identity characteristics to preserve. *Nguyen* merely provides that “the AC model 412 will be trained to retain the speaker’s identity characteristics.” EX1004-*Nguyen*, [0073]. *Nguyen* specifies that the identity characteristics include characteristics “such as timbre and pitch.” EX1004-*Nguyen*, [0073]. Accordingly, a PHOSITA would have been encouraged and motivated to look to other references, such as *Narayanan* and its use of voice profiles for details regarding additional identity characteristics for preservation of

the speaker's identity. *Narayanan* provides various audio, vocal, and linguistic characteristics that may be included in the voice profile. EX1006-Narayanan, [0075]. These characteristics include audio spectral patterns, vocal frequency, vocal pitch, speaking speed, intonation, loudness, amplitude, speech patterns, speech cadence, the number and/or duration of utterances, and the number and/or duration of breaks between utterances. EX1006-Narayanan, [0075]. Further, the use of voice profiles in speech processing was well known in the art before the '756 Patent. For example, *Huffman*, which was published in 2020, teaches a system that "creates a voice profile." EX1018-Huffman, 9:20.

124. A PHOSITA would have also been motivated to make the proposed combination because such a combination would have improved *Nguyen's* system by allowing it to retain vocal characteristics, beyond pitch and timbre, for each participant, by way of the voice profile, during accent conversion. EX1006-Narayanan, [0018]. Use of voice profiles as taught by *Narayanan* would have allowed *Nguyen's* system to retain characteristics such as intonation, speech patterns, cadence, and audio spectral patterns. EX1006-Narayanan, [0043]. This would have allowed the output of the accent conversion system to retain more identity characteristics of the second participant, improving the output of the accent conversion system by generating output audio that is closer to the voice of the second participant. A PHOSITA would have understood that if more identity characteristics

are preserved, the output will be more personalized and natural sounding. A PHOSITA would have appreciated that such a combination would have furthered *Nguyen*'s stated goal. *Nguyen* states that "[b]ecause the accent conversion retains the identity characteristics of the original speaker, but changes the accent, the requesting participant will simply perceive that the speaker's accent has changed, **but not their voice.**" EX1004-*Nguyen*, [0016]. "Such functionality can enable participants in a virtual conference to more easily understand each other, which can significantly enhance the virtual conference experience and enable clear communication between all participants." EX1004-*Nguyen*, [0017]. If more identity characteristics are preserved, the output will more accurately represent the parts of the user's voice that make their voice unique. This furthers the goal of enabling clear communication between all participants by changing the speaker's accent but not their voice.

125. Additionally, *Narayanan* teaches that its voice profile is stored and later updated based on further speech of a speaker. EX1006-*Narayanan*, [0047]. A PHOSITA would have understood that this would have provided an additional advantage to *Nguyen*'s system, allowing the voice profile, and subsequent speech modification, to improve over time. The voice profile would become more personalized over time, allowing for better retention of identity characteristics during accent conversion. *Narayanan* teaches "the spoken audio content in the input audio content 402 may be continuously monitored in real time ... with the voice

profile ... being updated accordingly.” EX1006-Narayanan, [0047]. A PHOSITA would have understood that the more speech data that is analyzed to generate speech characteristics, the more accurate the voice characteristics will be in representing a user’s voice. The more times the accent conversion functionality, or even just the virtual conference provider, is used, the more data the system would have to generate data from, allowing the voice profile to refine the characteristics contained within the profile. EX1005-Dirac, 7:56-60. As a result, an accent conversion or speech modification performed on the user’s speech will retain more accurate characteristics of the user’s voice, improving the similarity of the output to the user’s voice. As the retention of identity characteristics improves due to an improving voice profile, the output will increasingly sound like the original speaker. EX1006-Narayanan, [0043]-[0044], [0047]-[0049].

126. Further, such a combination would have constituted applying a known technique (e.g., *Narayanan’s* generating voice profiles for speakers) to a known device (e.g., *Nguyen’s* accent conversion system) ready for improvement to yield predictable results (e.g., using voice profiles to retain the specific identity characteristics of the speaker’s true voice for better retention during accent conversion).

127. Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because *Nguyen* already teaches the retaining

identity characteristics of the speaker during accent conversion, while *Narayanan* provides further implementation details of generating a voice profile containing the speaker characteristics to retain those characteristics in speech modification. *Nguyen* teaches “a virtual conference provider may maintain one or more trained machine-learning (“ML”) models that can accept speech that has one accent and convert it to speech having another accent, while retaining both the content of the original speech and the identity characteristics of the original speaker’s voice. EX1004-Nguyen, [0013]. *Narayanan* teaches that a voice profile is determined based on various characteristics, including intonation, and that speech may be generated based on the voice profile that sounds like the speaker the voice profile was based on. EX1006-Narayanan, [0004], [0075]. A PHOSITA would have understood that only minor changes to the software of *Nguyen* are required to generate *Narayanan*’s voice profiles to retain the identity characteristics of the speaker. Incorporating the use of *Narayanan*’s voice profiles into *Nguyen*’s system for accent conversion would have been within the knowledge and experience of a PHOSITA. In fact, *Nguyen* already teaches the use of profiles that include accent information, so a PHOSITA would have understood to modify these profiles to be *Narayanan*’s voice profiles. EX1004-Nguyen, [0090], [0095], [0108]. A PHOSITA that this would have required only simple changes to *Nguyen*’s system.

128. In the *Nguyen-Kee-Dirac-Narayanan* combination, the accent conversion system of *Nguyen* creates a voice profile by analyzing the speech of the second participant as taught by *Narayanan*. In the combination, the accent conversion system taught by *Nguyen* receives audio in the form of speech from a second participant. It analyzes the audio to generate a voice profile, including the second participant's audio characteristics such as intonation, pitch, and timbre. EX1006-Narayanan, [0004], [0006], [0023], [0044]; EX1004-Nguyen, [0073]. This voice profile is used to retain the identity characteristics of the second participant in the accent conversion as described below in Claim 1[c]. Thus, the *Nguyen-Kee-Dirac-Narayanan* combination teaches Claim 1[b], by analyzing the obtained input audio data to generate characteristics corresponding to the vocal traits of a second user contained in a voice profile.

129. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious Claim 1[b].

4. ***Claim 1[c]: “synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent; and”***

130. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *synthesiz[ing] a modified version of the second input speech* (e.g., *Nguyen's*

generating an output audio stream with modified speech of the second participant having a target accent) *by modifying the obtained second input audio data based on the generated characteristics* (e.g., *Narayanan's* voice profile which contains characteristics of the speaker's specific speech) *and the extracted accent features* (e.g., *Dirac's* audio characteristics of an accent), *wherein the modified version of the second input speech preserves aspects of the natural voice of the second user* (e.g., *Nguyen's* output audio stream which has the same voice characteristics as the original speaker) *and mimics the first accent* (e.g., *Nguyen's* accent conversion to the target accent), as claimed.

131. As discussed above, the combination of *Nguyen*, *Kee*, and *Dirac* teaches extracting accent features of a first input audio data associated with a first accent of a first participant as part of an accent detection process. *See Claim 1[a]*. The further combination of *Nguyen*, *Kee*, *Dirac*, and *Narayanan* teaches generating characteristics, stored in a voice profile, corresponding to the vocal traits of a second participant by analyzing a second input audio. *See Claim 1[b]*. As explained below, the combination further teaches generating *a modified version of the second input speech* that is based on *the extracted accent features* of the *first user* and *the generated characteristics* of the *second user*.

132. *Nguyen* teaches that accent detection is used to detect the source and target accents in order to determine which AC model to use for the accent

conversion. EX1004-Nguyen, [0090], [0098], [0108]; *see* Claim 1[a]. When performing the accent conversion, *Nguyen* teaches the system “generates, using an AC model 314, a second audio stream based on a received audio stream.” EX1004-Nguyen, [0112]. *Nguyen* teaches that the “AC functionality 352 can receive incoming audio streams ***that have speech with a source accent and convert it to speech having a target accent, while retaining the identity characteristics*** of the original speaker.” EX1004, [0072]. “Because the accent conversion retains the identity characteristics of the original speaker, but changes the accent, the requesting participant will simply perceive that the speaker’s accent has changed, but not their voice.” EX1004-Nguyen, [0016]. Thus, *Nguyen* teaches an output audio stream with a different accent than the speech of the input audio stream (*a modified version of the second input speech*).

133. A PHOSITA would have understood that an example of this functionality, consistent with the example provided in the mappings for Claims 1[a] and 1[b], is as follows: A first and second participant speak during a video conference. EX1004-Nguyen, [0090], [0095]; *see* Claim 1[a]. *Nguyen* teaches communication between participants. EX1004-Nguyen, [0017]. The accent conversion system of *Nguyen* detects the participants’ accents based on the accent characteristics of the participants’ respective speech. EX1004-Nguyen, [0090], [0095]; *see* Claim 1[a]. *Nguyen* teaches “the virtual conference provider may

automatically detect accents.” EX1004-Nguyen, [0090]. If the accents are different, the first participant can accept a prompt from the system to convert the speech of the second participant having a source accent into speech with a target accent (the accent of the first participant). EX1004-Nguyen, [0095], Fig. 5D; *see* Claim 1[b]. *Nguyen* teaches that the virtual conference provider provides an indication suggesting accent conversion. EX1004-Nguyen, [0095]. The accent conversion model is selected by the accent conversion system based on the detected accent of each participant. EX1004-Nguyen, [0108], [0065]. *Nguyen* teaches that accent conversion is performed “so long as a suitable AC model is available.” EX1004-Nguyen, [0108]. The audio stream of the second participant is modified by the accent conversion model to produce a second audio stream containing the identity characteristics of the second participant but with the accent of the first participant. EX1004-Nguyen, [0072], [0013], [0016]; *see* Claim 1[b]. *Nguyen* teaches that the virtual conference provider generates an audio stream that has the “selected accent while having the same voice characteristics as the original speaker.” EX1004-Nguyen, [0112], [0114].

134. The only differences between *Nguyen* and the *Nguyen-Kee-Dirac-Narayanan* combination are that the accent detection is performed using the method of *Dirac*, and identity characteristics are maintained by creating a voice profile as taught by *Narayanan*. A PHOSITA would have understood that in the combination,

the second audio stream would have been *based on* the generated voice profile (containing the identity characteristics) and the extracted accent features, which are included within the AC model as the output accent. A PHOSITA would have understood that the accent features extracted to determine the accent of the speakers are used to determine the appropriate AC model for performing the accent conversion. *See* Claim 1[a]; EX1005-Dirac, 8:58-64. A PHOSITA also would have understood that the characteristics contained in the voice profile would have been used to ensure the output sounds like the second user. *See* Claim 1[b]; EX1006-Narayanan, [0043]. Specifically, because the AC model is selected based on the extracted accent features, the output of the AC model is *based on* the extracted accent features. A PHOSITA would have understood that the output of the accent conversion functionality is based on the AC model, which is selected by processing the extracted accent features. Additionally, the output audio stream has speech with identity characteristics that are retained based on the identity characteristics within the voice profile that mimics the target accent. A PHOSITA would have understood that the voice profile is used to ensure that the output of the accent conversion system sounds like the original speaker (retains identity characteristics). Thus, *Nguyen-Kee-Dirac-Narayanan* teaches *synthesiz[ing] a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of*

the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.

135. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious Claim 1[c].

5. ***Claim 1[d]: “provide to the audio interface output audio data for output via the audio output device, wherein the output audio data is generated based on the modified version of the second input speech.”***

136. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *provid[ing] to the audio interface (e.g., Kee’s audio interface) output audio data (e.g., Nguyen’s generated second audio stream) for output via the audio output device (e.g., Nguyen’s speaker or headphones), wherein the output audio data is generated based on the modified version of the second input speech*, as claimed.

137. As established above, *Nguyen* teaches that a second audio stream containing modified speech is generated by the system. *See* Claim 1[c]. *Nguyen* further teaches that after accent conversion, “the generated audio output stream(s) can then be provided to the requesting participants instead of the original unaltered audio stream.” EX1004-*Nguyen*, [0070]. *Nguyen* teaches “transmit[ing] the second audio stream to speakers or headphones to be output to the user.” EX1004-*Nguyen*, [0115]. As established in Claim 1[Pre], the *Nguyen-Kee* combination teaches an audio interface that has analog-to-digital and digital-to-analog conversion

functionalities. *See* Claim 1[Pre]. A PHOSITA would have understood that in such a combination, the processed audio stream would have been provided to the audio interface to be converted to analog, allowing it to be output to the user. A PHOSITA would have understood that the processed audio would have to be converted from digital to analog form in order to be output to the user by means of a speaker or headphones. *See* section V.E. Digital data must be converted to analog form before it can be heard. EX1024-Ambardar, 1-2. In the *Nguyen-Kee-Dirac-Narayanan* combination the audio interface provides the digital-to-analog conversion. *See* Claim 1[Pre]. Further, a PHOSITA would have understood that the generated second audio stream (*output audio data*) contains the modified speech having the target accent and therefore *is generated based on the modified version of the second input speech*. A PHOSITA would have understood that the output audio is digital audio output from the accent conversion system (*the modified version*), which would have the accent of the first user and the natural voice characteristics of the second user. *See* Claim 1[c].

138. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious Claim 1[d].

B. Claim 2

1. ***Claim 2: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to extract from the first input audio data one***

or more prosodic features, linguistic features, or global speaker characteristics.”

139. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *wherein the one or more processors* (e.g., *Nguyen’s* processors) *are further configured to execute the instructions* (e.g., *Nguyen’s* processor-executable instructions) *to extract* (e.g., *Dirac’s* obtain) *from the first input audio data* (e.g., *Nguyen’s* audio stream of a first participant) *one or more prosodic features* (e.g., *Dirac’s* pitch), *linguistic features, or global speaker characteristics*, as claimed.

140. The ’756 Patent explains that prosodic features include pitch and timbre. EX1001, 4:20. As established in Claim 1[a], *Nguyen* in view of *Dirac* renders obvious extracting accent features from the first user’s speech using machine learning models. *See* Claim 1[a]. *Dirac* teaches that during accent detection “audio communications from the first party may be analyzed to obtain audio characteristics (e.g., ***pitch***, tone, stress, melody, etc.) of the first party’s accent” EX1005-*Dirac*, 8:59-61. Therefore, because pitch, as taught by the ’756 Patent, is a prosodic feature, *Dirac* teaches obtaining (*extracting*) one or more prosodic features.

141. In a *Nguyen-Kee-Dirac-Narayanan* combination, *Nguyen’s* system utilizes *Dirac’s* step of obtaining audio characteristics to *extract one or more prosodic features* (e.g., pitch).

142. Therefore, in my opinion, *Nguyen* in view of *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious Claim 2 of the ’756 Patent.

C. Claim 3

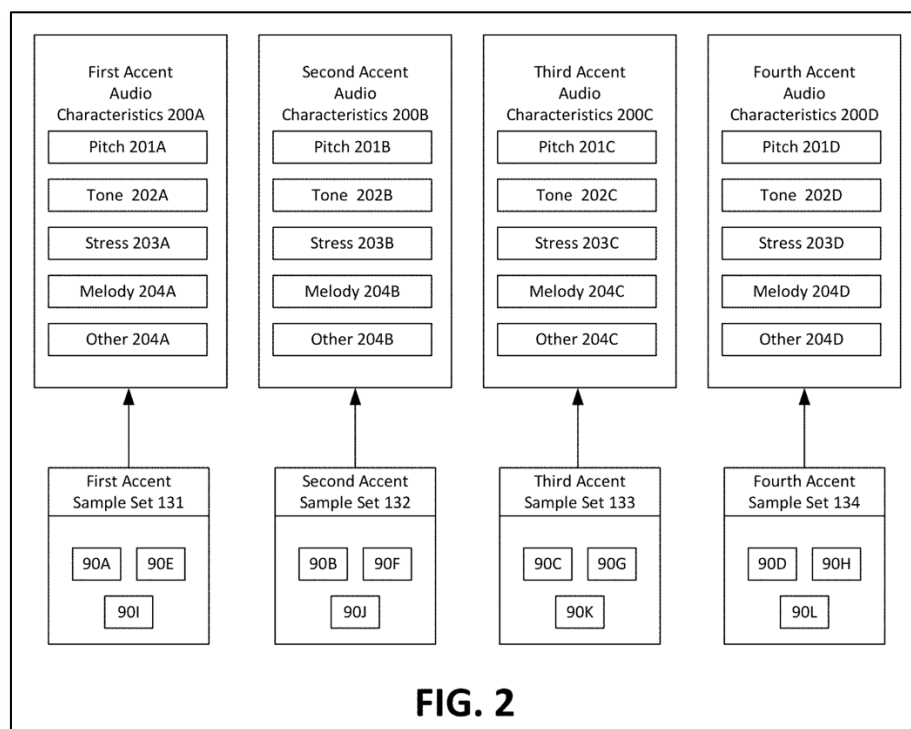
1. ***Claim 3: “The speech processing system of claim 1, wherein the accent features comprise one or more pitch contours, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”***

143. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *wherein the accent features* (e.g., *Dirac*’s audio characteristics of an accent) *comprise one or more pitch contours* (e.g., *Dirac*’s melody), *other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech* (e.g., *Dirac*’s melody throughout the speech), *the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent*, as claimed.

144. As established in Claim 1[a], *Nguyen* in view of *Kee* and *Dirac* renders obvious *apply[ing] one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user*. See Claim 1[a].

145. The ’756 Patent explains that “pitch contours comprise variations in pitch throughout the first input speech.” EX1001, 10:10-15, 4:10-11.

146. *Dirac* teaches to determine the accent, “audio communications from the first party may be analyzed to obtain audio characteristics (e.g., pitch, tone, stress, melody, etc.) of the first party’s accent.” EX1005-Dirac, 8:59-61; *see* Claim 1[a]. Examples of these characteristics, which represent an accent (*accent features*), are shown in Fig. 2. EX1005-Dirac, 5:9-27.



EX1005-Dirac, Fig. 2. As shown, these characteristics include melody. *Id.* A PHOSITA would have understood that melody, the rise and fall of pitch in a communication, is the pitch contours taught by Claim 3 of the '756 Patent. A PHOSITA would have understood that melody, when referred to in the speech processing field, is the rise and fall of pitch in speech communication. *See* section V.B.b; EX1015-Chow, 2. Accordingly, *Nguyen* in view of *Kee*, *Dirac*, and

Narayanan renders obvious wherein the accent features comprise one or more pitch contours ... and the pitch contours comprise variations in pitch throughout the first input speech.

147. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* render obvious Claim 3.

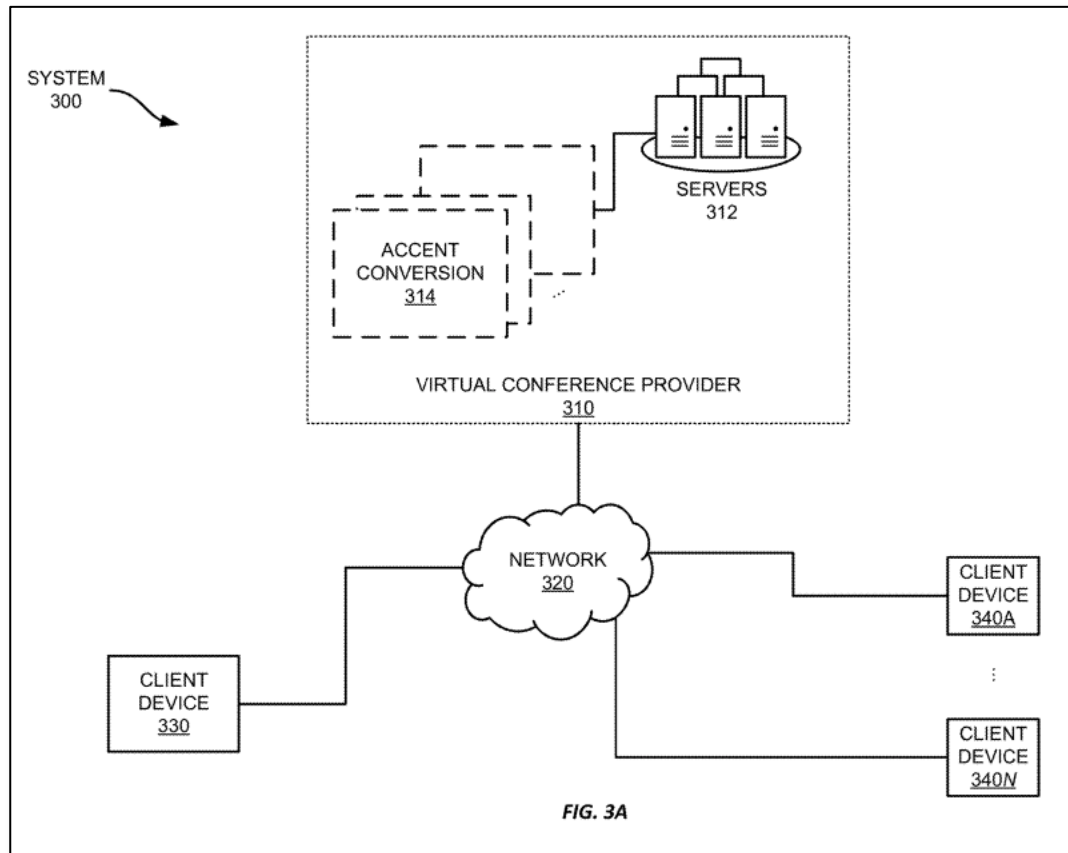
D. Claim 6

1. ***Claim 6: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”***

148. In my opinion *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *wherein the one or more processors (e.g., Nguyen’s processors) are further configured to execute the instructions (e.g., Nguyen’s processor-executable instructions) to receive the second input audio data (e.g., Nguyen’s audio stream from its second participant) via one or more communication networks (e.g., Nguyen’s network) and from a user computing device that is remote from the speech processing system (e.g., Nguyen’s remote client device), wherein the second input audio data is captured at the user computing device (e.g., Nguyen’s audio data that is received from the remote client device), as claimed.*

149. *Nguyen* teaches its system may be a single client device 330 (*speech processing system*). EX1004-*Nguyen*, Fig. 3B, [0071]-[0072]. *Nguyen* further teaches accent conversion “functionality 352 can operate on an audio stream received from the microphone 336 or it can ***operate on one or more audio streams received from other participants in a virtual conference.***” EX1004-*Nguyen*, [0072]. A PHOSITA would have understood that the audio stream received from other participants, such as a second participant, would have been *captured at a microphone of said participant’s device*. A PHOSITA would have understood that the audio must have been converted from analog to digital to be sent to another participant. For the computing device to process and transfer the data, it must have been in digital form. EX1024-Ambardar, 15; *see* Section V.E. A PHOSITA would have understood that analog audio is captured by a microphone and subsequently digitalized for processing. *See* Section V.E. *Nguyen* discloses that each client device transmits video and audio streams to other client devices through the virtual conference provider 310 via network 320 and receives streams from other client devices through the virtual conference provider via the network. EX1004-*Nguyen*, [0071], Fig. 3A. Further, a PHOSITA would have understood that the other client devices are *remote* from the first client device. EX1004-*Nguyen*, [0072]. The client devices receive streams through a network. Networks are used to transfer information between two devices that are physically remote from each other.

EX1004-Nguyen, [0012] (“virtual conferences allow participants to connect from anywhere in the world.”); Fig. 3A.



EX1004, Fig. 3A.

150. A PHOSITA would have understood that the second input audio data (the audio accent conversion is performed on) would have been received by the first client device from a remote device. A PHOSITA would have understood that the conference would have been taking place between participants that are in different physical locations and the audio and video streams are transferred from the capturing device to the devices of other participants via a network. EX1004-Nguyen, [0012],

Fig. 1. Accordingly, the second participant captures audio, which is sent to the first participant's device that contains the virtual conference application. The virtual conference application on the first client device performs the accent conversion. *Nguyen* teaches that performing the accent conversion on the client device allows for virtual conferences that are end-to-end encrypted, because in such conferences the provider does not have access to the unencrypted audio. EX1004-Ngyuen, [0072]. Thus, *Nguyen* teaches *receiv[ing] the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.*

151. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious Claim 6 of the '756 Patent.

X. OPINIONS REGARDING GROUND II: *NGUYEN*, *KEE*, *DIRAC*, *NARAYANAN*, AND *KAMRUZZAMAN*

A. Claim 4

1. ***Claim 4: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint."***

152. In my opinion *Nguyen* in view of *Kee*, *Dirac*, *Narayanan*, and *Kamruzzaman* renders obvious *wherein the one or more processors* (e.g., *Nguyen's*

processors) *are further configured to execute the instructions* (e.g., *Nguyen's* processor-executable instructions) *to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint* (e.g., *Kamruzzaman's* MFCC analysis to create a support vector machine) *of the voice of the second user* (e.g., *Nguyen's* second participant), *wherein the generated characteristics comprise the unique fingerprint* (e.g., *Kamruzzaman's* MFCC derived support vector incorporated into *Narayanan's* voice profile), as claimed.

153. The '756 Patent explains that "an MFCC analysis extracts a unique fingerprint." EX1001, 7:62.

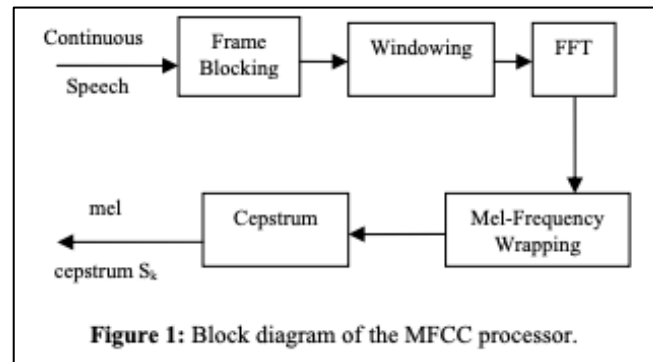
a. Kamruzzaman's Teachings

154. *Kamruzzaman* teaches the use of mel-frequency cepstrum coefficients for speaker identification. EX1011-Kamruzzaman, Abstract. The cepstrum coefficients represent speaker characteristics. EX1011-Kamruzzaman, Abstract. Regarding the use of MFCCs *Kamruzzaman* discloses:

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear prediction coding (LPC), mel-frequency cepstrum coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and this feature has been used in this paper. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency.

EX1011-Kamruzzaman, 1-2.

155. Fig. 1 is “a block diagram of the structure of an MFCC processor” which a PHOSITA would have understood outlines the process for performing an MFCC analysis. EX1011-Kamruzzaman, 2.



EX1011-Kamruzzaman, Fig. 1. A PHOSITA would have understood that generally, the steps for completing an MFCC analysis are: 1) Framing, 2) Windowing or Hamming, 3) Applying the Fourier Transform, 4) Applying mel-filter banks, and 5) Applying the Discrete Cosine Transformation. EX1021-Joshy, 2-4; EX1022-Giroti, 2. “Framing is the process of segmenting the digitally converted speech signal received as input into small frames.” EX1021-Joshy, 3. Hamming decreases the amplitude of the audio as it reaches the frame’s border “to avoid any discontinuities.” EX1022-Giroti, 2. Next, the Fast Fourier transform is applied. EX1021-Joshy, 3. This converts the data from the time domain to the frequency domain. EX1021-Joshy, 3. Next, is the mel filter banks or band-pass filters. EX1021-Joshy, 3; EX1022-Giroti, 2. The filter banks are triangular and are applied to the output of the

Fourier transform. EX1021-Joshy, 3. The filter banks are placed logarithmically, or equally spaced on the mel frequency scale, to be tailored to the spectrums humans hear with greater accuracy. EX1021-Joshy, 3; EX1022-Giroti, 2. The output of applying the filter banks is a set of numbers based on the number of filter banks used. EX1022-Giroti, 2. Finally, the cepstrum is determined by taking the log of the power spectrum and applying the Discrete Cosine transform. EX1021-Joshy, 3-4. The output of this analysis is a set of coefficients that represent the audio qualities of the audio data. EX1021-Joshy, 4; EX1022-Giroti, 2, EX1011-Kamruzzaman, 2.

156. *Kamruzzaman* discloses that by performing the steps of Fig. 1, “a set of mel-frequency cepstrum coefficients is computed.” EX1011-Kamruzzaman, 2. “This set of coefficients is called an acoustic vector,” which is “used to represent and recognize the voice characteristic of the speaker.” EX1011-Kamruzzaman, 2. A PHOSITA would have understood that the acoustic vector representation of the speaker’s voice characteristics contains speaker identifying information. See EX1001, 7:62. A PHOSITA would have understood that MFCCs represent the timbre signature of an individual, which is used for the recognition of individual speakers. *See* section V.B.c; EX1017-Piazza, 1, 4; EX1021-Joshy, 2. “Timbre as a feature of natural speech is consistent within individuals and distinct between individuals.” EX1017-Piazza, 8. Further, in my opinion, because *Kamruzzaman* describes using a Support Vector Machine (SVM) which uses the variability from

its MFCCs, *Kamaruzzaman* teaches the claimed (*MFCC*) *analysis to extract a unique fingerprint*. EX1011-Kamruzzaman, 1-2. Specifically, *Kamaruzzaman* discloses training an SMV which uses the variability in the MFCC samples for application in future speech recognition processing. EX1011-Kamruzzaman, 1-2, §3.1. Accordingly, this is one exemplary way which shows how *Kamruzzaman* applies MFCC techniques to capture speaker-specific information in order to construct a global support vector, such as for use in a voice profile, such as *Narayanan's*.

b. Motivation to Combine

157. *Nguyen-Kee-Dirac-Narayanan* does not expressly teach using voice characteristics extracted with an MFCC analysis in the voice profile. However, *Kamruzzaman* does teach using MFCC coefficients to represent the voice characteristics of a speaker.

158. A PHOSITA would have been motivated to incorporate *Kamruzzaman's* MFCC analysis into the *Nguyen-Kee-Dirac-Narayanan* system's voice profile with a reasonable expectation of success for several reasons. In addition to the reasons below, "MFCC is the most commonly used feature extraction technique in Speaker Recognition and Verification." EX1021-Joshy, 2. Naturally, a PHOSITA would have known to turn to MFCCs for generating characteristics

unique to a user. EX1011-Kamruzzaman, 2 (MFCC is perhaps the best known and most popular way to represent speech signal).

159. First, *Nguyen* specifically states that to alleviate communication issues, the system converts one accent to another accent while retaining the identity characteristics of the original speaker. EX1004-*Nguyen*, [0013]. However, *Nguyen* does not provide a detailed explanation of what characteristics are identity characteristics, nor does *Nguyen* provide a list of characteristics to preserve. *Nguyen* merely provides that “the AC model 412 will be trained to retain the speaker’s identity characteristics.” EX1004-*Nguyen*, [0073]. The only two identity characteristics provided are “timbre and pitch.” EX1004-*Nguyen*, [0073]. While *Narayanan* provides characteristics, such as audio spectral patterns and intonation, to be included in a speaker’s voice profile, *Narayanan* does not provide specifics for how the various traits are determined and represented. *Narayanan* simply teaches that “the voice profile module 106 may determine the voice profile based on the spoken audio content.” EX1006-*Narayanan*, [0023], [0024], [0043], [0044], [0075]. Accordingly, and based on these references’ disclosures, a PHOSITA would have been motivated to look to an additional reference, such as *Kamruzzaman*, for specific implementation details of obtaining features/characteristics from a speech signal to incorporate into the voice profile. EX1011-Kamruzzaman, 1. *Kamruzzaman* is one example which provides that, and explains how MFCCs represent the voice

characteristics of a speaker. EX1011-Kamruzzaman, 2. Thus, a PHOSITA would have looked to *Kamruzzaman* for specifics on how features are determined and represented. EX1011-Kamruzzaman, 2. *Kamruzzaman* provides step-by-step instructions on how to obtain MFCCs and how they can be used. *See generally* EX1011-Kamruzzaman. Use of MFCCs to represent the voice characteristics of a speaker was well known in the art at the time of the '756 Patent. For example, "MFCC is the most commonly used feature extraction technique in Speaker Recognition and Verification." EX1021-Joshy, 2.

160. Second, such a combination would have constituted applying a known technique (e.g., *Kamruzzaman's* use of MFCCs) to a known device (e.g., *Nguyen-Narayanan's* voice profiles) ready for improvement to yield predictable results (e.g., extracting improved identity characteristics). Because MFCCs are based on the human ear's ability to hear certain frequencies, a PHOSITA would have understood that providing the voice profile with data that represents the voice characteristics of the speaker based on humans' frequency perception would have improved the performance of the accent conversion model. EX1011-Kamruzzaman, 2. *Joshy* teaches that "MFCC is based on human hearing system which cannot exactly perceive frequencies above 1 KHz." EX1021-Joshy, 2. *Joshy* further teaches that "the Mel Scale Filtering is applied so as to obtain frequency ranges resembling that of normal human voice signals." EX1021-Joshy, 3. Retaining voice characteristics

perceptible to a human would have improved the resulting output audio with a modified accent to sound more like the original speaker. A PHOSITA would have understood that the user of the accent conversion system would have been more perceptive of changes in the frequency range in which MFCCs put the most emphasis. EX1022-Giroti, 2 (MFCCs “use non-linear frequency scale based on auditory perceptions of human ear.”).

161. Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because the *Nguyen-Kee-Dirac-Narayanan* combination already teaches generating characteristics stored in the voice profile, and *Kamruzzaman* simply provides an additional, and well-known, method to generate characteristics, which a PHOSITA would have readily been able to implement. A PHOSITA would have understood that *Narayanan's* voice profile may be based on vocal characteristics, and the MFCC coefficients represent such voice characteristics. EX1006-Narayanan, [0043]; EX1011-Kamruzzaman, 2. A PHOSITA would have understood how to incorporate MFCCs to the existing data in *Narayanan's* voice profiles. A PHOSITA would have understood that it would have required only minor changes to the software of the *Nguyen-Kee-Dirac-Narayanan* combination to implement an MFCC analysis to generate characteristics for the voice profile. The addition of MFCCs into *Narayanan's* voice profiles would have been well within the skill of a PHOSITA at the time of the '756 Patent.

162. In the *Nguyen-Kee-Dirac-Narayanan-Kamruzzaman* combination, the accent conversion system taught by *Nguyen* generates a voice profile (of *Narayanan*) by extracting voice characteristics of the second participant through an MFCC analysis, as taught by *Kamruzzaman*. Specifically, the accent conversion functionality of the client device taught by *Nguyen* receives audio in the form of speech from a second participant, performs the MFCC analysis taught by *Kamruzzaman*, and includes the resulting acoustic vectors (*unique fingerprint*) in the voice profile. Therefore, in the *Nguyen-Kee-Dirac-Narayanan-Kamruzzaman* combination, the generated characteristics comprise the unique fingerprint. Accordingly, *Nguyen* in view of *Kee*, *Dirac*, *Narayanan*, and *Kamruzzaman* teaches this claim.

163. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, *Narayanan*, and *Kamruzzaman* render obvious Claim 4 of the '756 Patent.

XI. OPINIONS REGARDING GROUND III: NGUYEN, KEE, DIRAC, NARAYANAN, AND WANG

A. Claim 5

1. *Claim 5: “The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.”*

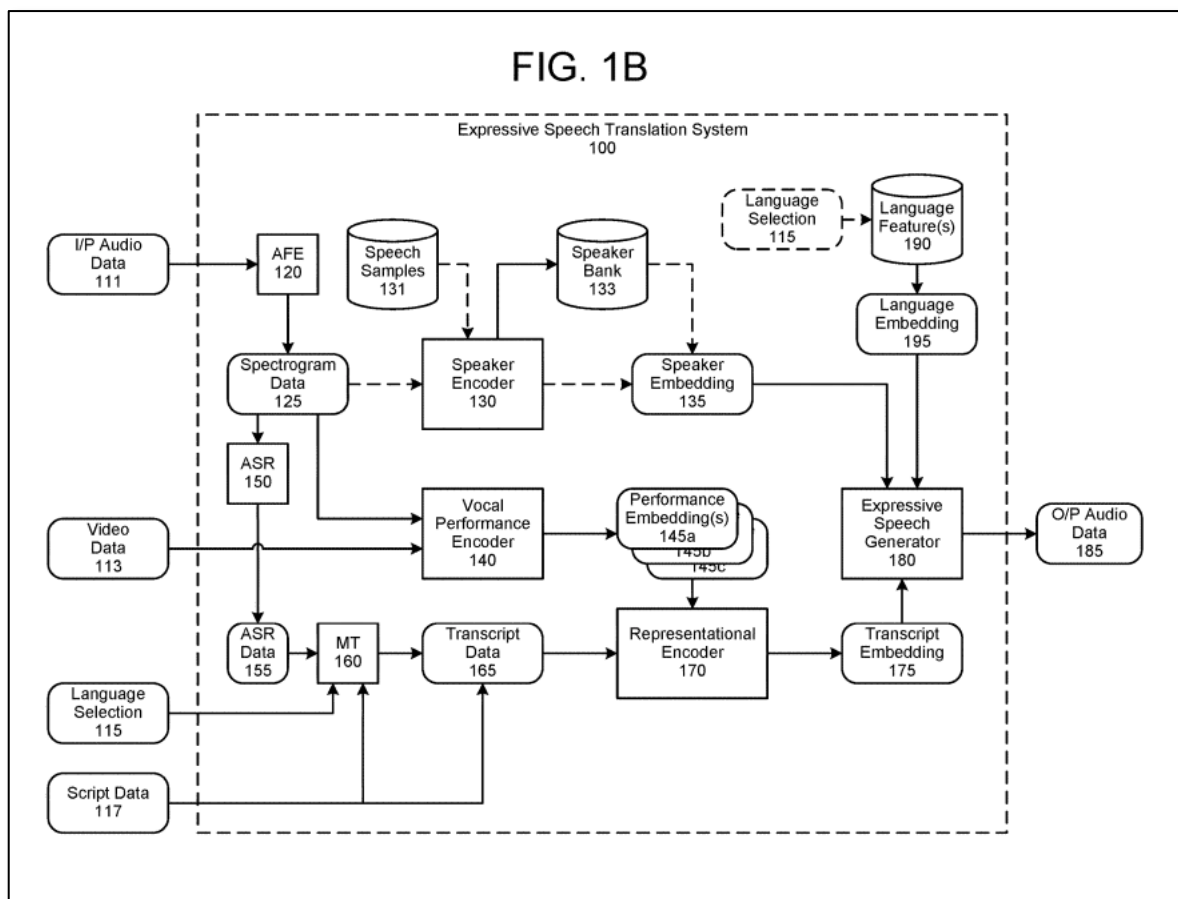
164. *Nguyen in view of Kee, Dirac, Narayanan, and Wang renders obvious wherein the one or more processors (e.g., Nguyen's processors) are further configured to execute the instructions (e.g., Nguyen's processor-executable instructions) to apply a speaker identity encoding technique (e.g., Wang's use of a speaker encoder) to encode speaker-specific voice characteristics (e.g., Narayanan's characteristics of the speaker's specific speech are encoded with Wang's speaker encoder to generate speaker embedding data), wherein the generated characteristics comprise the speaker-specific voice characteristics (e.g., Narayanan's voice profile comprises the speaker embedding data generated from applying Wang's speaker encoder to Narayanan's characteristics of the speaker's specific speech), as claimed.*

165. The '756 Patent explains that the speaker-specific voice characteristics may include pitch, timbre, and/or rhythm. EX1001, 6:35-36. *Narayanan* teaches that the voice characteristics stored in the voice profile “may include audio spectral patterns, vocal frequency, **vocal pitch**, speaking speed, **intonation**, loudness, amplitude, **speech patterns**, **speech cadence**, the number and/or duration of utterances, the number and/or duration of breaks between utterances, etc.” Thus, *Narayanan's* characteristics of the speaker's specific speech stored in the voice profile are *speaker-specific voice characteristics* and are part of the *generated*

characteristics. Narayanan does not expressly teach *apply[ing] a speaker identity encoding technique to encode speaker-specific voice characteristics*.

a. Wang's Teachings

166. Wang teaches “an expressive speech translation system [that] may process source speech in a source language and output synthesized speech in a target language while retaining vocal performance characteristics such as intonation, emphasis, rhythm, style, and/or emotion.” EX1007-Wang, Abstract. Fig. 1B depicts the expressive speech translation system.



EX1007-Wang, Fig. 1B. *Wang* teaches that the system may use the various components to “disentangle” the voice and speech characteristics from the identity of the speaker and the language spoken. EX1007-Wang, [0028]. This functionality is accomplished using “a language feature storage component 190 that stores language embedding data 195 for various languages/dialects/locales/**accents**/etc.” EX1007-Wang, [0029]. Therefore, *Wang* teaches that the speech translation system is applied to speech with different accents.

167. Specifically, regarding the speaker encoding, *Wang* teaches:

The expressive speech translation system 100 may include a **speaker encoder 130** for generating **speaker embedding data 135**. The speaker embedding data 135 represents **voice identity characteristics** that differ from speaker to speaker such as **fundamental frequency (e.g., pitch), timbre, cadence**, etc. In some implementations, the expressive speech translation system 100 may use the speaker encoder 130 to generate speaker embedding data 135b using the input audio data; thus, generating a speaker embedding that represents a speaker of the source speech.

EX1007-Wang, [0030]. The voice identity characteristics of *Wang* overlap with the characteristics described in *Narayanan*. *Wang* describes **pitch**, timbre, and **cadence** as voice identity characteristics. EX1007-Wang, [0030]. *Narayanan*’s characteristics include “audio spectral patterns, vocal frequency, **vocal pitch**,

speaking speed, intonation, loudness, amplitude, speech patterns, *speech cadence*, the number and/or duration of utterances, the number and/or duration of breaks between utterances, etc.” EX1006-Narayanan, [0075]. *Wang* and *Narayanan* each disclose pitch, cadence, and timbre characteristics (such as audio spectral patterns and vocal frequency). Thus, the characteristics in *Narayanan* and *Wang* are largely duplicative. Like the characteristics of *Narayanan*, which are “generally representative of the speaker’s speech,” the voice identity characteristics of *Wang*, which “differ from speaker to speaker,” are vocal traits that are distinct to a speaker and represent speaker-specific voice characteristics. *See* Claim 1[b].

168. *Wang* also teaches “the expressive speech translation system 100 may include a speaker bank 133 that stores speaker embedding data 135a corresponding to different voice identities.” EX1007-Wang, [0030].

169. Finally, *Wang* teaches that “using the speaker embedding data 135 ... an expressive speech generator 180 may synthesize speech that reflects the speech characteristics of the target language while reproducing the original speaker’s expressiveness and preserving their voice identity characteristics.” EX1007-Wang, [0047].

b. Motivation to Combine

170. In the *Nguyen-Kee-Dirac-Narayanan-Wang* combination, the system encodes the voice identity characteristics of the second speaker taught by *Narayanan*

into a speaker embedding, as taught by *Wang*. In such a combination, the speaker embedding would have been included in the voice profile of the second user, as taught by *Narayanan*.

171. *Nguyen-Kee-Dirac-Narayanan* does not expressly teach using speaker encoding to retain speaker identity characteristics. However, *Wang* teaches retaining the speaker identity characteristics through speaker encoding.

172. A PHOSITA would have been motivated to incorporate *Wang's* technique of encoding the speaker identity characteristics into a speaker embedding into the *Nguyen-Kee-Dirac-Narayanan* system's voice profile with a reasonable expectation of success for several reasons. In addition to the reasons stated below, both *Narayanan* and *Wang* teach a voice profile or speaker bank. Compare EX1006-Narayanan, [0075]; with EX1007-Wang, [0030].

173. First, while *Narayanan* provides specific identity characteristics as candidates for preservation, a PHOSITA would have understood that for the identity characteristics to be usable by a machine learning model, the identity characteristics must first be encoded in a usable medium. Computers accept digital data, and characteristics describing vocal quality, such as timbre, are not naturally digital. See Section V.C; EX1024-Ambardar, 16. Naturally, a PHOSITA would have looked to a reference, such as *Wang*, which provides a method for converting identity characteristics into embeddings usable by a machine learning model. A PHOSITA

would have understood that, to preserve the identity characteristics throughout the audio processing, the characteristics would have had to be encoded into a computer-readable format, such as the embeddings of *Wang*. *Wang* teaches an encoder that encodes the identity characteristics into a speaker embedding usable by the processing system to preserve those characteristics when synthesizing data. EX1007-Wang, [0030], [0049].

174. Second, such a combination would have constituted applying a known technique (e.g., *Wang*'s technique for encoding identity characteristics into a speaker embedding) to a known method (e.g., *Nguyen-Kee-Dirac-Narayanan*'s identity characteristics contained in the voice profiles) ready for improvement to yield predictable results (e.g., *Nguyen-Kee-Dirac-Narayanan*'s accent translation system would improve in its ability to process and preserve the identity characteristics contained in the voice profile because via their encoding into a speaker embedding). Further, using encoding to convert data to a form usable by machine learning models was well known in the art at the time of the '756 Patent. For example, *Kashet* teaches using an autoencoder and creating an "embedding of voice characteristics of the speaker". EX1019-Kashet, [0060], [0095].

175. Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because the *Nguyen-Kee-Dirac-Narayanan* combination already teaches generating and preserving characteristics stored in the

voice profile using a machine learning model, and *Wang* simply provides a well-known way to prepare the characteristics for processing by a machine learning model. A PHOSITA would have understood that converting the characteristics into a form usable by machine learning models would have been essential to the audio processing system. *See* Section V.C. A PHOSITA would have further understood that the speaker embeddings taught by *Wang* would have been included in the voice profile taught by *Narayanan*. A PHOSITA would have understood that only minor changes to the software of the *Nguyen-Kee-Dirac-Narayanan* combination are required to encode the identity characteristics into a speaker embedding, as taught by *Wang*, and to store the speaker embedding in a voice profile. This modification would have been well within the skills and knowledge of a PHOSITA at the time of the '756 Patent.

176. Therefore, in my opinion, *Nguyen* in view of *Kee*, *Dirac*, *Narayanan*, and *Wang* renders obvious Claim 5 of the '756 Patent.

XII. OPINIONS REGARDING GROUND IV: *NGUYEN, DIRAC, NARAYANAN, AND KAMRUZZAMAN*

A. Claim 7

- 1. Claim 7[Pre]:** “*A method for real-time accent mimicking, the method implemented by a speech processing system and comprising:*”

177. In my opinion, *Nguyen* discloses, or at least renders obvious, *a method for real-time accent mimicking* (e.g., *Nguyen's* accent conversion method), *the*

method implemented by a speech processing system (e.g., *Nguyen's system for accent conversion for virtual conferences*), as claimed.

178. As established in Claim 1[Pre], *Nguyen* discloses a speech processing system. *See* Claim 1[Pre]. *Nguyen* teaches a *method* for accent conversion to convert a second participant's accent to a first participant's accent (e.g., *accent mimicking*) for virtual conferences that is implemented by the *speech processing system*. EX1004-*Nguyen*, Title, Abstract; *see* Claim 1[Pre]. *Nguyen* also teaches that during a conference, participants communicate with each other and the virtual conference provider. EX1004-*Nguyen*, [0025], [0027]. This communication happens over a “real-time media server.” EX1004-*Nguyen*, [0052]-[0054]. *Nguyen* further teaches that the accent conversion process receives “audio streams from the virtual conference and each generates an output audio stream having the corresponding target accent.” EX1004-*Nguyen*, [0070]. *Nguyen* teaches that “the generated output audio stream(s) can then be provided to the requesting participants instead of the original unaltered audio stream.” EX1004-*Nguyen*, [0070]. Because the system is performing the accent conversion during a virtual conference, over a real-time media server, a PHOSITA would have understood that the accent conversion method occurs in real-time. A PHOSITA would have understood that a video conference requires real-time audio and video, so that participants can effectively communicate with one another. For example, the virtual conferencing system of *Nguyen* uses the

real-time media server to “enable multiple different streams of varying quality, thereby allowing client devices to change streams in real-time as needed, e.g., based on network bandwidth, latency, etc.” which allows for effective real-time communication. EX1004-Nguyen, [0044]. The real-time media server realizes this requirement by being adaptive, “reacting to real-time network and client changes” and monitoring “parameters such as a client’s bandwidth CPU usage, memory and network I/O as well as network parameters such as packet loss, latency and jitter to determine how to modify the way in which streams are provided.” EX1004-Nguyen, [0041]. A PHOSITA would have understood that the accent conversion functionality of *Nguyen* is implemented in the virtual conference provider and thus, must be performed in real-time.

179. Therefore, in my opinion, *Nguyen* discloses Claim 7[Pre].

2. ***Claim 7[a]: “applying one or more trained machine learning models to first input audio data to extract accent features of first input speech associated with a first accent of a first user;”***

180. See Claim 1[a].

3. ***Claim 7[b]: “analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise a unique fingerprint of the voice of the second user;”***

181. See Claim 4.

4. *Claim 7[c]: “synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and”*

182. See Claim 1[c].

5. *Claim 7[d]: “providing output audio data generated based on the modified version of the second input speech.”*

183. See Claim 1[d].

B. Claim 8

1. *Claim 8: “The method of claim 7, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.”*

184. See Claims 1[c] and 7.

C. Claim 9

1. *Claim 9: “The method of claim 7, further comprising extracting from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.”*

185. See Claims 2 and 7.

D. Claim 10

1. *Claim 10: “The method of claim 7, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech. The intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”*

186. See Claims 3 and 7.

E. Claim 11

1. *Claim 11: “The method of claim 7, further comprising applying a mel frequency cepstral coefficient (MFCC) analysis to extract the unique fingerprint of the voice of the second user.”*

187. See Claims 4 and 7.

F. Claim 13

1. *Claim 13: “The method of claim 7, further comprising receiving the second input audio data from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”*

188. See Claims 6 and 7.

XIII. OPINIONS REGARDING GROUND V: NGUYEN, DIRAC, NARAYANAN, AND WANG

A. Claim 14

1. *Claim 14[Pre]: “A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:”*

189. In my opinion, *Nguyen* discloses, or at least renders obvious, *a non-transitory computer-readable medium (e.g., Nguyen’s non-transitory computer readable media) comprising instructions (e.g., Nguyen’s processor-executable instructions stored on the media) that, when executed by at least one processor (e.g., Nguyen’s processor which executes the stored instructions), cause the at least one processor (e.g., Nguyen’s processor that performs a function according to the instructions) to, as claimed.*

190. *Nguyen* teaches that “one or more non-transitory computer-readable media, that may store processor-executable instructions that, when executed by the processor, can cause the processor to perform methods according to this disclosure as carried out, or assisted, by a processor.” EX1004-*Nguyen*, [0121]; *see* Claim 1[Pre].

2. ***Claim 14[a]: “apply one or more trained machine learning models to first input audio data obtained via a microphone to extract accent features of first input speech associated with a first accent of a first user;”***

191. *See* Claim 1[a].

3. ***Claim 14[b]: “analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise speaker-specific voice characteristics;”***

192. *See* Claim 5.

4. ***Claim 14[c]: “synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and”***

193. *See* Claim 1[c].

5. ***Claim 14[d]: “provide for output via an audio output device output audio data generated based on the modified version of the second input speech.”***

194. *See* Claim 1[d].

B. Claim 15

1. *Claim 15: “The non-transitory computer-readable medium of claim 14, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first user.”*

195. See Claims 1[c] and 14.

C. Claim 16

1. *Claim 16: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.”*

196. See Claims 2 and 14.

D. Claim 17

1. *Claim 17: “The non-transitory computer-readable medium of claim 14, wherein the accent features comprise one or more pitch contour, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.”*

See Claims 3 and 14.

E. Claim 19

1. *Claim 19: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor to apply a speaker identity encoding technique to encode the speaker-specific voice characteristics.”*

197. See Claims 5 and 14.

F. Claim 20

1. *Claim 20: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.”*

198. See Claims 6 and 14.

XIV. OPINIONS REGARDING GROUND VI: NGUYEN, DIRAC, NARAYANAN, KAMRUZZAMAN, AND WANG

A. Claim 12

1. *Claim 12: “The method of claim 7, further comprising applying a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.”*

199. See Claims 5 and 7. Claim 12 requires that the generated characteristics comprise “a unique fingerprint of the voice of the second user” (in Claim 7) and “the speaker-specific voice characteristics.” As explained above, a POSITA would have been motivated to include both the MFCCs of *Kamruzzaman* and the encoded speaker characteristics taught by *Wang* in the voice profile of *Narayanan* for the same reasons and with the same reasonable expectation of success provided in Claims 4 and 5.

B. Claim 18

1. *Claim 18: “The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.”*

200. See Claims 4, 12, and 14.

XV. CONCLUSION

201. 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 30, 2026

By: 
Cristopher 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. Anilionyte, 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.
