

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.
Petitioner

v.

Sanas.ai Inc.
Patent Owner

Inter Partes Review Case No. IPR2026-00273
U.S. Patent No. 12,125,496

DECLARATION OF CHRISTOPHER SCHMANDT

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS.....	10
A.	Educational Background and Professional Experience	10
II.	METHODOLOGY AND MATERIALS CONSIDERED	14
III.	OVERVIEW AND LEGAL STANDARDS	18
A.	Person of Ordinary Skill in the Art.....	19
B.	Obviousness	20
C.	Analogous Art	25
D.	Claim Construction.....	26
IV.	LEVEL OF A PERSON OF ORDINARY SKILL.....	26
V.	OVERVIEW OF TECHNOLOGY	27
A.	Neural Networks	27
B.	Feature Extraction and Dimensionality Reduction Techniques...28	
1.	<i>Principal Component Analysis.....</i>	28
2.	<i>Hierarchical Feature Extraction</i>	29
C.	Digital Noise Removal.....	30
VI.	OVERVIEW OF THE '496 PATENT.....	32
A.	Description of the '496 Patent.....	32
B.	Field of Endeavor of the '496 Patent.....	33
C.	Problem Purportedly Solved by the Inventors of the '496 Patent	34
D.	Summary of the Prosecution History.....	37
VII.	OVERVIEW OF THE PRIOR ART.....	38

A.	Overview of Dantrey.....	39
B.	Overview of Strake	42
C.	Overview of Chen.....	45
D.	Overview of Quillen	47
E.	Overview of Heckmann	49
F.	Overview of Liu.....	50
VIII.	SUMMARY OF UNPATENTABILITY	52
IX.	OPINIONS REGARDING GROUND I: <i>Dantrey, Strake, and Chen</i>	53
A.	<i>Claim 1</i>	53
1.	<i>Claim 1[pre]: A voice enhancement system, comprising memory having instructions stored thereon and one or more processors coupled to the memory and configured to execute instructions to:</i>	<i>53</i>
2.	<i>Claim 1[a]: fragment input audio data into a plurality of input speech frames, wherein the input audio data comprises foreground speech content, one or more non-content elements, and one or more speech characteristics;.....</i>	<i>54</i>
3.	<i>Claim 1[b]: convert the input speech frames to low-dimensional representations of the input speech frames, wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data and the low-dimensional representations of the input speech frames omit one or more of the non-content elements;</i>	<i>65</i>
4.	<i>Claim 1[c]: apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; and</i>	<i>75</i>
5.	<i>Claim 1[d]: combine the target speech frames to generate output audio data, wherein the output audio data further</i>	

	<i>comprises one or more portions of the foreground speech content and one or more of the speech characteristics.</i>	<i>80</i>
B.	Claim 2	84
	1. Claim 2[Pre]: The voice enhancement system of claim 1, further comprising a physical microphone and an audio output device, wherein the one or more processors are further configured to execute the instructions to:.....	84
	2. Claim 2[a]: digitize analog input audio signals obtained via the physical microphone to generate the input audio data;	84
	3. Claim 2[b]: convert the output audio data to analog audio output signals; and	86
	4. Claim 2[c]: provide the analog audio output signals to the audio output device via one or more of a virtual microphone or a communication application executed by the voice enhancement system.	89
C.	Claim 3: The voice enhancement system of claim 1, wherein the non-content elements comprise one or more of background noise, microphone pops, low-fidelity audio, or audio clippings and the speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech.....	90
D.	Claim 7: The voice enhancement system of claim 1, wherein the second neural network comprises one or more of a diffusion probabilistic model, a flow-based model, or a generative adversarial network-based model.	91
E.	Claim 8: The voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to pre-process the input audio data by applying one or more of a noise reduction algorithm to remove background noise from the input audio data or a filtering technique to remove high-frequency noise or pops from the input audio data.	92
F.	Claim 9	95

1.	<i>Claim 9[a]: The voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to: extract one or more features from the input audio data, wherein the features comprise one or more of pitch, intonation, or formants; and</i>	<i>95</i>
2.	<i>Claim 9[b]: encode the extracted features into one or more of the low-dimensional representations of the input speech frames using a dimensionality reduction technique.....</i>	<i>97</i>
G.	Claim 16	99
1.	<i>16[pre] A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:</i>	<i>99</i>
2.	<i>16[a] digitize analog input audio signals to generate input audio data;</i>	<i>100</i>
3.	<i>16[b] fragment the input audio data into a plurality of input speech frames, wherein the input audio data comprises foreground speech content, one or more non-content elements, and one or more speech characteristics;.....</i>	<i>100</i>
4.	<i>16[c] convert the input speech frames to low-dimensional representations of the input speech frames, wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data and the low-dimensional representations of the input speech frames omit one or more of the non-content elements;.....</i>	<i>100</i>
5.	<i>16[d] apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames;</i>	<i>100</i>
6.	<i>16[e] combine the target speech frames to generate output audio data, wherein the output audio data further comprises one or more portions of the foreground speech content and one or more of the speech characteristics; and</i>	<i>100</i>

7.	<i>16[ff] convert the output audio data to analog audio output signals before providing the analog audio output signals to an audio output device.</i>	<i>101</i>
H.	Claim 17: The non-transitory computer-readable medium of claim 16, wherein the non-content elements comprise one or more of background noise, microphone pops, low-fidelity audio, or audio clippings and the speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech.....	101
I.	Claim 20: The non-transitory computer-readable medium of claim 16, wherein the second neural network comprises one or more of a diffusion probabilistic model, a flow-based model, or a generative adversarial network-based model.....	101
X.	OPINIONS REGARDING GROUND II: <i>Dantrey, Strake, Chen, and Quillen</i>	101
A.	Claim 4: The voice enhancement system of claim 1 wherein the one or more processors are further configured to execute the instructions to train the first neural network using input audio training data, one or more augmentations, and one or more transcripts, wherein the first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and low-dimensional representations of input audio training data speech frames.....	103
B.	Claim 5: The voice enhancement system of claim 4, wherein the augmentations simulate one or more degraded speech characteristics and comprise one or more of background noise, masked data, microphone pops, smooth speech, or convolving speech.	108
C.	Claim 6: The voice enhancement system of claim 4, wherein the one or more processors are further configured to execute the instructions to train the second neural network using a target speech sample and the low-dimensional representation of input audio training data speech frames, wherein the second neural network is trained to use dynamic conversion to learn a mapping between each of the low-dimensional representation of input	

	audio training data speech frames and a corresponding one of a plurality of target training speech frames.	110
D.	Claim 11	112
	1. <i>Claim 11[Pre]: A method for real-time voice enhancement, the method implemented by a voice enhancement system and comprising:</i>	<i>112</i>
	2. <i>Claim 11[a]: training a first neural network using input audio training data, one or more augmentations, and one or more transcripts and a second neural network using a target speech sample and a plurality of low-dimensional representation of input audio training data speech frames,</i>	<i>112</i>
	3. <i>Claim 11[b]: applying the trained first neural network to convert input speech frames fragmented from input audio data to low-dimensional representations of the input speech frames, wherein the low-dimensional representations of the input speech frames omit one or more non-content elements of the input audio data;.....</i>	<i>113</i>
	4. <i>Claim 11[c]: applying the trained second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; and</i>	<i>113</i>
	5. <i>Claim 11[d]: combining the target speech frames to generate output audio data, wherein the output audio data further comprises one or more portions of foreground speech content of the input audio data and one or more speech characteristics of the input audio data.....</i>	<i>113</i>
E.	Claim 12: The method of claim 11, wherein the trained first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and the low-dimensional representation of input audio training data speech frames	113
F.	Claim 13: The method of claim 11, wherein the augmentations simulate one or more degraded speech characteristics and comprise one or more of background noise, masked data, microphone pops, smooth speech, or convolving speech.....	113

G.	Claim 14: The method of claim 11, further comprising pre-processing the input audio data by applying one or more of a noise reduction algorithm to remove background noise from the input audio data or a filtering technique to remove high-frequency noise or pops from the input audio data.....	114
H.	Claim 15	114
	1. <i>Claim 15[a]: The method of claim 11, further comprising: extracting one or more features from the input audio data, wherein the features comprise one or more of pitch, intonation, or formants; and</i>	114
	2. <i>Claim 15[b]: encoding the extracted features into one or more of the low-dimensional representations of the input speech frames using a dimensionality reduction technique.....</i>	114
I.	Claim 18: The non-transitory computer-readable medium of claim 16, wherein the instructions, when executed by the at least one processor further causes the at least one processor to train the first neural network using input audio training data, one or more augmentations, and one or more transcripts, wherein the first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and low-dimensional representation of input audio training data speech frames.....	114
J.	Claim 19: The non-transitory computer-readable medium of claim 18, wherein the instructions, when executed by the at least one processor further causes the at least one processor to train the second neural network using a target speech sample and the low-dimensional representation of input audio training data speech frames, wherein the second neural network is trained to use dynamic conversion to learn a mapping between each of the low-dimensional representation of input audio training data speech frames and a corresponding one of a plurality of target training speech frames.	115
XI.	OPINIONS REGARDING GROUND III: <i>Dantrey, Strake, Chen, and Heckmann</i>.....	115

A.	Claim 10: The voice enhancement system of claim 9, wherein the one or more processors are further configured to execute the instructions to extract the features using a hierarchical feature extraction network comprises a plurality of levels, wherein each of the levels is configured to capture a different one or more of the features and the captured different one or more of the features are compressed at each of the levels.....	115
XII.	OPINIONS REGARDING GROUNDS IV-VI: <i>Dantrey, Strake, Liu</i> (Ground 4), AND ADDITIONALLY <i>Quillen</i> (Ground 5), and additionally <i>Heckmann</i> (Ground 6).....	123
A.	Liu’s Teachings	123
B.	Specific Claim Limitation Applications for “low-dimensional representation”.....	125
1.	<i>Ground 4 (Dantrey-Strake-Liu)</i>	125
2.	<i>Ground 5 (Dantrey-Strake-Liu-Quillen)</i>	130
3.	<i>Ground 6 (Dantrey-Strake-Liu-Heckmann)</i>.....	132
XIII.	CONCLUSION.....	132

I, Christopher Schmandt, declare as follows:

I. INTRODUCTION AND QUALIFICATIONS

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

A. Educational Background and Professional Experience

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

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

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

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

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

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

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

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

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

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

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

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

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

II. METHODOLOGY AND MATERIALS CONSIDERED

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

13. For this work, I have been asked to review U.S. Patent No. 12,125,496 (“’496 Patent”) (EX1001), including the specification and claims, and the ’496 Patent’s prosecution history (“’496 File History”) (EX1002). In developing my opinions relating to the ’496 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. 12,125,496 (“’496 Patent”)
Exhibit 1002	Prosecution History for the 12,125,496 Patent (“File History”)
Exhibit 1004	U.S. Patent No. 12,412,590 to Dantrey et al. (“Dantrey”)
Exhibit 1005	WIPO Publication No. 2020/199990 to Strake et al. (“Strake”)

Exhibit 1006	U.S. Patent No. 10,867,616 to Chen et al. (“Chen”)
Exhibit 1007	U.S. Patent Application Publication No. 2021/0241780 to Quillen (“Quillen”)
Exhibit 1008	Martin Ernst Heckmann, et al., <i>A Hierarchical Framework for Spectro-Temporal Feature Extraction</i> , Speech Communication, vol. 53, no. 5, pp. 736 - 752, 2011 (“Heckmann”)
Exhibit 1009	R. E. Crochiere, <i>A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis</i> , IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, No. 1, February 1980 (“Crochiere”)
Exhibit 1010	Christopher Schmandt, <i>Voice Communication with Computers Conversational Systems</i> , Van Nostrand Reinhold, (1994) (“Schmandt”)
Exhibit 1011	U.S. Patent Application Publication No. 2024/0098218 to Nguyen et al. (“Nguyen”)
Exhibit 1012	Martinek, R.; Ladrova, M.; Sidikova, M.; Jaros, R.; Behbehani, K.; Kahankova, R.; Kawala-Sterniuk, A. <i>Advanced Bioelectrical Signal Processing Methods: Past, Present and Future Approach—Part II: Brain Signals</i> , Sensors 2021, 21, 6343. https://doi.org/10.3390/s21196343 (“Martinek”)
Exhibit 1013	Casey O’Callaghan, <i>Pitch</i> , http://caseyocallaghan.com/research/papers/Pitch.pdf (“O’Callaghan”)
Exhibit 1014	W. Q. Zheng, J. S. Yu, Y. X. Zou, <i>An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks</i> , International Conference on Affective Computing and Intelligent Interaction (ACII) (2015) (“Zheng”)
Exhibit 1015	Xuechuan Wang, Douglas O’Shaughnessy, <i>Improving the Efficiency of Automatic Speech Recognition by Feature Transformation And Dimensionality Reduction</i> , 10.21437/Eurospeech.2003-204 (2003) (“Wang”)
Exhibit 1016	Shuhua Gao, Xiaoling Wu, Cheng Xiang, and Dongyan Huang, <i>Development of a Computationally Efficient Voice Conversion System on Mobile Phones</i> , https://doi.org/10.1017/ATSIP.2018.23 (2018) (“Wu”)
Exhibit 1017	U.S. Patent Application Publication No. 2020/0066296 to Sargsyan et al. (“Sargsyan”)
Exhibit 1018	Su-Hyun Han , Ko Woon Kim, SangYun Kim , Young Chul Youn, <i>Artificial Neural Network: Understanding the Basic Concepts</i>

	<i>without Mathematics</i> , Dement Neurocognitive Disord. 2018 Sep;17(3):83-89 (“ <i>Han</i> ”)
Exhibit 1019	Maurya Vijayaramachandran, and Siddique Afraaz N, <i>Impact of Hidden Layer in Artificial Neural Networks</i> , IOSR Journal of Engineering, Vol. 10, Issue 11, November 2020, Series -I, 33-38 (“ <i>Siddique</i> ”)
Exhibit 1020	Maad M. Mijwel, Adam Esen and Aysar Shamil, <i>Overview of Neural Networks</i> , Babylonian Journal of Machine Learning, 1:2, April 2019 (“ <i>Mijwel</i> ”)
Exhibit 1021	Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, <i>Generative Adversarial Networks: An Overview</i> , 1710.07035v1 [cs.CV] 19 Oct 2017 (“ <i>Creswell</i> ”)
Exhibit 1022	Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou, <i>Feature Dimensionality Reduction: A Review</i> , Complex & Intelligent Systems (2022) 8:2663–2693 (“ <i>Jia</i> ”)
Exhibit 1023	Hyun Ah Song and Soo-Young Lee, <i>Hierarchical Data Representation Model - Multi-layer NMF</i> , arXiv:1301.6316v3 [cs.LG] 18 Mar 2013 (“ <i>Song</i> ”)
Exhibit 1024	Premananda B S and Dr. Uma B V, <i>Speech Enhancement Algorithm to Reduce the Effect of Background Noise in Mobile Phones</i> , International Journal of Wireless & Mobile Networks (IJWMN) Vol. 5, No. 1, February 2013 (“ <i>Premananda</i> ”)
Exhibit 1025	Yi Xu, <i>Prosody, Tone, and Intonation</i> , Routledge Handbook of Phonetics. W. F. Katz and P. F. Assmann: Routledge, New York. pp. 314-356 (2019) (“ <i>Xu</i> ”)
Exhibit 1026	PCT Publication No. 2022/168102 to Keshet et al. (“ <i>Keshet</i> ”)
Exhibit 1027	Tal Peer and Timo Gerkmann, <i>Phase-Aware Deep Speech Enhancement: It’s All About The Frame Length</i> , arXiv:2203.16222v2 [eess.AS] 4 Oct 2022 (“ <i>Peer</i> ”)
Exhibit 1028	U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“ <i>Prabhavalkar</i> ”)
Exhibit 1029	Qifeng Zhu and Abeer Alwan, <i>On the Use of Variable Frame Rate Analysis in Speech Recognition</i> , 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 3, 1783-1786 (2000) (“ <i>Zhu</i> ”)
Exhibit 1030	Qi Li, <i>An Auditory-Based Transform For Audio Signal Processing</i> , 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (“ <i>Li</i> ”)

Exhibit 1031	U.S. Patent No. 7,328,153 to Wells et al. (“ <i>Wells</i> ”)
Exhibit 1032	U.S. Patent No. 10,796,686 to Arik et al. (“ <i>Arik</i> ”)
Exhibit 1033	U.S. Patent Application Publication No. 2003/0088408 to Thyssen et al. (“ <i>Thyssen</i> ”)
Exhibit 1034	U.S. Patent Application Publication No. 2005/0066209 to Kee et al. (“ <i>Kee</i> ”)
Exhibit 1035	Abdullah Zaini Alsheibi, Unsupervised Learning Algorithm for Noise Suppression and Speech Enhancement Applications, Electronic Theses and Dissertations. 2168. https://digitalcommons.du.edu/etd/2168 (2023) (“ <i>Alsheibi</i> ”)
Exhibit 1036	U.S. Patent Application Publication No. 2015/0371655 to Gao (“ <i>Gao</i> ”)
Exhibit 1037	U.S. Patent No. 10,561,361 to Quatieri et al. (“ <i>Quatieri</i> ”)
Exhibit 1038	U.S. Patent No. 9,195,649 to Neuhauser et al. (“ <i>Neuhauser</i> ”)
Exhibit 1039	European Patent Application Publication No. 0756172 to Demmin et al. (“ <i>Demmin</i> ”)
Exhibit 1040	U.S. Patent No. 6,269,351 to Black (“ <i>Black</i> ”)
Exhibit 1041	U.S. Patent Application Publication No. 2017/0193066 to Zhu et al. (“ <i>Guo</i> ”)
Exhibit 1042	Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai, <i>Real-Time Neural Text-To-Speech with Sequence-To-Sequence Acoustic Model and WaveGlow or Single Gaussian WaveRNN Vocoders</i> , Interspeech 2019 (“ <i>Okamoto</i> ”)
Exhibit 1043	Haohe Liu et al., <i>VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration</i> , Interspeech 2022 (“ <i>Liu</i> ”)
Exhibit 1044	Leyuan Sheng et al., <i>Reducing sover-smoothness in speech synthesis using Generative Adversarial Networks</i> , IEEE 2018 (“ <i>Sheng</i> ”)
Exhibit 1045	Declaration of Mina Ching, Records Request Processor at the Internet Archive (“ <i>Ming</i> ”)
Exhibit 1046	Francois Waldner et al., <i>Deep learning on edge: extracting field boundaries from satellite images with a convolutional neural network</i> , February 4, 2020 (“ <i>Waldner</i> ”)
Exhibit 1047	Patrik O. Hoyer, <i>Non-Negative Sparse Coding</i> , Neural Networks Research Centre Helsinki University of Technology (“ <i>Hoyer</i> ”)
Exhibit 1048	Haohe Liu et al., <i>VoiceFixer: Toward General Speech Restoration with Neural Vocoder</i> , October 5, 2021

14. I have considered these materials from the viewpoint of a PHOSITA as of the priority date of the '496 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the '496 Patent is May 5, 2023. I note that my opinions provided in this Declaration are made from the perspective of a PHOSITA as of this priority date of the '496 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 '496 Patent’s priority date (again, unless expressly stated otherwise). I merely use the present verb tense for ease of reading.

III. OVERVIEW AND LEGAL STANDARDS

15. In formulating my opinions, I have been instructed to apply certain legal standards. I am not a lawyer. I do not expect to offer any testimony regarding what the law is. Instead, the following sections summarize the law as I have been instructed to apply it in formulating and rendering my opinions found later in this declaration. I understand that, in an *inter partes* review (“IPR”) proceeding, patent claims may be deemed unpatentable if it is shown that they were anticipated or rendered obvious in view of the prior art. I understand that prior art in an IPR review is limited to patents or printed publications that predate the priority date of the patent at issue. I understand that questions of claim clarity (definiteness) and enablement

cannot be considered as a ground for considering the patentability of a claim in these proceedings.

A. Person of Ordinary Skill in the Art

16. I understand that the '496 Patent, the record of proceedings at the Patent Office (which I understand is called the "File History" or "Prosecution History"), and the teachings of the prior art are evaluated from the perspective of a person of ordinary skill in the art ("PHOSITA"). I understand that the factors considered in determining the ordinary level of skill in the art include: (i) the levels of education and experience of persons working in the field; (ii) the types of problems encountered in the field; and (iii) the sophistication of the technology. I may also consider, if available, the education level of the inventor, prior art solutions to the problems encountered in the art, and the rapidity with which innovations are made in the relevant art.

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

B. Obviousness

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

19. I also understand that an analysis of whether a claimed invention would have been obvious should be considered in light of the scope and content of the prior art, the differences (if any) between the prior art and the claimed invention, and the level of ordinary skill in the pertinent art involved. I understand as well that a prior art reference should be viewed as a whole. I understand that in considering whether an invention for a claimed combination would have been obvious, I may assess whether there are apparent reasons to combine known elements in the prior art in the manner claimed in view of interrelated teachings of multiple prior art references, the effects of demands known to the design community or present in the marketplace,

and/or the background knowledge possessed by a PHOSITA. I also understand that other principles may be relied on in evaluating whether a claimed invention would have been obvious, and that these principles include the following:

- A combination of familiar elements according to known methods is likely to be obvious when it does no more than yield predictable results;
- When a device or technology is available in one field of endeavor, design incentives and other market forces can prompt variations of it, either in the same field or in a different one, so that if a PHOSITA can implement a predictable variation, the variation is likely obvious;
- If a technique has been used to improve one device, and a PHOSITA would have recognized that it would improve similar devices in the same way, using the technique is obvious unless its actual application is beyond his or her skill;
- An explicit or implicit teaching, suggestion, or motivation to combine two prior art references to form the claimed combination may demonstrate obviousness, but proof of obviousness does not depend on or require showing a teaching, suggestion, or motivation to combine;
- Market demand, rather than scientific literature, can drive design trends and may show obviousness;

- In determining whether the subject matter of a patent claim would have been obvious, neither the particular motivation nor the avowed purpose of the named inventor controls;
- One of the ways in which a patent's subject can be proved obvious is by noting that there existed at the time of invention a known problem for which there was an obvious solution encompassed by the patent's claims;
- Any need or problem known in the field of endeavor at the time of invention and addressed by the patent can provide a reason for combining the elements in the manner claimed;
- "Common sense" teaches that familiar items may have obvious uses beyond their primary purposes, and in many cases a PHOSITA will be able to fit the teachings of multiple patents together like pieces of a puzzle;
- A PHOSITA is also a person of ordinary creativity, and is not an automaton;
- A patent claim can be proved obvious by showing that the claimed combination of elements was "obvious to try," particularly when there is a design need or market pressure to solve a problem and there are a finite number of identified, predictable solutions such that a PHOSITA

would have had good reason to pursue the known options within his or her technical grasp; and

- One should not use hindsight in evaluating whether a claimed invention would have been obvious.

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

21. I am informed that it is improper to combine references where the references teach away from their combination. I am informed that a reference may be said to teach away when a person of ordinary skill in the relevant art, upon reading the reference, would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path that was taken by the patent applicant. In general, a reference will teach away if it suggests that the line of development flowing from the reference's disclosure is unlikely to be productive of the result sought by the patentee. I am informed that a reference teaches away, for example, if (1) the combination would produce a seemingly inoperative device, or (2) the references leave the impression that the product would not have the property

sought by the patentee. I also am informed, however, that a reference does not teach away if it merely expresses a general preference for an alternative invention but does not criticize, discredit, or otherwise discourage investigation into the invention claimed.

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

23. I am further informed that secondary considerations evidence is only relevant if the offering party establishes a connection, or nexus, between the evidence and the claimed invention. The nexus cannot be based on prior art features. The establishment of a nexus is a question of fact. While I understand that the Patent

Owner here has not offered any secondary considerations at this time, I will supplement my opinions in the event that the Patent Owner raises secondary considerations during the course of this proceeding.

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

C. Analogous Art

25. I have been informed that for a prior art reference to be proper for use in an obviousness analysis, the reference must be “analogous art” to the claimed invention. I have been informed that a reference is analogous art to the claimed invention if: (1) the reference is from the same field of endeavor as the claimed invention (even if it addresses a different problem); or (2) the reference is reasonably pertinent to the problem faced by the inventor (even if it is not in the same field of endeavor as the claimed invention). In order for a reference to be “reasonably pertinent” to the problem, it must logically have commended itself to an inventor’s attention in considering his problem. In determining whether a reference is reasonably pertinent, one should consider the problem faced by the inventor, as reflected either explicitly or implicitly, in the specification. I believe that all of the references that my opinions in this IPR are based upon are well within the range of references a person having ordinary skill in the art would consult to address the type of problems described in the Challenged Claims.

D. Claim Construction

26. I understand that the claim terms in this IPR will be construed according to their ordinary and customary meaning as understood in light of the claim language, the patent's description, and the prosecution history viewed from the perspective of a PHOSITA. I further understand that claims may be construed based on the specification.

27. In this proceeding, I have been instructed to construe the term "low-dimensional representation" as "a compressed representation of an input speech frame which is a result of a dimensionality reduction technique" in Grounds 1-3. And I have been instructed to apply the express claim language "low-dimensional representation" without such construction in Grounds 4-6.

IV. LEVEL OF A PERSON OF ORDINARY SKILL

28. Based on my review and analysis of the '496 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a PHOSITA in the field of the '496 Patent at the time of the earliest possible priority date (May 5, 2023) would have been knowledgeable regarding the field of audio speech signal processing. In my experience working in this field, most workers of ordinary skill in the art as of the earliest possible priority date of May 5, 2023, would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of speech

processing and machine learning. *See, e.g.*, the '496 Patent (Ex. 1001) at 1:17-2:15 (describing the “Background”). A person with less relevant practical experience but with additional education can also qualify as a PHOSITA in the field of the '496 Patent provided the additional education focused on speech processing technology. When I refer to the understanding of a PHOSITA, I am referring to the understanding of such a person as of May 5, 2023.

V. OVERVIEW OF TECHNOLOGY

29. I was asked to briefly summarize the background of the technology from the standpoint of a PHOSITA prior to May 5, 202.

A. Neural Networks

30. An artificial neural network is a machine learning algorithm based on the concept of a human neuron. EX1018-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. EX1019-Siddique, 1. Typical applications of neural networks include voice recognition and speech synthesis. EX1020-Mijwel, 3. A neural network consists of neurons (also called nodes) that “pick up information from the outside or from other neurons and pass it on to other neurons or output it as a final result.” EX1020-Mijwel, 2.

31. The structure of a neural network contains input, hidden and output neurons. EX1020-Mijwel, 2. Input neurons receive outside data, hidden neurons map

internal patterns, and output neurons relay information to the world. *Id.* Weights control the influence one neuron exerts over another neuron, and during training weightings of the connections change. *Id.* Thus, “[n]eural Networks are capable of learning and can be trained to produce desired results.” EX1019-Siddique, 2.

32. There are many types of different neural networks. One pertinent to the disclosure is a Generative Adversarial Network (GAN) model. A GAN model is adversarial and involves two submodels, a generator and a discriminator. EX1021-Creswell, 1 rhc. The generator learns to create fake data that resembles the original domain data. *Id.* The discriminator attempts to distinguish between the fake and real data. *Id.* The generator improves until the discriminator cannot easily distinguish the fake and real data. EX1021-Creswell, 6 lhc.

B. Feature Extraction and Dimensionality Reduction Techniques

33. Feature extraction generates features from the original data. EX1022-Jia, 2 rhc. Feature extraction helps separate effective information from redundant data. *Id.* Dimensionality reduction techniques can compress features and representations, which include principal component analysis (PCA) and linear discriminant analysis (LDA). EX1022-Jia, 7-9.

1. Principal Component Analysis

34. PCA “allows the original complex variable to be represented by several integrated factors that reflect the information contained in the original variable as

much as possible, and these factors do not relate to each other, to achieve the purpose of simplification” EX1022-Jia, 8 lhc. The “high-dimensional observation data is transformed into sub-spaces with lower dimensions through projection.” *Id.* The “principal components are required to reflect as much as possible the information contained in the original data[.]” *Id.* PCA addresses the “curse of dimensionality” where significant “computing time and storage space are spent on the processing of the data.” EX1022-Jia, 1 lhc. Further, “this problem also has a bad influence on the accuracy of the recognition.” EX1022-Jia, 1. Thus dimensionality reduction techniques such as PCA “obtain useful reduction data from the high-dimensional data set to meet the recognition accuracy and storage requirements under the premise of maintaining the essential characteristics of the original data optimally.” EX1022-Jia, 1 rhc. Specifically, “PCA allows the original complex variable to be represented by several integrated factors that reflect the information contained in the original variable as much as possible, and these factors do not relate to each other, to achieve the purpose of simplification[.]” EX1022-Jia, 8 lhc.

2. Hierarchical Feature Extraction

35. Hierarchical feature extraction is a strategy that has been in use since at least 2006. EX1023-Song, 1. It is recognized that artificial neural networks can be used to process data features hierarchically “by several transformation stages, finally, the feature representation of the sample in the original space is transformed

into a new feature space. Then combining low-level features to form more abstract high-level representations and attribute categories or features, hierarchical feature representation is obtained, which is more conducive to the classification or visualization of features” EX1022-Jia, 21-22. For example, *Song* proposes a “hierarchical multi-layer NMF structure comprise of several layers of unit algorithm” where each layer is trained separately.” EX1023-Song, 2. *Song*’s network “Proposed network is constructed by stacking nsNMF [non-smooth non-negative matrix factorization] into several layers.” EX1023-Song, 1. Hierarchical feature extraction “reveals intuitive feature hierarchies (subcategories) by learning feature relationships throughout the layers, and [] learns more meaningful features compared to one-step learning.” EX1023-Song, 3. For example *Song*’s multi-layer NMF model “results in much better classification and reconstruction performance, provided small number of dimensions for data representation.” EX1023-Song, 3.

C. Digital Noise Removal

36. Techniques for removing unwanted noise from audio have long been known in the art for various purposes. For example, a user would want to remove background noise in a phone call because “[a] high background acoustic noise level is annoying to the listener side.” EX1024-Premananda, 1. Further, “[l]istening to speech or audio signals becomes more difficult as the background noise level dominates.” EX1024-Premananda, 1. Additionally, in other applications,

“[background] noise may prevent speech recognition software from correctly identifying the speech audio.” EX1017-Sargsyan, [0003]. “Hence, there is a strong need to improve the quality of the speech signal in noisy conditions by developing speech enhancement algorithms to minimize the effect of background noise.” EX1024-Premananda, 1. As a result, a variety of methods and technologies have been developed to reduce background noise from speech, including through the use of machine learning neural networks.

37. For example, *Sargsyan* teaches “speech enhancement and noise suppression using a deep neural network.” EX1017-Sargsyan, [0002]. The neural network receives frames of audio data in the frequency domain and outputs a ratio mask. EX1017-Sargsyan, [0004]. The ratio mask is used to compute clean voice coefficients that allow the system to output an audio file “having enhanced speech and suppressed noise[.]” EX1017-Sargsyan, [0004].

38. It was known that natural speech contains non-phonetic voice characteristics or components that communicate information to a listener. EX1025-Xu, 1. Examples of these characteristics include pitch, formant structure, prosody, intonation, rate, stress, articulation, etc. EX1005-Strake, [031]; EX1011-Nguyen, [0073]; EX1025-Xu, 1. PHOSITAs understood that replicating natural speech is important for speech modification systems as it is easier for a listener to understand and pay attention to. *See* E1010-Schmandt, 119-123 (explaining the benefits of

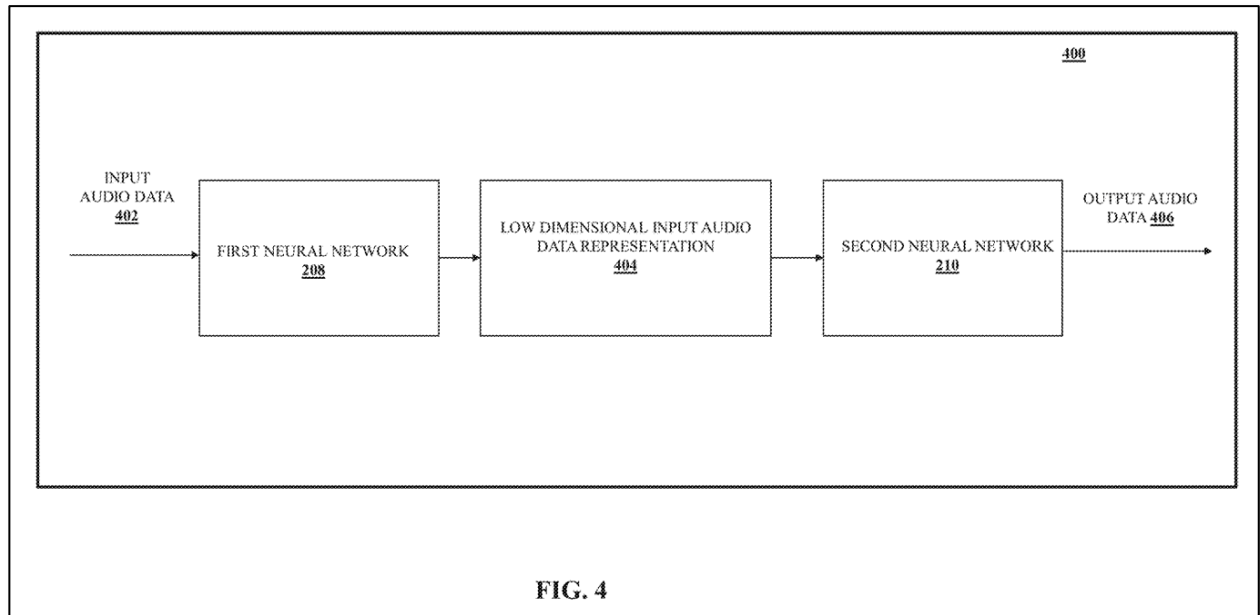
natural-sounding speech). Thus, speech processing systems were known to retain some or all of these characteristics in modified or synthesized speech. *See* EX1005-Strake, [031]; *see also* EX1011-Nguyen [0073]; *see also* EX1026-Keshet, [0009], [0025], [0057].

VI. OVERVIEW OF THE '496 PATENT

39. I have reviewed the '496 Patent (EX1001). Descriptions of the '496 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the '496 Patent are intended only as a summary of the '496 Patent and/or intended as being pertinent to the Challenged Claims of the '496 Patent. My opinions provided herein are not intended as a discussion of my entire understanding of the '496 Patent or any other claims in any related patent to the '496 Patent.

A. Description of the '496 Patent

40. The '496 Patent is generally related to “audio analysis and, more particularly to methods and systems for voice enhancement using neural networks.” EX1001, 1:11-13. To accomplish the voice enhancement, the patent proposes a voice enhancement system that consists of two neural networks. EX1001, Abstract, Fig. 4.



EX1001. Fig. 4.

41. A first neural network of the system “converts the input audio data 402 to a low-dimensional input audio data representation 404.” EX1001, 5:52-55. Then, the system applies a “second neural network 210 to the low-dimensional input audio data representation 404.” EX1001, 5:56-58. The second neural network generates output audio data which is then be converted to analog before being output as output audio. EX1001, 5:58-60.

B. Field of Endeavor of the '496 Patent

42. I have been informed that the field of endeavor of the claimed invention can be determined by reference to explanations of the invention’s subject matter in

the patent application, including the embodiments, function, and structure of the claimed invention.

43. A PHOSITA would have understood that the '496 Patent is directed to audio/speech processing. For example, the '496 Patent is titled "Methods for Neural Network-Based Voice Enhancement And Systems Thereof" and states that the "technology is generally related to audio analysis and, more particularly, to methods and systems for voice enhancement using neural networks." EX1001, Title, 1:11-13. After reviewing the '496 Patent, it is my opinion that a PHOSITA would understand that the field of endeavor of the '496 Patent is speech audio processing. *See* EX1001, Abstract, 1:11-13, Fig. 2.

C. Problem Purportedly Solved by the Inventors of the '496 Patent

44. I have been informed that a prior art reference is "reasonably pertinent" if a PHOSITA would have consulted it and applied its teachings when faced with the problem that the inventor was trying to solve. As such, I have been asked to analyze the '496 Patent and determine the problems that the inventors were trying to solve.

45. In my opinion, there are multiple problems the inventors of the '496 Patent were involved with solving.

46. For example, the inventors addressed problems related to how neural networks are applied in noise removal/voice enhancement systems. EX1001,

Abstract, 2:42-45, 4:38-41, Fig. 4, Claims 1,11, 16. The '496 Patent discloses “systems for neural network-based voice enhancement and noise suppression.” EX1001, 2:43-45.

47. Additionally, the inventors addressed problems with how to remove noise from audio signals, particularly via the application of certain pipelined processes, and how data should be broken up and segmented for such pipelined processes. EX1001, 1:16-26, 1:54-64, 2:9-15, 2:42-45, 4:38-48, Claims 3, 5, 13, 14, 17. For example, the first neural network “receives input audio data, fragments the input audio data into frames, and converts the frames to low-dimensional representations[.]” EX1001, 4:41-44. “The low-dimensional input audio data representation 404 may omit any number of the non-content elements of the audio data received in step 502 (e.g., background noise, and other elements such as microphone pops, low-fidelity audio, and audio clippings).” EX1001, 7:2-6.

48. The inventors also addressed problems related to the dimensionality of representations and how such representations should be compressed, for example through dimensionality reduction techniques. EX1001, Abstract, 4:42-46, 7:22-54, Claims 1, 9, 11, 15, 16. The '496 Patent teaches that features can be compressed into low-dimensional features through “techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), or other dimensionality reduction techniques, for example.” EX1001, 7:25-27. In embodiments where a

hierarchical feature extraction network is used, “[a]t each level of the hierarchical feature extraction network, the extracted features could be compressed into a low-dimensional input audio data representation 404 using a compression algorithm such as principal component analysis (PCA) or non-negative matrix factorization (NMF), for example.” EX1001, 7:39-44.

49. The inventors also addressed problems related to how to train neural networks in noise removal systems. EX1001, 6:8-57, 7:57-8:67, Figs. 6, 7, Claims 4, 6, 11, 12, 18, 19. For example, “the first neural network 208 may be optimized by the voice enhancement system 100 to learn a mapping between the input training speech frames and the low-dimensional input audio data training data representation 632, using techniques such as supervised learning or reinforcement learning, for example.” EX1001, 6:48-53. Further, “the second neural network 210 may be initially trained using supervised learning to convert the low-dimensional representation of input audio training data speech frames 634(1)-634(n) in real-time. The second neural network 210 may be trained to learn the conversion between the low-dimensional representation of input audio training data speech frames 634(1)-634(n) and the target training speech frames 712(1)-712(n) using a loss function that minimizes the difference between the predicted and actual target speech frames, for example.” EX1001, 8:26-35.

50. The inventors further addressed problems related to feature extraction, particularly when processing speech in noisy environments. EX1001, 1:19-26, 1:54-64, 2:9-15, 7:16-22, 7:32-39, Claims 9, 10, 15. The '496 Patent teaches that “features may be extracted by the voice enhancement system 100 such as by using Fourier Transform, Mel-Frequency Cepstral Coefficients (MFCC), or other techniques.” EX1001, 7:17-19. Further, the system may use “a hierarchical feature extraction network that extracts multiple levels of features from the input audio data 402.” EX1001, 7:34-35.

D. Summary of the Prosecution History

51. The Application that resulted in the '496 Patent was filed on April 24, 2024, and purports to claim priority to provisional patent application No. 63/464,432, filed on May 5, 2023. EX1001, (22), (60). For the purposes of this proceeding, I have been asked to apply a May 5, 2023 priority date for the Challenged Claims.

52. The '496 Patent received no rejections throughout prosecution, receiving a Notice of Allowance on July 5, 2024. EX1002, 125. The Examiner stated the reason for allowance is that the prior art fails to teach, disclose, or suggest the claimed limitations to “apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames.” EX1002, 130.

VII. OVERVIEW OF THE PRIOR ART

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

54. The prior art is analogous to the claimed invention of the '496 Patent, as discussed with greater detail for each prior art reference included. In general, the prior art is from the same field of endeavor of the '496 Patent, namely speech audio processing. '496 Patent, Abstract, 1:11-13, 2:42-45. Moreover, a PHOSITA would have found the prior art to be of the same field of endeavor and thus analogous to the '496 Patent.

55. The prior art is also reasonably pertinent to at least one problem facing the '496 Patent inventors, namely at least one of:

(A) Problems related to how neural networks are applied in noise removal/voice enhancement systems;

(B) Problems with how to remove noise from audio signals, particularly via the application of certain pipelined processes, and how data should be broken up and segmented for such pipelined processes;

(C) Problems related to the dimensionality of representations and how such representations should be compressed, for example through dimensionality reduction techniques;

(D) Problems related to how to train neural networks in noise removal systems;

(E) Problems related to feature extraction, particularly when processing speech in noisy environments;

As described further below with respect to each denoted prior art reference.

A. Overview of *Dantrey*

56. U.S. Patent No. 12,412,590 (“*Dantrey*”) was filed on December 19, 2019 and issued as a U.S. Patent on September 9, 2025. *Dantrey* was not cited or discussed during prosecution of the ’496 Patent. *See generally* EX1001, EX1002.

57. *Dantrey* teaches apparatuses, systems, and techniques for reducing noise in audio, and specifically a two neural network process to do such. EX1004-*Dantrey*, Abstract, Title. Fig. 1 illustrates a system that processes audio to remove noise. EX1004-*Dantrey*, 2:64-66.

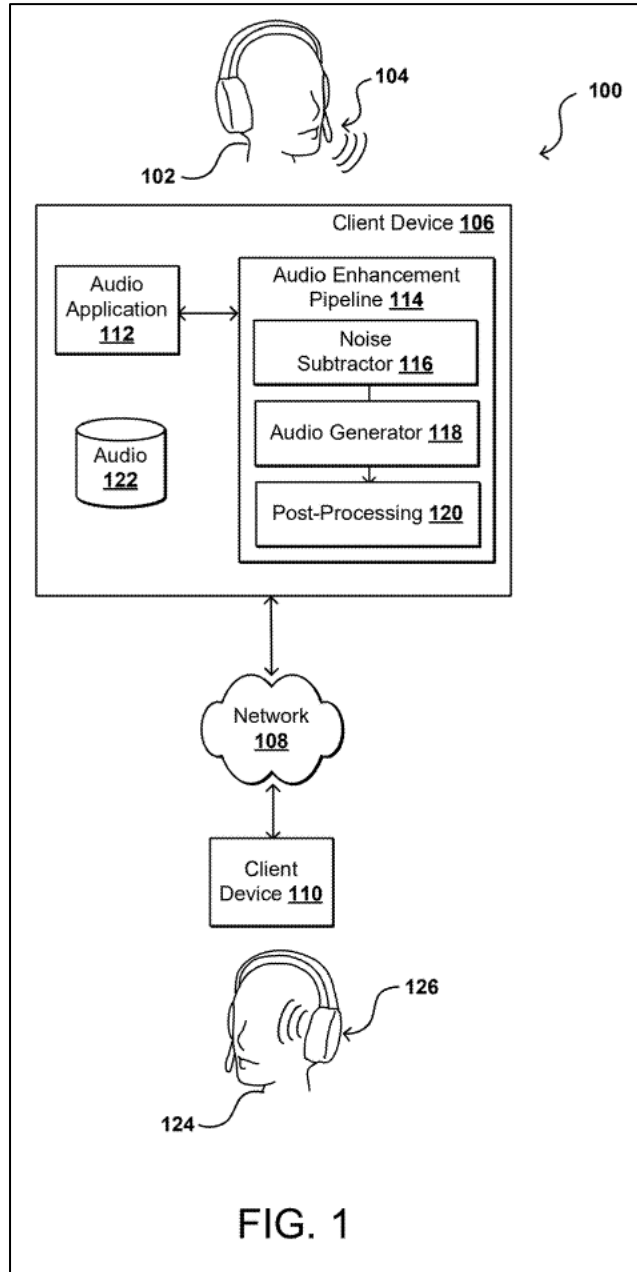


FIG. 1

EX1004-Dantrey, Fig. 1. *Dantrey* teaches that a digital audio signal is input into its pipeline (where the first component in the pipeline is a noise subtractor neural network). EX1004-Dantrey, 3:57-62, Claims 7, 12, Fig. 1. The noise subtractor's output is input to the second neural network, an audio generator 118, which generates

clean speech audio. EX1004-Dantrey, 5:4-10. Both the noise subtractor and the audio generator are neural networks. EX1004-Dantrey, 4:45-47, 5:7-13.

58. A PHOSITA would have identified *Dantrey* as being within the same field of endeavor as the '496 Patent because both *Dantrey* and the '496 Patent relate to speech audio processing. *Compare* EX1004-Dantrey, 4:45-53, 5:4-10; *with* '496 Patent, Abstract, 1:11-13, 2:42-45. *Dantrey* discloses a system where audio data [] can be processed to remove noise[.]” EX1004-Dantrey, 2:64-65. Specifically, *Dantrey* teaches transforming input audio to an audio spectrogram, which “can be provided as input to a noise subtractor 116, which can reduce a presence of background noise.” EX1004-Dantrey, 3:60-62. The “audio spectrogram with reduced background noise can be passed as input to an audio generator 118, which can reduce a presence of foreground noise.” EX1004-Dantrey, 3:62-65.

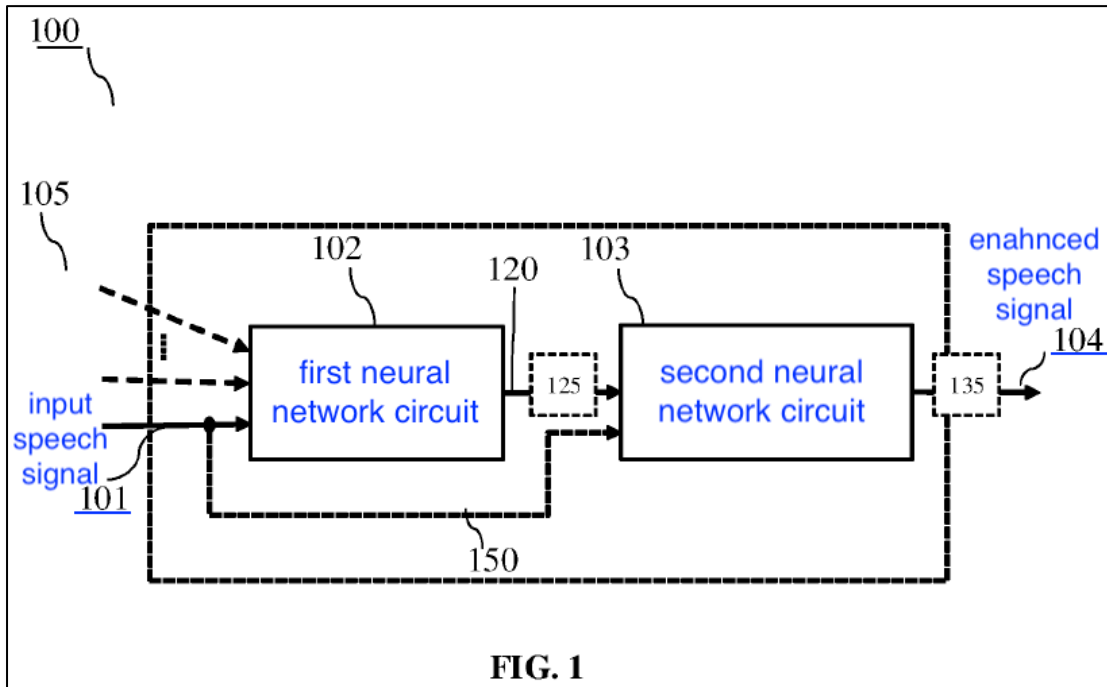
59. A PHOSITA would have found *Dantrey* reasonably pertinent to the problem faced by the inventors of the '496 Patent because both *Dantrey* and the '496 Patent describe an improvement upon existing solutions, that is, applying neural networks to noise removal. *Compare* EX1004-Dantrey, Abstract, 2:64-66, 4:22-27; *with* EX1001, 2:42-45, 2:51-56, 4:38-41. For example, *Dantrey* describes “removing background noise using a first neural network that is able to extract background noise having a frequency range and pattern that is differentiable from that of this primary audio.” EX1004-Dantrey, 4:24-27. “[A] second neural network is able to

differentiate between audio sources and generate an audio signal that predominantly corresponds to this primary audio. EX1004-Dantrey, 4:30-32. Further, *Dantrey* and the '496 Patent each account for problems with how to remove noise from audio signals, particularly via the application of certain pipelined processes, and how data should be broken up and segmented for such pipelined processes. *Compare* EX1004-Dantrey, 3:54-62, 4:20-40; *with* EX1001, 1:16-26, 1:54-64, 2:9-15, 2:42-45, 4:38-48.

B. Overview of Strake

60. International Application Publication No. 2020/199990 (“*Strake*”) was filed on March 24, 2020, and published on October 8, 2020. EX1005-Strake, (22), (43). *Strake* was not cited or considered during the prosecution of the '496 Patent. *See generally* EX1001, 1002.

61. *Strake* teaches a “speech processing system” that employs “neural network topologies...to obtain a strong acoustic interference suppression.” EX1005-Strake, [019].



EX1005-Strake, Fig. 1¹.

62. Specifically, *Strake* teaches “a two-stage neural network-based speech enhancement circuit” that “uses neural network models in both processing stages,”

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

where “the first neural network...can be trained to suppress any acoustic background interference,” and “the second neural network is configured to restore natural sounding speech.” EX1005-Strake, [029]-[030].

63. *Strake* further specifically teaches the framing functionality for its processing stages/pipeline, including breaking up frames at the front end, processing frames throughout, and finally combining such resulting frames via well-known overlap add (OLA) type techniques. EX1005-Strake, [037], [042].

64. A PHOSITA would have identified *Strake* as being within the same field of endeavor as the '496 Patent because both *Strake* and the '496 Patent relate to speech audio processing. *Compare* EX1005-Strake, Title, Abstract, [02], [019]; *with* '496 Patent, Abstract, 1:11-13, 2:42-45. *Strake* discloses a “speech processing system and method therefor.” EX1005-Strake, [02]. Specifically, *Strake* teaches “a neural-network-based approach that uses two distinct processing stages, namely a first neural network circuit (or stage) that is configured for acoustic interference suppression and a second neural network circuit (or stage) that is configured for restoration of a natural sounding speech signal.” EX1005-Strake, [019].

65. A PHOSITA would have found *Strake* reasonably pertinent to the problem faced by the inventors of the '496 Patent because both *Strake* and the '496 Patent describe an improvement upon existing solutions, that is, how to remove noise from audio signals, particularly via the application of certain pipelined processes,

and how data should be broken up and segmented for such pipelined processes. Compare EX1005-Strake, [037]-[042]; with EX1001, 1:16-26, 1:54-64, 2:9-15, 2:42-45, 4:38-48. For example, *Strake* teaches framing audio data into frames of 32ms. EX1005-Strake, [037]. A time domain second stage estimate of the clean signal is created by “applying an inverse transform from the processing domain back to the time domain, together with a subsequent combination of the time domain frames, e.g. by Inverse Fast Fourier Transform (IFFT) and an overlap add (OLA) operation, in one example embodiment.” EX1005-Strake, [042].

C. Overview of Chen

66. U.S. Patent No. 10,867,616 (“*Chen*”) was filed on October 10, 2019, and issued December 15, 2020. *Chen* was not cited or considered during prosecution of the ’496 Patent. See generally EX1001, EX1002.

67. *Chen* teaches “solutions for eliminating undesired audio artifacts, such as background noises, on an audio channel.” EX1006-*Chen*, Abstract. *Chen* also teaches the use of dimensionality reduction approaches by machine learning models. EX1006-*Chen*, 3:66-4:4.

68. A PHOSITA would have identified *Chen* as being within the same field of endeavor as the ’496 Patent because both *Chen* and the ’496 Patent relate to speech audio processing. Compare EX1006-*Chen*, Abstract, 1:13-17; with ’496 Patent, Abstract, 1:11-13, 2:42-45. *Chen* discloses “systems and methods for

reducing background noise and in particular, for deploying machine learning models to detect and attenuate unwanted background noises (audio artifacts) in teleconference and videoconference settings.” EX1006-Chen, 1:13-17. Specifically, *Chen* teaches multi-layered approach where “sounds having a low-probability of being background noises can be quickly filtered using a light-weight preliminary (first) ML model” while “higher-probability background events can be efficiently screened using a subsequent (second) ML model that is more accurate and robust than the first ML model.” EX1006-Chen, 3:7-18.

69. A PHOSITA would have found *Chen* reasonably pertinent to the problem faced by the inventors of the '496 Patent because both *Chen* and the '496 Patent describe an improvement upon existing solutions, that is, how to remove noise from audio signals, particularly via the application of certain pipelined processes, and how data should be broken up and segmented for such pipelined processes. *Compare* EX1006-Chen, Abstract, 2:44-57, 3:7-18, Fig. 3; *with* EX1001, 1:16-26, 1:54-64, 2:9-15, 2:42-45, 4:38-48. For example, *Chen* teaches that “[a]udio segments can be sampled from audio information passing over a communications channel, for example, as between two or more of [the] devices[.]” EX1006-Chen, 5:55-58. “After the real-time audio segments are generated, the segments are provided to a first ML model (204).” EX1006-Chen, 5:64-65. If the first and second neural network determine that there is a high probability that a background

feature is unwanted noise, “the background event may be reduced in volume (dB) using an attenuation module[.]” EX1006-Chen, 6:56-57; *see* EX1006-Chen, 6:21-64.

70. Additionally, a PHOSITA would have found *Chen* reasonably pertinent to problems related to the dimensionality of representations and how such representations should be compressed, for example through dimensionality reduction techniques. *Compare* EX1006-Chen, 3:66-4:4; *with* EX1001, Abstract, 4:42-46, 7:22-54. Particularly, *Chen* teaches a machine learning model employing Principal Component Analysis (PCA) as a dimensionality reduction technique. EX1006-Chen, 3:66-4:4.

D. Overview of Quillen

71. U.S. Patent Application Publication No. 2021/0241780 (“*Quillen*”) was filed on October 31, 2020, and published on August 5, 2021. *Quillen* was not cited or considered during prosecution of the ’496 Patent. *See generally* EX1001, EX1002.

72. *Quillen* teaches “[e]mbodiments improve speech data quality through training a neural network for de-noising audio enhancement” and *Quillen* specifically provides detailed training disclosures for how to train its neural networks. EX1007-*Quillen*, Abstract.

73. A PHOSITA would have identified *Quillen* as being within the same field of endeavor as the '496 Patent because both *Quillen* and the '496 Patent relate to speech audio processing. *Compare* EX1007-*Quillen*, Abstract, [0001]-[0002], [0016], [0018]; *with* '496 Patent, Abstract, 1:11-13, 2:42-45. *Quillen* discloses “techniques for speech enhancement through the training and use of a neural network.” EX1007-*Quillen*, [0016]. Specifically, *Quillen* teaches training neural networks for de-noising where the “training includes minimizing errors in the neural network according to at least one of (i) a decoding error of an Automatic Speech Recognition (ASR) system processing current de-noised speech data results generated by the neural network during the training and (ii) spectral distance between the high quality speech data and the current de-noised speech data results generated by the neural network during the training.” EX1007-*Quillen*, Abstract.

74. A PHOSITA would have found *Quillen* reasonably pertinent to the problem faced by the inventors of the '496 Patent because both *Quillen* and the '496 Patent describe an improvement upon existing solutions, that is, how to train neural networks in noise removal systems. *Compare* EX1007-*Quillen*, Abstract, [0003]-[0006], [0024]; *with* EX1001, 6:8-57, 7:57-8:67. For example, *Quillen* teaches “[e]mbodiments [that] train a neural network using normalizing flow techniques and employ this trained neural network to enhance audio data.” EX1007-*Quillen*, [0022]. Specifically, “performing the training 332, e.g., deep normalizing flow training,

trains the neural network to determine an invertible one-to-one mapping of high quality (clean) speech to noise[.]” EX1007-Quillen, [0022].

E. Overview of Heckmann

75. “A hierarchical framework for spectro-temporal feature extraction” to Heckmann et al. (“*Heckmann*”) was publicly available at least as early as May 2011. *Heckmann* was not cited or considered during the prosecution of the ’496 Patent. *See generally* EX1001, 1002.

76. Heckmann teaches “a hierarchical framework for the extraction of spectro-temporal acoustic features,” consisting of a “first layer” for extracting local features and a “second layer” for extracting complex features. EX1008-Heckmann, Abstract.

77. A PHOSITA would have identified *Heckmann* as being within the same field of endeavor as the ’496 Patent because both *Heckmann* and the ’496 Patent relate to speech audio processing. *Compare* EX1008-Heckmann, Abstract, Introduction; *with* ’496 Patent, Abstract, 1:11-13, 2:42-45. *Heckmann* discloses “a hierarchical framework for the extraction of spectro-temporal acoustic features[.]” EX1008-Heckmann, Abstract. The features produced are tested in a “continuous digit recognition task in noise.” *Id.*

78. A PHOSITA would have found *Heckmann* reasonably pertinent to the problem faced by the inventors of the ’496 Patent because both *Heckmann* and the

'496 Patent describe an improvement upon existing solutions, that is, problems related to feature extraction, particularly when processing non-clear speech from users. *Compare* EX1008-Heckmann, Abstract, Introduction, 6-9; *with* EX1001, 1:19-26, 1:54-64, 2:9-15, 7:16-22, 7:32-39. For example, *Heckmann* teaches hierarchical feature extraction to extract features where “[t]he design of the features targets higher robustness in dynamic environments.” EX1008-Heckmann, Abstract. Heckmann notes its teachings are “[m]otivated by the large gap between human and machine performance in such conditions[.]” *Id.* Specifically, *Heckmann* teaches extracting local features at a first layer. EX1008-Heckmann, 6-8. Complex features are extracted at a second layer. EX1008-Heckmann, 8-9.

F. Overview of Liu

79. “VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration” to Haohe Liu et al. (“Liu”) was publicly available at least as early as April 13, 2022. EX1043-Liu; EX1045-Ching, 1-2, 11-12, 48-54 (declaration of Mina Ching, Records Request Processor at the Internet Archive). Liu was not cited or discussed during prosecution of the '496 Patent. EX1001; EX1002.

80. Liu teaches a two neural network pipeline for speech restoration/noise removal which expressly uses a “low dimensional mel spectrogram as the intermediate-level feature,” for example as seen by Liu’s Figure 1:

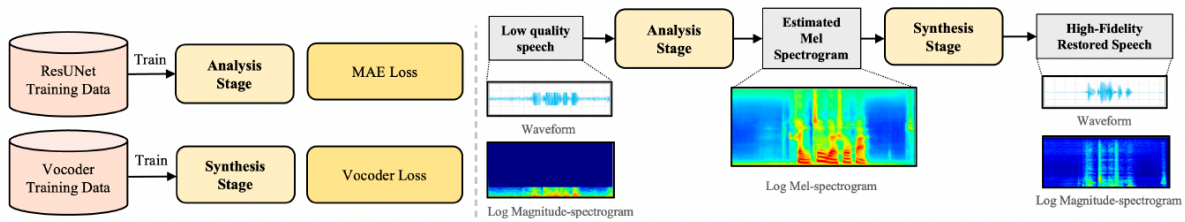


Figure 1: Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.

EX1043-Liu, Fig. 1, Abstract, §§ 1, 3.1.

81. In my opinion, Liu is in the same field of invention as the '496 Patent because both Liu and the '496 Patent relate to speech audio processing. Compare EX1043-Liu, Title, Abstract, Introduction; with '496 Patent, Abstract, 1:11-13, 2:42-45.

82. Additionally, Liu is reasonably pertinent to problems faced by the inventors of the '496 Patent. For example, Liu relates to how neural networks are applied in noise removal system. EX1001, 2:42-45, 2:51-56, 4:38-41, Fig. 4, Claims 1, 11, 16; EX1043-Liu, Title, Abstract, Fig. 1. §§ 1, 3.1, 3.2. Indeed, Liu specifically addresses and contemplates a two neural network pipeline for such. *Id.* Additionally, Liu relates to problems with how to remove noise from audio signals. EX1001, 1:16-26, 1:54-64, 2:9-15, 2:42-45, 4:38-48, 7:2-6, Claims 3, 5, 13, 14, 17. Liu in fact describes how to deal with “multiple distortions simulatneously” in noise removal systems. EX1043-Liu, Abstract, § 1. Further Liu is related to problems related to the dimensionality of representations at least because Liu states “VoiceFixer uses the low dimensional mel spectrogram as the intermediate-level feature, which alleviates

the difficulties of restoring multiple distortions simultaneously.” PatentEX1001, Abstract, 4:42-46, 7:22-54, Claims 1, 9, 11, 15, 16; EX1043-Liu, Abstract, § 1. Further, Liu addresses problems for how to train neural networks because Liu expressly discloses how its training was conducted in addition to making its pretrained models available. EX1001, 6:8-57, 7:57-8:67, Figs. 6, 7, Claims 4, 6, 11, 12, 18, 19; EX1043-Liu, § 4. Liu also addressed problems with feature extraction in noisy environments because Liu again teaches how to remove multiple distortions at the same time. EX1001, 1:19-26, 1:54-64, 2:9-15, 7:16-22, 7:32-39, Claims 9, 10, 15; EX1043-Liu, §§ 1, 3.1-3.2, Fig. 1.

83. Accordingly, and for the reasons explained above, it is my opinion that Liu is analogous art because it is in the same field of endeavor as the '496 Patent as well as reasonable pertinent to at least one problem faced by the inventors of the '496 Patent.

VIII. SUMMARY OF UNPATENTABILITY

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

Proposed Grounds of Unpatentability	Exhibits
Ground 1: Claims 1-3, 7-9, 16-17, and 20 are obvious over Dantrey (EX1004), Strake (EX1005), and Chen (EX1006)	EX1004, EX1005, EX1006

Ground 2: Claims 4-6, 11-15, 18-19 are obvious over Dantrey, Strake, Chen, and Quillen (EX1007)	EX1004, EX1005, EX1006, EX1007
Ground 3: Claim 10 is obvious over Dantrey, Strake, Chen, and Heckmann (EX1008)	EX1004, EX1005, EX1006, EX1008
Ground 4: Claims 1-3, 7-9, 16-17, and 20 are obvious over Dantrey, Strake, and Liu (EX1043)	EX1004, EX1005, EX1043
Ground 5: Claims 4-6, 11-15, 18-19 are obvious over Dantrey, Strake, Liu, and Quillen	EX1004, EX1005, EX1043, EX1007
Ground 6: Claim 10 is obvious over Dantrey, Strake, Liu, and Heckmann	EX1004, EX1005, EX1043, EX1008

IX. OPINIONS REGARDING GROUND I: *DANTREY, STRAKE, AND CHEN*

A. Claim 1

1. Claim 1[pre]: A voice enhancement system, comprising memory having instructions stored thereon and one or more processors coupled to the memory and configured to execute instructions to:

85. In my opinion, *Dantrey* discloses, or at least renders obvious, *a voice enhancement system* (e.g., *Dantrey's* system to reduce noise and enhance received audio), *comprising memory* (e.g., *Dantrey's* memory) *having instructions* (e.g., *Dantrey's* instructions) *stored thereon and one or more processors* (e.g., *Dantrey's*

processor) *coupled to the memory and configured to execute instructions*, as claimed.

86. A PHOSITA would have understood that the claimed “voice enhancement system” purportedly enhances audio data by removing unwanted noise. For example, the ’496 Patent states its “technology advantageously improves speech clarity and intelligibility in various applications by utilizing noise suppression algorithms that more accurately estimate the background noise signal from a single microphone recording, thereby suppressing noise without distorting the target or output enhanced speech data.” EX1001, 2:51-56.

87. *Dantrey* similarly discloses “a system for enhancing audio” by removing unwanted noise. *Dantrey*, Title, Abstract, 1:27-28, Fig. 1.

88. *Dantrey* also discloses the claimed hardware components in the preamble. For example, *Dantrey* discloses its invention uses a “computer system 800” which includes “memory 820” storing “instruction(s)” to be “executed by processor 802.” EX1004-*Dantrey*, 15:3-15, 13:52-14:2, 14:21-28, Fig. 8.

2. Claim 1[a]: fragment input audio data into a plurality of input speech frames, wherein the input audio data comprises foreground speech content, one or more non-content elements, and one or more speech characteristics;

89. In my opinion, *Dantrey* in view of *Strake* renders obvious *fragment[ing] input audio data* (e.g., *Dantrey*’s digital input audio signal) *into a*

plurality of input speech frames (e.g., Strake's frames), wherein the input audio data comprises foreground speech content (e.g., Dantrey's captured human speech), one or more non-content elements (e.g., Dantrey's background noise), and one or more speech characteristics (e.g., Strake's speech components), as claimed.

90. I first explain how *Dantrey* in view of *Strake* teaches how the claimed input audio includes *foreground speech content, one or more non-content elements, and one or more speech characteristics*, along with why a PHOSITA would have been motivated to make this obvious combination. I then explain how *Dantrey* in view of *Strake* teaches how the claimed input audio data is framed, along with why a PHOSITA would have made such obvious combination.

(1) Input audio data comprising foreground speech content, non-content element[s], and speech characteristic[s]

91. Regarding the claimed *input audio data comprising foreground speech content*, in my opinion, *foreground speech content* refers to the prominent content in speech captured from a user. This is clear from the claim language and the specification. For example, the claim language states the *input audio data* comprises *foreground speech content, one or more non-content elements, and one or more speech characteristics*. EX1001, Claim1. *Non-content elements comprise one or more of background noise*, etc. EX1001, Claim 3. And *speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation,*

annunciation, voice identity, or unintelligible speech. EX1001, Claim 3. Accordingly, what remains in captured noisy human speech, *foreground speech content*, is the prominent content of the speech or in other words what is being said.

92. This is because a PHOSITA would have understood that captured human speech (or speech one would hear) contains both what is being said and how it is being said (in addition to noise present in the captured audio). For example, speech characteristics can include pitch, stress, articulation, etc. *See* Section IV.C.

93. *Dantrey* discloses that its system uses a microphone to “*capture speech* or other utterances produced by a first [human] person” and that such captured speech is provided to a computing device to “produce *a digital audio signal*.” EX1004-Dantrey, 3:3-9, 7:45. *Dantrey* further explains that “there may be additional audio or sounds captured by the microphone” which “may be separate from [the] speech of the first person [] and undesirable[,]” and that “this additional audio can include *background noise*[.]” EX1004-Dantrey, 3:19-25.

94. A PHOSITA would have understood that captured speech from a human user includes both prominent content of the speech as well as characteristics of speech such as pitch, articulation, formant structure, etc. For instance, *Nguyen* teaches incoming audio streams include speech from a speaker (i.e., human user) that include speech content and characteristics. EX1011-Nguyen, [0013]. *Nguyen* further teaches characteristics of speech, such as timbre and pitch, and speech

patterns such as pronunciation patterns, cadence, and prosody. EX1011-Nguyen, [0073]. Furthermore, *Xu* teaches that speech is “a system for transmitting communicative meanings with human articulators.” EX1025-Xu, 1. *Xu* teaches that some aspects of speech include intonation, pitch, duration, amplitude, tone, and formant structure. EX1025-Xu, 1, 26-27. Accordingly, *Dantrey’s* captured noisy speech from a human user teaches that *the input audio data comprises foreground speech content, one or more non-content, and one or more speech characteristics*. A PHOSITA would have understood that for a process in which speech is heard or recorded, the primary speech content as well as the speech characteristics would be present and audible.

95. To the extent *Dantrey’s* captured human speech disclosures (which comprise captured speech from a human user along with background noise) do not teach *speech characteristics*, *Strake* expressly discloses such.

96. For example, *Strake* also discloses a noise removal speech processing system which takes in “an input speech signal that comprises clean speech and acoustic interference” (i.e., “background interference”). EX1005-Strake, [021]-[022]. And *Strake* further discloses that its speech includes “one or more *speech components*” such as “pitch frequency, pitch harmonic structure, formant structure, spectral envelope, [and] spectral phase.” EX1005-Strake, [031].

b) Motivation to Combine

97. To the extent not already present, a PHOSITA would have been motivated to include *Strake's* speech components of pitch frequency, pitch harmonic structure, formant structure, etc. as part of *Dantrey's* captured human speech with background noise with a reasonable expectation of success.

98. For example, including speech characteristics in captured speech is well-known in the art. For instance, U.S. Patent Application Publication No 2024/0304200 to Keshet et al. ("*Keshet*") discloses a system for "real-time, automated re-synthesis of a speaker's speech...in a way which preserves **characteristics of the original speech**" including, but not limited to, "perceived speaker voice and identity, naturalness, intonation, and/or rhythm." EX1026-Keshet, [0058].

99. As such, *Keshet's* method necessitates the collection of the original speech characteristics with the "speaker's speech" because doing so allows the "perceived" speech to be maintained. In other words, if the speech characteristics were not included, any captured speech would be monotone or robotic. Thus, for at least this reason, a PHOSITA would have understood that including the aforementioned speech characteristics with captured human speech is both advantageous and trivial. Accordingly, a PHOSITA would have been motivated to include *Strake's* speech components including pitch frequency and pitch harmonic structure, formant structure, etc. in captured human speech as previously discussed

to improve the authenticity of human speech used in subsequent speech processing efforts.

100. Additionally, *Dantrey* and *Strake's* express disclosures would have motivated such combination. *Dantrey* discloses its system processes audio data from people using digital communications such as teleconferencing and describes methods to remove “undesirable” sound or sound which “may degrade a quality or clarity of captured speech,” for example, sound which “does not match a pattern of human speech.” EX1004-*Dantrey*, 2:64-3:29. And *Strake* discloses that its speech components “are relevant to the speech quality and intelligibility[.]” EX1005-*Strake*, [031]. Accordingly, a PHOSTIA would have understood that *Dantrey* strives to maintain the quality of human voices in live teleconferencing contexts and that *Strake's* speech components are one example of how to maintain such quality. By maintaining the speech components, *Strake* improves the listenability of the modified speech. See Section IV.C. These components improve the naturalness of the speech and improve the listenability of the speech to a human listener. EX1010-Schmandt, 120-121. Maintaining the speech components while reducing noise would have maintained the quality in contexts such as teleconferencing.

101. Further, such combination would have constituted combining prior art elements (*Dantrey's* captured human speech and *Strake's* speech components) according to known methods (*Strake's* maintaining of pitch/formant structure/etc.-

related information) to yield predictable results (preserving human aspects of speech). By retaining speech components such as pitch, formant structure, etc. of the speaker, the output speech would sound more natural and realistic, improving listenability of the modified speech. *See* Section. IV.C. As such, a PHOSITA would have understood the combination improves *Dantrey's* system in the same way as the *Strake's* system is improved by retaining pitch related information, namely keeping the speech sounding like the human it was captured from.

102. A PHOSITA would have had a reasonable expectation of success in making such combination at least because it was well-known how to maintain pitch, formant structure, etc. information in captured speech and because such would have only required minor modifications in the signal processing of the captured speech. *Dantrey* and *Strake* each describe neural network-based noise removal systems *Compare* EX1004-*Dantrey*, Abstract, 4:1-4, 4:22-27, 5:64-6:1, 7:41-45, 7:66-8:6; *with* EX1005-*Strake*, Abstract, [019]-[022], [025], [029], [084]. A PHOSITA would have understood how to make the necessary programming changes required to implement the functionality of *Strake* in the similar system of *Dantrey*.

(1) *Fragmenting input audio data into frames*

103. *Dantrey* teaches a feature extraction process which operates on *Dantrey's* input audio signal, as explained in detail below in Claim 1[b]. *Dantrey* specifically discloses such “feature extraction process can support a sample rate of

about 16000 Hz, with a FFT size of 1024 samples and Hop size of 256 samples.”

EX1004-Dantrey, 6:28-30. A PHOSITA would have understood such sample parameters teach breaking up the input audio into *speech frames* for *Dantrey’s* subsequent feature extraction process. Input audio having specified speech frames taught or otherwise indicated by sample parameter disclosures was well known in the art, including for systems utilizing a sample rate and sample quantity within a similar range as that which is described by *Dantrey* for example in *Phase-Aware Deep Speech Enhancement: It’s All About the Frame Length* to Peer et al. (“*Peer*”).

EX1027-Peer. *Peer* discloses that sample parameter disclosures are indicative of frame size and may be calculated according to the following equation and corresponding variable definition: “As M is the number of samples in a single frame, we define $M_t = \frac{M}{f_s}$ (where f_s is the sampling frequency) as the physical frame length, measured in second. The term frame length will refer to M_t ...[.]” EX1027-Peer, 3.

According to *Dantrey’s* disclosed speech frame characteristics and using basic mathematics in accordance with the algebra taught by *Peer*, wherein the disclosed sample quantity divided by sample rate yields the frame size, a PHOSITA would have determined a speech frame as having a frame size of 64 ms.

$$M_t = \left(\frac{1,024 \text{ samples}}{16,000 \frac{\text{samples}}{\text{s}}} \right) * \frac{1,000 \text{ ms}}{1 \text{ s}} = 64 \text{ ms}$$

104. To the extent *Dantrey's* sample parameter disclosures (which teach a specific exemplary frame size) do not teach *fragment[ing] input audio into a plurality of input speech frames*, *Strake's* disclosures do.

105. For example, *Strake* expressly discloses that its “input signal $y(n)$ [] is input to a framing circuit ... and processed in frames of 32msec.” EX1005-*Strake*, [037]. This is an additional express disclosure that audio data in a neural network based noise removal system is broken up into frames for subsequent processing.

c) Motivation to Combine

106. To the extent not already present, a PHOSITA would have been motivated to incorporate *Strake's* framing functionality into *Dantrey's* system with a reasonable expectation of success, as systems and methods were well-known within the art regarding, for example, how to divide input audio data into frames for subsequent processing. For instance, U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“*Prabhavalkar*”) teaches receiving audio data for an audio utterance, determining speech frames representing audio characteristics from the audio data, and successively processing the speech frames via a speech recognition model. EX1028-Prabhavalkar, 6:56-7:4. Likewise, *Strake* discloses an input signal divisible and processed in frames (e.g., 32 millisecond frames having a 50% frame shift). EX1005-*Strake*, [037]. As such, a PHOSITA would have understood the

commonality of dividing audio data into frames and been motivated to combine *Strake's* framing with *Dantrey's* system for at least this reason.

107. As another example, *Dantrey* describes its system as a “pipeline” and describes exemplary sample sizes and rates of data for input into such pipeline. EX1004-Dantrey, 3:54-57, 6:27-30, Fig. 1. A PHOSITA would have understood that “[i]n most speech-processing systems, speech signals are first windowed into frames[,]” as taught by *On the Use of Variable Frame Rate Analysis in Speech Recognition* to Zhu et al. (“*Zhu*”). EX1029-Zhu, 1 lhc; *see also*, EX1028-Prabhavalkar, 6:56-7:4. *Zhu* further teaches that “[t]he justification for such a segmentation is that speech signals are non-stationary and exhibit quasi-stationary behavior at the shorter durations.” EX1029-Zhu, 1 lhc. Accordingly, a PHOSITA would have understood that framing the input speech would have allowed processing on “quasi-stationary behavior” thus allowing for reliable processing/prediction as well as keeping a finite length of data segments to maintain the pipeline flow. And a PHOSITA would have understood that continuous audio streams are advantageously broken up into frames before subsequent processing.

108. Further, *Dantrey* teaches processing based on representations. EX1004-Dantrey, 4:50-53, 6:5-8, 6:35-40. A PHOSITA would have understood that a continuous audio stream would need to be broken up into parts (frames) in order to create and subsequently process such representations.

109. Such a combination would have also constituted combining prior art elements (*Dantrey's* pipeline system and *Strake's* framing functionality) according to known methods (*Strake's* framing process) to yield predictable results (allowing processing on quasi-stationary audio data and maintaining pipeline flow). As explained above, it was well-known both to frame audio data for subsequent processing and how to do so. A POHSITA would have understood the inclusion of framing functionality in a pipelined system to be trivial for all of the reasons explained above. And a PHOSITA would have understood that doing so would have provided the benefit and predictable result of allowing processing on discrete data portions.

110. A PHOSITA would have had a reasonable expectation of success in making such combination at least because it was well-known how to divide input audio data into frames for subsequent processing, and because such would have only required minor modifications in the signal processing of the captured speech. A PHOSITA would have understood and found it simple to frame audio data prior to further processing. *See* EX1029-Zhu, 1 lhc.

3. Claim 1[b]: convert the input speech frames to low-dimensional representations of the input speech frames, wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data and the low-dimensional representations of the input speech frames omit one or more of the non-content elements;

111. In my opinion, *Dantrey-Strake* in view of *Chen* renders obvious *conver[ting] the input speech frames* (e.g., *Dantrey-Strake*'s frames of input speech) *to low-dimensional representations of the input speech frames* (e.g., *Dantrey*'s reduced noise spectrograms that have *Chen*'s PCA dimensionality reduction technique applied to them), *wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data* (e.g. *Dantrey*'s noise subtractor neural network applying *Chen*'s machine learning PCA technique) *and the low-dimensional representations of the input speech frames omit one or more of the non-content elements* (e.g. *Dantrey*'s reduced noise spectrograms that have *Chen*'s ML PCA technique applied to them omit background noise), as claimed.

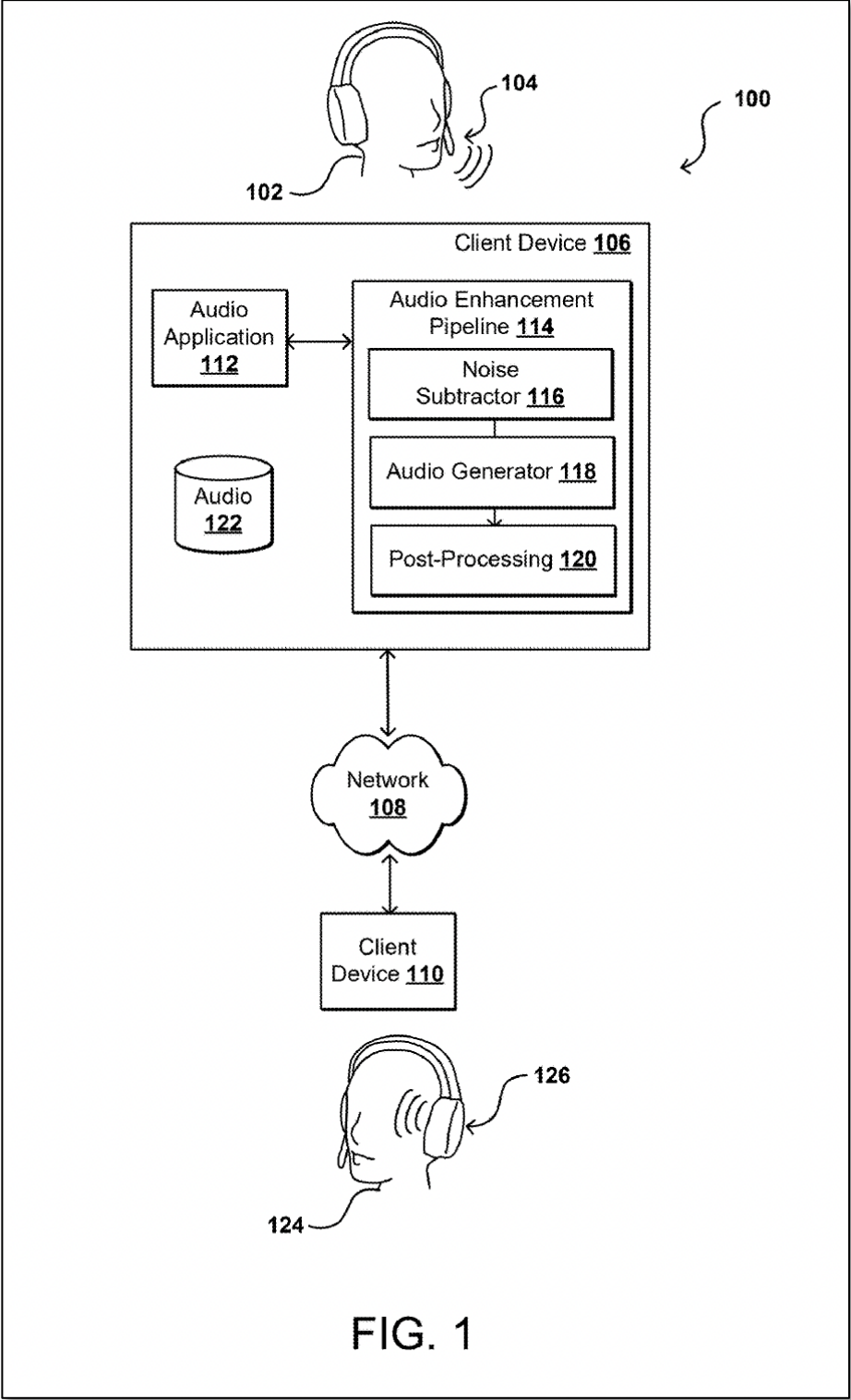
112. The '496 patent explains that "features may be extracted by the voice enhancement system 100 such as by using Fourier Transform, **Mel-Frequency Cepstral Coefficients (MFCC)**, or other techniques" and that "[t]he extracted features may be encoded by the voice enhancement system 100 into the **low-dimensional input audio data representation** 404 in step 506 using techniques such

as *Principal Component Analysis (PCA)*, Linear Discriminant Analysis (LDA), or other dimensionality reduction techniques.” EX1001, 7:22-27.

113. As mentioned above, for this Ground (as well as Grounds 2-3 which are further based on this Ground), I have been instructed to apply the construction of “a compressed representation of an input speech frame which is a result of a dimensionality reduction technique” for the term *low-dimensional representation*. In my opinion, the *Dantrey-Strake-Chen* combination teaches *low-dimensional representation* under such construction as explained below.

114. *Dantrey* teaches its “digital audio signal” from the captured speech is fed into a noise subtractor 116 because it is fed into *Dantrey’s* audio pipeline, which comprises noise subtractor 116 (*a first neural network*), audio generator 118 (*a second neural network*), and post-processing 120. EX1004-Dantrey, Fig. 1, 3:54-65, claim 7.

115. For example, *Dantrey* discloses human speech is captured and digitized. EX1004-Dantrey, 3:2-10. *Dantrey* next discloses that its “digital audio signal [is] provided as input to an audio enhancement pipeline 114.” EX1004-Dantrey, 3:55-57. As seen in *Dantrey’s* Figure 1, the first stage in this pipeline is “noise subtractor 116”:



EX1004-Dantrey, Fig. 1. Accordingly, such digital audio signal is fed directly into noise subtractor 116. I further note that the distinction between *Dantrey's* Claims 7

and 12 cements this understanding. For example, Claim 7 recites “cause one or more digital signals to be filtered, where one or more portions of speech are removed from the one or more filtered digital signals; generate, by one or more neural networks, the one or more portions of speech filtered from the one or more digital signals based, at least in part, on the one or more filtered digital signals; and add the one or more portions of speech, generated by the one or more neural networks, to the one or more filtered digital signals.” And Claim 12 recites “12. The system of claim 7, wherein the one or more processors are further to provide the one or more filtered digital signals as input to the one or more neural networks as one or more spectrograms.” Accordingly, a PHOSITA would have understood that Claim 7 states that the first neural network operates on digital audio data, while Claim 12 adds the additional requirement (which is not a requirement of Claim 7) that the input to the neural network can be a spectrogram.

116. *Dantrey* teaches that its “noise subtractor” is a neural network. EX1004-Dantrey, 4:45-47 (“noise subtractor 116 is a frequency-domain deep learning-based noise subtractor (or suppressor) network.”), 4:1-4 (“noise subtractor 116 and audio generator 118 can involve neural network-based tasks”), 4:22-27 (“primary audio [] can be enhanced by removing background noise using a first neural network”).

117. *Dantrey* explains that “feature extraction is performed by noise subtractor 116 using *Mel Frequency Cepstral Coefficient (MFCC)*,” and that “this [noise subtractor] network reduces and blurs noise energy” by producing *Dantrey’s* “spectrogram 204” based on such MFCCs. Specifically, “a spectrogram 202 for input audio is passed as input to a noise subtractor 116.” EX1004-Dantrey, 4:44-45. “[F]eature extraction is performed by noise subtractor 116 using Mel Frequency Cepstral Coefficients (MFCC), a delta of MFCC, and band energy stacked together.” EX1004-Dantrey, 4:50-53. The noise subtractor “reduces and blurs noise energy in input audio spectrogram 202 without reducing speech energy.” EX1004-Dantrey, 4:53-55. The noise subtractor 116 outputs “a spectrogram 204 having background noise, and at least some foreground noise, reduced by noise subtractor 116” that “can be provided as input to an audio generator 118.” EX1004-Dantrey, 5:4-7. Accordingly, *Dantrey* teaches a noise reduced spectrogram is passed from its first neural network to its second neural network, which such noise reduced spectrogram is based on features, for example extracted via MFCC.

118. *Dantrey* teaches “noise subtractor 116” is a neural network which “reduce[s] a presence of background noise” and outputs “an audio spectrogram with reduced background noise” to audio generator 118. EX1004-Dantrey, Fig. 1, 3:54-65, 4:45-47 (“noise subtractor 116 is a frequency-domain deep learning-based noise subtractor (or suppressor) network.”), 4:1-4 (“noise subtractor 116 and audio

generator 118 can involve neural network-based tasks”), 4:22-27 (“primary audio [] can be enhanced by removing background noise using a first neural network”).

119. *Dantrey* explains that “feature extraction is performed by noise subtractor 116 using *Mel Frequency Cepstral Coefficient (MFCC)*,” and that “this [noise subtractor] network reduces and blurs noise energy” by producing *Dantrey’s* noise reduced spectrogram from such MFCCs. That is, *Dantrey* teaches processing “spectrograms for an input audio signal” using “an audio processing pipeline[.]” EX1004-Dantrey, 4:41-43. The spectrogram “is passed as input to a noise subtractor” which is “a frequency-domain deep learning-based noise subtractor (or suppressor) network.” EX1004-Dantrey, 4:43-48. The noise subtractor performs feature extraction “using Mel Frequency Cepstral Coefficients (MFCC), a delta of MFCC, and band energy stacked together.” EX1004-Dantrey, 4:50-53. As described, the noise subtractor “reduces and blurs noise energy in input audio spectrogram 202 without reducing speech energy.” EX1004-Dantrey, 4:53-55.

120. Accordingly, *Dantrey’s* noise subtractor teaches *a first neural network* which produces *representations of the input speech* from the *Dantrey-Strake input speech frames*.

121. Notably, *Dantrey* discloses that its “pre-trained model(s) [] may include any type[] of machine learning models depending on implementation or embodiment” and that such machine learning models “include machine learning

model(s) using linear regression, logistic regression, ... *dimensionality reduction algorithms* ... and/or other types of machine learning models.” EX1004-Dantrey, 78:51-64. *Dantrey* does not expressly provide an express example of such a dimensionality reduction algorithm. However, *Chen* does.

122. *Chen* teaches “Noise Mitigation Using Machine Learning” and specifically discloses that “*ML models can employ a dimensionality reduction approach*, such as, one or more of ... an Incremental *Principal Component Analysis (PCA)* algorithm[.]” EX1006-Chen, Title, 3:66-4:4. I note that PCA was a very well-known dimensionality reduction technique in the prior art. EX1014-Zheng, 2-4; EX1031-Wells, 13:36-42; EX1015-Wang, 3 (see §4.3). And PCA was also well-known to be applied to a variety of different data formats, specifically including spectrograms and MFCCs. EX1008-Heckmann, 737, Fig.1; EX1014-Zheng, 2-4; EX1031-Wells, 13:36-42; EX1015-Wang, 3 (see §4.3). As the prior art explains, PCA has been known to be “useful to retain the most important components from the signal, and to denote both noise and the background EEG.” EX1012-Martinek, 18 (see §3.3.3).

123. Accordingly, the *Dantrey-Strake-Chen* combination teaches Claim 1(b), including the construction of *low-dimensional representation* because *Dantrey’s* “noise subtractor 116” neural network applies *Chen’s* ML PCA dimensionality reduction technique to *Dantrey’s* noise reduced spectrogram

representation, which is based on *Dantrey's* MFCC extracted features, before sending such PCA-reduced representation to the second neural network (i.e., *Dantrey's* audio generator 118) in *Dantrey's* pipeline, and *Dantrey's* first noise subtractor 116 module in the pipeline receives the digital audio signal of the captured speech. EX1004-Dantrey, 3:54-64, 4:41-5:7, Fig. 1, Claim 7. PCA was a well-known technique to reduce dimensionality and was specifically well-known to be applied to spectrograms in speech processing, as mentioned in the previous paragraph. See Section IV.B.1. For example, *Wang* teaches that “Linear Discriminant analysis (LDA) and Principal Component Analysis (PCA) are the two popular feature transformation methods.” EX1015-Wang, 1 (see Abstract). *Wang* “investigates their performances in dimensionality reduction tasks in continuous speech recognition systems.” *Id.* Reducing the dimensionality through these techniques leads to “less computational cost and system complexity” making said techniques “necessary for robust speech recognition.” EX1015-Wang, 1 (see §1).

124. *Dantrey's* reduced noise frame spectrograms representations (based on the extracted MFCCs) with their dimensionalities reduced by the well-known PCA technique (the claimed *low-dimensional representations*) of the *Dantrey-Strake-Chen* system omit one or more of the non-content elements (i.e., *background noise*) because *Dantrey* expressly states that such representations “hav[e] background noise

[] reduced by noise subtractor 116[.]” EX1004-Dantrey, 4:41-5:7, 3:60-61, 4:23-25, Fig. 1.

a) Motivation to Combine

125. A PHOSITA would have been motivated to incorporate *Chen’s* PCA technique into *Dantrey’s* noise subtractor neural network with a reasonable expectation of success for several reasons.

126. As a first example, a PHOSITA would have understood that “Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are the two popular feature transformation methods” and that using the PCA dimensionality reduction techniques in “continuous speech recognition (CRS) systems” have been established to be advantageous. EX1015-Wang, 1 (see Abstract). These dimensionality reduction methods “concentrate the energy distributions of a speech signal onto fewer dimensions than those of parameter extraction and thus reduce the dimensionality of the system.” EX1015-Wang, 1 (see Abstract). The reduced dimensionality lessens “computational cost and system complexity” making implementation of said techniques “necessary for robust speech recognition.” EX1015-Wang, 1 (see §1). Applying a dimensionality reduction technique such as PCA saves the most important data characteristics while reducing the complexity by reducing redundant data coding information. And as mentioned above, PCA has been known to be “useful to retain the most important components from the signal,

and to denote both noise and the background EEG.” EX1012-Martinek, 18 (see §3.3.3). Accordingly, a PHOSITA would have understood that PCA is a well-known dimensionality reduction technique which allows for efficient processing by removing unnecessary information while maintaining the most important components of the signal.

127. As another example, such a combination would have constituted combining prior art elements (*Dantrey’s* noise subtractor neural network and *Chen’s* ML PCA technique) according to known methods (the well-known PCA technique) to yield predictable results (reducing the dimension of feature representations to improve on computation expense and maintain the most important components). *See* EX1015-Wang, 1 (see §1); EX1006-Chen, 3:66-4:4; EX1012-Martinek, 18 (see §3.3.3, PCA is “useful to retain the most important components from the signal, and to denote both noise and the background EEG.”). A PHOSITA would have recognized that using PCA to encode the features allows for faster processing because a PCA reduced data set has lower dimensions than its previous form, and therefore the computer does not have to pass along as many binary numerals to represent the same information (the important components of the signal).

128. A PHOSITA would have had a reasonable expectation of success in making such combination at least because PCA was such a well-known technique a PHOSITA would have known how to implement, *Chen* establishes evidence that it

was known how to apply PCA using machine learning models, *Zheng/Wang* establishes that it was known how to apply PCA to spectrogram representations/MFCCs, and such a combination would have only required minor programming modifications. EX1006-Chen, 3:66-4:4 (“ML models can employ a dimensionality reduction approach, such as, ... an Incremental Principal Component Analysis (PCA)[.]”); EX1014-Zheng, 2 rhc(“The spectrogram is further log-transformed and processed using PCA whitening (with 60 components) to reduce the dimensionality and some interference[.]”); *see* EX1015-Wang, 1.

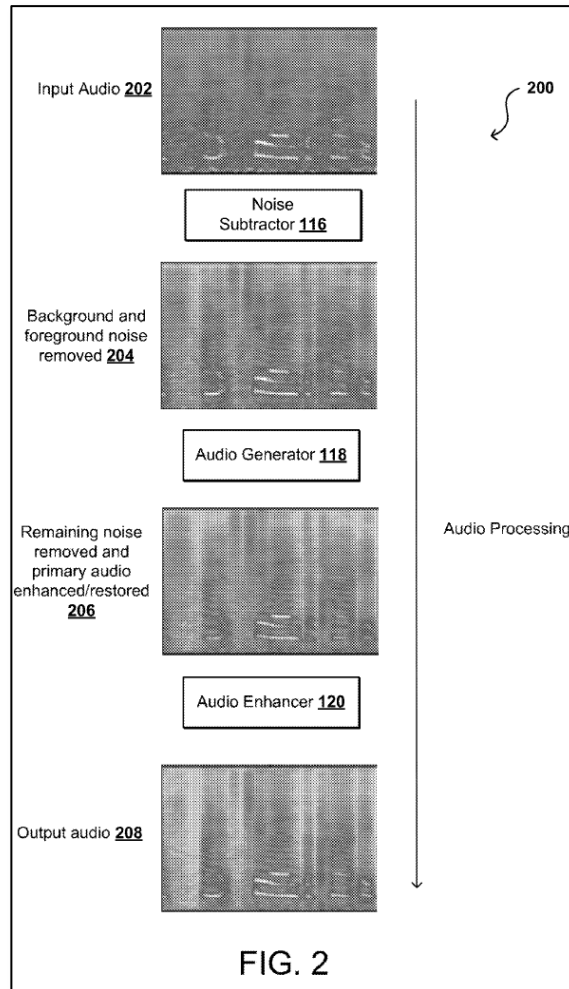
4. Claim 1[c]: apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; and

129. In my opinion, *Dantrey-Strake-Chen* renders obvious *apply[ing] a second neural network (e.g., Dantrey’s audio generator) to the low-dimensional representations of the input speech frames (e.g., Dantrey’s reduced noise spectrograms that have Chen’s PCA dimensionality reduction technique applied to them) to generate target speech frames (e.g., Strake’s time domain frames), as claimed.*

130. *Dantrey* teaches that its “**audio generator**” (which receives the output from *Dantrey’s* noise subtractor as frames of PCA-reduced spectrograms with reduced noise) “**is a frequency-domain to audio, deep learning-based clean speech generator (e.g., generator network) and an auto-encoder (e.g., speech enhancer).**”

EX1004-Dantrey, 5:7-10. *Dantrey* continues that “a generative deep learning network (DNN) can be used, such as a latent modified flow-based WaveGlow network from NVIDIA Corporation. In at least one embodiment, such a generative network can be used *to generate clean speech audio from noise-reduced spectrograms[.]*” EX1004-Dantrey, 5:11-16, 4:1-4 (“noise subtractor 116 and audio generator 118 can involve neural network-based tasks”), 6:35-46.

131. *Dantrey* does not expressly disclose all the details for how its audio generator produces clean speech audio from noise-reduced spectrograms or expressly state that its clean speech is produced in output speech *frames*. However, in the *Dantrey-Strake-Chen* system explained above, *Dantrey*'s pipeline is processed in frames per *Strake*, particularly based on *Dantrey*'s express sample size, sample rate, and hop size. Accordingly, a PHOSITA would have understood that a pipeline such as *Dantrey*'s that takes in frames and passes frames of data from *Dantrey*'s noise subtractor to *Dantrey*'s audio generator also produces its output clean speech as *frames*. Specifically, as each PCA-reduced spectrogram received by the audio generator represents a frame of audio data, output of the generator for each spectrogram similarly represents the same frame of audio data. For example, Fig. 2 of *Dantrey* shows the audio generator produces a spectrogram 206 with “remaining noise removed and primary audio enhanced/restored[.]”



EX1004-Dantrey, Fig. 2.

132. To the extent such does not teach *generat[ing] target speech frames*, *Strake* expressly discloses techniques for both *generate target speech frames* and subsequently *combin[ing] such frames*, as claimed (I note the latter is the subject of Claim 1[d]).

133. Specifically, *Strake* discloses “[a] second neural network circuit [] configured to receive [a] representation of [] estimated clean speech and restore speech components of the clean speech in the input speech signal, and suppress any

residual acoustic interference, and output a reconstructed speech signal.” EX1005-Strake, [022]. Within such “second neural network circuit,” *Strake* discloses taking “*frame representations*” of the clean speech signal and “*applying an inverse transform from the processing domain back to the time domain*, together with a subsequent combination *of the time domain frames*, e.g., by Inverse Fast Fourier Transform (IFFT) and an overlap add (OLA) operation, in one example embodiment.” EX1005-Strake, [042].

a) **Motivation to Combine**

134. A PHOSITA would have been motivated to incorporate *Strake*’s frame-based inverse transform technique (which produces time domain frames) into the *Dantrey-Strake-Chen* system explained above with a reasonable expectation of success.

135. For example, and as previously mentioned, speech processing pipelines such as *Dantrey* advantageously use frames to help keep the processing pipeline appropriately filled and because “speech signals are non-stationary and exhibit quasi-stationary behavior at the shorter durations.” EX1029-Zhu, 1 lhc. In other words, speech is constantly changing over time, but if you look at it in very short time slices, it stays roughly the same within each slice, and this allows for reliable processing. Accordingly, a PHOSITA would have been motivated to process speech data in *Dantrey*’s audio generator also in frames.

136. Additionally, processing data in frames at the output of a speech generation model was known to improve the real time operation of speech processing systems. For example, frame-based processing was known to reduce delays. EX1016-Wu, 18 (“To further reduce the delay of voice conversion, frame-based approaches capable of converting spectral parameters frame by frame are more desirable”). A PHOSITA would have understood that such a desirable frame-based approach improves the real time operation specifically because it reduces delays.

137. Further, such a combination would have constituted combining prior art elements (*Dantrey’s* audio generator and *Strakes’* frame-based inverse transform technique) according to known methods (well-known frame-based inverse transform technique) to yield predictable results (allowing pipelined speech processing systems to temporally operate predictably and reduce delay). As explained above, frame-based processing in speech systems was both well-known and specifically well-known to reduce delay. *Strake’s* frame-based inverse transform technique was also well-known. For example, *Li* teaches that the inverse Fourier transform is used “when the processed frequency decomposed signals need to be converted back real signals, such as speech and music synthesis and noise reduction[.]” EX1030-Li, 2-3. And further, as explained in Paragraph 128, because speech is constantly changing

over time, dividing a speech signal into frames is advantageous because it allows for predictable operation on such.

138. A PHOSITA would have had a reasonable expectation of success in making such combination at least because applying frame-based inverse transforms was well-known and well-within a PHOSITA's capabilities, and such combination would have only required minor programming modifications. As *Li* explains, the inverse transform is an equation readily applied to audio. EX1030-Li, 2-3. As such, a PHOSITA would have understood how to implement an inverse transform so the *Dantrey-Strake-Chen* system produces time domain frames.

5. Claim 1[d]: combine the target speech frames to generate output audio data, wherein the output audio data further comprises one or more portions of the foreground speech content and one or more of the speech characteristics.

139. In my opinion, *Dantrey-Strake-Chen* renders obvious *combin[ing] the target speech frames* (e.g., *Strake's* overlap add operation which combines time domain frames) *to generate output audio data* (e.g., *Dantrey's* clean speech audio), *wherein the output audio data further comprises one or more portions of the foreground speech content* (e.g., *Dantrey's* captured human speech) *and one or more of the speech characteristics* (e.g., *Strake's* speech components), as claimed.

140. As explained above, *Dantrey* teaches that its audio generator produces clean speech audio but does not go into detail (on its own) for how such audio

generator goes from the “frequency-domain to audio” to do so. *See* Claim 1[c]. EX1004-Dantrey, 5:7-16 (“**audio generator 118 is a frequency-domain to audio, deep learning-based clean speech generator (e.g., generator network) and an auto-encoder (e.g., speech enhancer)** ... such a generative network can be used **to generate clean speech audio from noise-reduced spectrograms**”), 6:35-46.

141. However, and as also discussed above, the *Dantrey-Strake-Chen* system *generate[s]* *target speech frames* with *Dantrey’s* audio generator neural network via *Strake’s* frame-based inverse transform technique (which produces time domain frames of output data). *See* Claim 1[c].

142. *Strake* further discloses such time domain frames are subsequently combined via an overlap add (OLA) operation. EX1005-Strake, [042] (discussing how the frame representations of clean speech are combined: “this may be achieved by applying an inverse transform from the processing domain back to the time domain, **together with a subsequent combination of the time domain frames**, e.g. by Inverse Fast Fourier Transform (IFFT) and **an overlap add (OLA) operation**”).

143. Accordingly, the *Dantrey-Strake-Chen* system teaches *combin[ing]* the *target speech frames to generate output audio data*, as claimed.

144. The *Dantrey-Strake-Chen* system further teaches *the output audio data further comprises one or more portions of the foreground speech content and one or more of the speech characteristics*, as claimed. For example, the *Dantrey-Strake-*

Chen system teaches that noisy speech from a user captured with a microphone (which includes prominent speech content, speech characteristics, and noise) is fed through *Dantrey's* pipeline to remove the noise and retain the captured speech (including the prominent speech content and speech components/characteristics), resulting in *Dantrey's* clean speech. EX1004-Dantrey, Title, Abstract, 2:64-3:6, 3:61-64, 4:22-25, 4:53-55, 5:4-16.

a) **Motivation to Combine**

145. A PHOSITA would have been motivated to incorporate *Strake's* overlap add (OLA) technique into *Dantrey's* clean speech generation pipeline with a reasonable expectation of success.

146. For example, a PHOSITA would have understood that when working with a pipeline which processes samples/frames based on a hop size (such as *Dantrey*), an OLA technique (such as *Strake's*) provides benefits of maintaining continuity and preventing information from being lost at the edges of frames. EX1017-Sargsyan, [0065] (“us[ing] the overlap-add method” and “overlapping frames maintains continuity between the frames, and prevents information at the edges of the frames from being lost.”). Using such an overlap add technique helps ensure that content is captured which could otherwise be missed if it falls near the edge of a frame. Further, a PHOSITA would have known that using OLA is advantageous because it smooths the frames boundaries thus reducing the possibility

of artifacts. EX1033-Thyssen, [0112] (“any waveform discontinuity in the signal ... will be smoothed out by the overlap-add operation”).

147. Additionally, such a combination would have constituted combining prior art elements (*Dantrey’s* speech processing pipeline and *Strake’s* OLA technique) according to known methods (the well-known OLA technique) to yield predictable results (maintaining continuity, preventing information from being lost at the edges of frames, smoothing, and preventing artifacts). As previously explained the OLA technique, and its benefits, were both well-known in the art. PHOSITAs were very aware of OLA, and its benefits, and regularly employed such for its predictability.

148. A PHOSITA would have had a reasonable expectation of success in making such combination at least because OLA was well-known and well-within a PHOSITA’s capabilities to implement, and because such would have only required minor programming modifications. In fact, OLA was known at least as early as 1980 as a method combining segments after an inverse Fourier transformation EX1009-Crochiere 1 (“Allen [3] carefully discussed a method of synthesizing a signal from its short-time Fourier spectra by inverse transforming each sample of the short-time spectra to recover the short-time segments of the signal in time. These overlapped signal segments are then appropriately summed (overlapped and added) to reproduce the time signal. *This method is referred to as the overlap-add synthesis method.*”)

B. Claim 2

1. **Claim 2[Pre]: The voice enhancement system of claim 1, further comprising a physical microphone and an audio output device, wherein the one or more processors are further configured to execute the instructions to:**

149. In my opinion, *Dantrey-Strake-Chen* renders obvious *the voice enhancement system of claim 1, further comprising a physical microphone (e.g., Dantrey’s microphone) and an audio output device (e.g., Dantrey’s speaker), wherein the one or more processors (e.g., Dantrey’s processor) are further configured to execute the instructions, as claimed.*

150. The '496 Patent provides examples of “headphones or speaker(s)” as “an audio output device.” EX1001, 3:57-58.

151. *Dantrey* discloses that the speech of a user is captured via “a **microphone** 104 or other audio capture device, as may be part of a headset of the computing device” and that such speech may “presented to second person 124 using at least **one speaker 126** or presentation mechanism, as may be part of a **headset or audio speaker.**” EX1004-Dantrey, 3:2-15, Fig. 1.

2. **Claim 2[a]: digitize analog input audio signals obtained via the physical microphone to generate the input audio data;**

152. In my opinion, *Dantrey-Strake-Chen* renders obvious, *digitiz[ing] analog input audio signals (e.g., Dantrey’s producing of its digital audio signal)*

obtained via the physical microphone (e.g., Dantrey's microphone) to generate the input audio data (e.g. Dantrey's digital input audio signal), as claimed.

153. *Dantrey* discloses that its human speech captured from a user by microphone 104 is “provided to a client device 106 [] which can **produce a digital audio signal**[.]” EX1004-*Dantrey*, 3:2-9, 69:59 (disclosing “a digital-to-analog converter (‘DAC’), **and like**.”). 95:21-40 (discussing how “obtaining, acquiring, receiving, or inputting **analog or digital data** ... can be accomplished in a variety of ways”). A PHOSITA would have understood that “produc[ing] a digital audio signal” from speech captured from a microphone (as taught by *Dantrey*) teaches *digitiz[ing] analog input audio signals*, as claimed, because microphones were well-known to capture analog signals. For instance, *Kee* teaches a audio interface that “includes a **digital-to-analog converter** having a pair of outputs that are coupled to the speakers” and “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”. EX1034-*Kee*, [0027].

154. To the extent such does not expressly teach Claim 2[a], a PHOSITA would have found it obvious to “produce [such] digital audio signal” of *Dantrey's* from an analog signal, for example via's *Dantrey's* “digital-to-analog converter (‘DAC’), **and like**” where the “**like**” would have been the reciprocal analog-to-

digital converter. EX1004-Dantrey, 3:2-9, 69:59; EX1034-Kee, [0027] (teaching an audio interface that “includes a *digital-to-analog converter* having a pair of outputs that are coupled to the speakers” and “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[.]”). A PHOSITA would have had a reasonable expectation of success in doing so because analog-to-digital converters were well known and easily implementable by PHOSITAs. *See* EX1034-Kee, [0027].

3. Claim 2[b]: convert the output audio data to analog audio output signals; and

155. In my opinion, *Dantrey-Strake-Chen* renders obvious *convert[ing] the output audio data* (e.g., *Dantrey’s* clean speech audio) *to analog audio output signals* (e.g., *Dantrey’s* audio that is to be transmitted or presented), as claimed.

156. *Dantrey* discloses that its digital audio signal can be sent to “another client device 110, which can cause this digital audio signal to be presented to second person 124 using at least one speaker 126 or presentation mechanism, as may be part of a headset or audio speaker.” EX1004-Dantrey, 3:10-15, 3:44-49 (explaining the audio signal can be improved “before transmitting that speech to client device 110 for presentation (e.g., *providing playback through at least one speaker 126*) to second user 124.”), 4:11-17 (“post-processing can involve ... *adjusting a format of*

*an audio signal for playback[] this enhances [sic] audio signal can then be transmitted for presentation to second person 124 through an appropriate speaker 126 or playback mechanism”), 69:59 (disclosing “a digital-to-analog converter (‘DAC’), and like.”); see also 95:21-40 (discussing how “obtaining, acquiring, receiving, or inputting **analog or digital data** ... can be accomplished in a variety of ways”). A PHOSITA would have understood that these *Dantrey* disclosures render obvious *convert[ing] the output audio data to analog audio output signals*, as claimed.*

157. A PHOSITA would have been motivated to adjust the format of the digital output to analog based on *Dantrey’s* express statements about providing playback through a second user’s speaker, adjusting a format of an audio signal for playback, and digital-to-analog converters. EX1004-*Dantrey*, 3:10-15 (“In at least one embodiment, this digital audio signal can be received to another client device 110, which can cause this digital audio signal to be presented to second person 124 using at least one speaker 126 or presentation mechanism, as may be part of a headset or audio speaker.”), 3:44-49 (“In at least one embodiment, an audio application 112 executing on client device 106 can attempt to improve a quality of speech contained in a digital audio signal before transmitting that speech to client device 110 for presentation (e.g., providing playback through at least one speaker 126) to second user 124.”), 4:11-14 (“In at least one embodi-ment, post-processing can involve

improving clarity of audio, adjusting one or more values of an output audio signal, or adjusting a format of an audio signal for playback.”), 69:59 (“a digital-to-analog converter (“DAC”), and like.”), 95:21-41 (“In present document, references may be made to obtain-ing, acquiring, receiving, or inputting analog or digital data into a subsystem, computer system, or computer-implemented machine. Obtaining, acquiring, receiving, or inputting analog and digital data can be accomplished in a variety of ways such as by receiving data as a parameter of a function call or a call to an application programming interface. In some implementations, process of obtaining, acquiring, receiving, or inputting analog or digital data can be accomplished by transferring data via a serial or parallel 30 interface. In another implementation, process of obtaining, acquiring, receiving, or inputting analog or digital data can be accomplished by transferring data via a computer network from providing entity to acquiring entity. References may also be made to providing, outputting, transmitting, 35 sending, or presenting analog or digital data. In various examples, process of providing, outputting, transmitting, sending, or presenting analog or digital data can be accomplished by transferring data as an input or output parameter of a function call, a parameter of an application programming interface or interprocess communication mechanism.”). Because DACs were so well-known and easy to implement, a PHOSITA would have been motivated to use such a DAC (as expressly mentioned by *Dantrey*) to readily provide playback on a speaker of a

second user's device (as also expressly mentioned by *Dantrey*). See above at, for instance, EX1004-Dantrey, 3:10-15, 3:44-49, 4:11-14, 69:59, 95:21-41. A PHOSITA would have understood that speakers advantageously play audio in the form of analog. For example, *Kee* teaches an audio interface that “includes a digital-to-analog converter having a pair of outputs that are coupled to the speakers[.]” EX1034-Kee, [0027]. A PHOSITA would have also had a reasonable expectation of success using a DAC to convert Dantrey's digital audio output to analog form at least because the implementation of such was so well-known and would have only required minor software modifications. For example, *Kee* teaches an audio interface containing a digital-to-analog converter is part of a computer system. See EX1034-Kee, [0023], [0027], Fig. 4. A PHOSITA would have understood how a digital-to-analog would be implemented in similar devices.

4. Claim 2[c]: provide the analog audio output signals to the audio output device via one or more of a virtual microphone or a communication application executed by the voice enhancement system.

158. In my opinion, *Dantrey-Strake-Chen* renders obvious *provid[ing] the analog audio output signals (e.g., Dantrey's audio that is to be transmitted or presented) to the audio output device (e.g., Dantrey's speaker) via one or more of a virtual microphone or a communication application (e.g., Dantrey's audio*

application) *executed by the voice enhancement system (Dantrey's system to reduce noise and enhance received audio), as claimed.*

159. *Dantrey discloses “an **audio application** 112 executing on client device 106” which “**transmit[s] speech to client device 110 for presentation (e.g., providing playback through at least one speaker 126)** to a second user.” EX1004-Dantrey, 3:44-49.*

160. In the *Dantrey-Strake-Chen* combination described above in Claim 2[b], *Dantrey's* output audio is converted from digital to analog as part of *Dantrey's* “post-processing” where *Dantrey* discloses “this enhance[d] audio **can then be transmitted** for presentation to second person 124 through an appropriate speaker 126 or playback mechanism.” EX1004-Dantrey, 4:12-17.

C. Claim 3: The voice enhancement system of claim 1, wherein the non-content elements comprise one or more of background noise, microphone pops, low-fidelity audio, or audio clippings and the speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech.

161. In my opinion, *Dantrey-Strake-Chen* renders obvious *wherein the non-content elements comprise one or more of background noise, microphone pops, low-fidelity audio, or audio clippings (e.g., Dantrey's background noise) and the speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech (e.g., Strake's*

speech components which include pitch frequency, pitch harmonic structure, formant structure, etc.), as claimed.

162. See Claim 1[a].

D. Claim 7: The voice enhancement system of claim 1, wherein the second neural network comprises one or more of a diffusion probabilistic model, a flow-based model, or a generative adversarial network-based model.

163. In my opinion, *Dantrey-Strake-Chen* renders the voice enhancement system of claim 1, wherein the second neural network (e.g., *Dantrey's* audio generator) comprises one or more of a diffusion probabilistic model, a flow-based model, or a generative adversarial network-based model (e.g., *Dantrey's* flow-based network), as claimed.

164. *Dantrey* discloses that its audio generator can be a “flow-based” network. EX1004-*Dantrey*, 5:7-16 (“**audio generator** 118 is a frequency-domain to audio, deep learning-based clean speech generator (e.g., generator network) and an auto-encoder (e.g., speech enhancer). In at least one embodiment, a generative deep learning network (DNN) can be used, such as a latent modified **flow-based** WaveGlow network from NVIDIA Corporation. In at least one embodiment, such a generative network can be used to generate clean speech audio from noise-reduced spectrograms”), 6:35-39 (“a speech generator network can use a configuration 350 as illustrated in FIG. 3B. In at least one embodiment, this configuration corresponds

to a *flow-based network* capable of generating high quality speech from spectrograms”), Claim 9.

E. Claim 8: The voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to pre-process the input audio data by applying one or more of a noise reduction algorithm to remove background noise from the input audio data or a filtering technique to remove high-frequency noise or pops from the input audio data.

165. In my opinion, *Dantrey-Strake-Chen* renders obvious *the voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to pre-process the input audio data by applying one or more of a noise reduction algorithm to remove background noise from the input audio data or a filtering technique to remove high-frequency noise or pops from the input audio data* (e.g., *Strake’s* pre-processing to remove echos and reverberation), as claimed.

166. I first note that *Dantrey* discloses its system “filter[s] noise” from “one or more digital signals,” where *Dantrey’s* “one or more neural networks” then subsequently perform further noise reduction processing “based, at least in part, on the one or more filtered digital signals.” EX1004-*Dantrey*, Claims 7-8, 12. *Dantrey* later discloses that its “data may undergo pre-processing as part of data processing pipeline to prepare data for processing by one or more applications.” EX1004-*Dantrey*, 74:50-52.

167. *Strake* discloses a “pre-processing system” for removing certain noise aspects such as “echo[s]” and “reverberation” from an input audio signal before sending such to neural networks for subsequent noise removal processing. EX1005-*Strake*, [034]-[036]. A PHOSITA would have understood that echo refers to signal feedback from the far end of a communication network, and reverberation refers to room acoustics, such as from a speaker phone, and that both of these constitute unwanted background noise.

168. A PHOSITA would have been motivated to incorporate *Strake’s* pre-processing to remove echos and reverberation into the *Dantrey-Strake-Chen* system described above with a reasonable expectation of success. For example, a PHOSITA would have understood that *Dantrey* expressly discloses that its “data may undergo pre-processing as part of data processing pipeline to prepare data for processing by one or more applications” and that *Strake’s* preliminary removal of noise (before audio data is sent through subsequent neural network processing) would be one way to accomplish such pre-processing. EX1004-*Dantrey*, 74:50-52; EX1005-*Strake*, [034]-[036]. Additionally, a PHOSITA would have understood that *Strake’s* “pre-processing” system which removes “echo” and “reverberation” would be advantageous because such would allow the neural networks to focus their processing on more complex noise data. The application of a pre-processing system to improve the performance of subsequent speech enhancement (e.g., allowing the

neural networks to focus their processing on more complex noise data) was known in the art. *Unsupervised Learning Algorithm for Noise Suppression and Speech Enhancement Applications* to Alsheibi (“*Alsheibi*”) teaches that most real-time speech enhancement algorithms lose effectiveness when their input consists of too much noise and that inclusion of “a pre-processing step before applying speech enhancement algorithms...improves considerably the performance of speech enhancement algorithms when compared to other approaches with no pre-processing steps.” EX1035-Alsheibi, 7. *Alsheibi* teaches to introduce “a single pre-processing step in the overall speech enhancement framework to improve the input SNR[,]” that is to reduce noise as a pre-processing step. EX1035-Alsheibi, 18. It was also well known in the art that reducing noise would include removing echo and reverberation, as taught in U.S. 2015/0371655 to Gao (“*Gao*”), and taught by *Strake* as previously discussed. Specifically, *Gao* teaches methods for cancelling/reducing acoustic echos in speech/audio signal enhancement processing based on a echo tail length or the echo reverberation time, where noise is reduced. EX1036-Gao, Abstract, [0010], [0023]. A PHOSITA would have had a reasonable expectation of success at least because such techniques were well-known and a PHOSITA would have known how to implement such, particularly given the detailed architectural diagram in *Strake*’s Fig. 2 for “pre-processing system 250.” EX1005-Strake, Fig. 2.

F. Claim 9

1. **Claim 9[a]:** *The voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to:* extract one or more features from the input audio data, wherein the features comprise one or more of pitch, intonation, or formants; and

169. In my opinion, *Dantrey-Strake-Chen* renders obvious *the voice enhancement system of claim 1, wherein the one or more processors are further configured to execute the instructions to: extract one or more features (e.g., Dantrey’s extracted features) from the input audio data (e.g., Dantrey’s digital input audio signal), wherein the features comprise one or more of pitch, intonation, or formants (e.g., Strake’s pitch/formant structure/etc.-related information maintained in the Dantrey-Strake-Chen representations), as claimed.*

170. As mentioned above in Claims 1[a]-[b], in the *Dantrey-Strake-Chen* combination “**feature extraction** is performed by noise subtractor 116 **using Mel Frequency Cepstral Coefficient (MFCC)**” from *Dantrey’s* digital input audio signal, in the same manner as the ‘496 Patent. EX1004-Dantrey, 4:50-52; *see* Claims 1[a]-1[b]; EX1001, 7:17-19 (“**features may be extracted** ... by using Fourier Transform, **Mel-Frequency Cepstral Coefficients (MFCC)**, or other techniques.”); EX1037-Quatieri, 5:12-32 (““Mel Frequency Cepstral Coefficients’ (MFCC) refers to the coefficients that collectively make up a ‘mel-frequency cepstrum’ (MFC), which is a representation of the short-term power spectrum of a sound signal. The

term ‘cepstrum’ refers to the result of taking the Inverse Fourier transform (IFT) of the logarithm of the spectrum of a signal. The term ‘mel’ refers to the use of the ‘mel scale’ or similar filterbank by methods that obtain MFCC. The ‘mel scale’ is a perceptual scale of pitches judged by the listeners to be equal in distance from one another. The MFCCs are commonly derived as follows: (1) Take the Fourier transform of a windowed excerpt of a signal. (2) Apply the mel filterbank to the power spectrum obtained in (1), sum the energy in each filter. (The mel-scale filterbank is commonly implemented as triangular overlapping windows.) (3) Take the logarithm of all filterbank energies. (4) Take the discrete cosine transform (DCT) of the list of values obtained in (3) to arrive at the MFCCs. The number of the filters in the mel-scale filter bank dictates the number of MFCCs.” Further, MFCCs were well-known to capture speech characteristics, for example formants. EX1038-Neuhauser, 7:56-61 (“MFCCs ... represent the formant peaks of the spectrum”).

171. As also explained above in Claims 1[a]-1[b], the extracted features from the input audio data are prominent speech content and speech components (pitch, formant structure, and spectrum-related information) from the originally captured speech, because *Dantrey’s* noise subtractor neural network removes noise when creating the noise reduced spectrogram based on the MFCCs. *See* Claims 1[a]-1[b]. Additionally, Strake's speech components (which include pitch, formant structure, and spectral information) are referred to as components of the speech

which are restored. EX1005-Strake, [031]. Accordingly, a PHOSITA would have understood that such components/characteristics are maintained, in order to be subsequently restored.

2. Claim 9[b]: encode the extracted features into one or more of the low-dimensional representations of the input speech frames using a dimensionality reduction technique.

172. *Dantrey-Strake-Chen* renders obvious *encod[ing] the extracted features (e.g., Dantrey's extracted features) into one or more of the low-dimensional representations of the input speech frames (e.g., Dantrey's reduced noise spectrograms that have Chen's PCA dimensionality reduction technique applied to them) using a dimensionality reduction technique (e.g., Chen's PCA technique), as claimed.*

173. As explained above, in the *Dantrey-Strake-Chen* system, a noise reduced spectrogram is produced via the neural network “noise subtractor” neural network based on the features extracted via MFCC. Dantrey, 4:41-5:7, Fig. 1; *see* Claims 1[a]-1[b], 9[a]. *Dantrey's* neural network noise subtractor 116, which is the first stage in *Dantrey's* pipeline (Fig. 1), ultimately provides a spectrogram, “having background noise [] reduced by noise subtractor 116,” to *Dantrey's* second neural network audio generator 118. EX1004-Dantrey, 5:4-7, Fig. 1. To accomplish such, *Dantrey's* noise subtractor 116 performs “feature extraction [] using Mel Frequency Cepstral Coefficients (MFCC), a delta of MFCC, and band energy stacked together.”

EX1004-Dantrey, 4:50-53. Accordingly, a PHOSITA would have understood that in the *Dantrey-Strake-Chen* combination, *Dantrey* first extracts features, for example using MFCC, and then ultimately passes a noise reduced spectrogram from *Dantrey's* first neural network to *Dantrey's* second neural network, and that such extracted features which are passed along are the non-noise features in view of *Strake's* teaching to “restore one or more speech components of the clean speech signal.” EX1005-Strake, [031]. *Chen's* ML PCA dimensionality reduction technique is then applied to such noise reduced spectrogram in the *Dantrey-Strake-Chen* combination. See Claim 1[b]; EX1006-Chen, 3:66-4:4; EX1004-Dantrey, 4:41-5:7. Both the '96 Patent and the prior art explain that PCA is a well-known dimensionality reduction technique, particularly as applied to speech spectrograms. EX1001, 7:22-27; EX1014-Zheng, 2-4; EX1031-Wells, 13:36-42 (“Time frequency matrix 204 is essentially a spectrogram. The next step reduces the spectrogram to the least number of values which best represent it. There are numerous methods for doing so, including time and frequency marginals, principal component analysis, singular value decomposition, and moments of the spectrogram in time and frequency.”); EX1015-Wang, 3 (see §4.3 (“Our experiments include baseline, LDA transformation, PCA transformation and LP transformation experiments. In baseline experiments, MFCC features are used. The dimensionality of MFCC vectors varies from 3 to 39 in each sub-experiment. In LDA, PCA and LP experiments, 39-

dimensional MFCC vectors are obtained first. Then they are transformed to feature vectors in each sub-experiment with the dimensionality corresponding to that of baseline experiments. Features and the dimensionality of feature vectors used in these experiments are shown in Table 1.”).

174. I note that such is directly in line with the '496 Patent which discloses that features are extracted via MFCC before such are encoded via a PCA dimensionality reduction technique. '496 Patent, 7:16-27.

G. Claim 16

1. 16[pre] A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:

175. In my opinion, to the extent the preamble is limiting, *Dantrey-Strake-Chen* renders obvious *[a] non-transitory computer-readable medium (e.g., Dantrey's memory) comprising instructions (e.g., Dantrey's instructions) that, when executed by at least one processor (e.g., Dantrey's processor), cause the at least one processor to: (e.g., Dantrey's processor executing the instructions). See Claim 1[Pre].*

176. As described in Claim 1, *Dantrey* discloses its invention uses a “computer system 800” which includes “memory 820” storing “instruction(s)” to be “executed by processor 802.” EX1004-Dantrey, 15:3-15, 13:52-14:2, 14:21-28, Fig.

8. Accordingly, a PHOSITA would have understood that such memory teaches a *non-transitory computer-readable medium*, as claimed.

2. 16[a] digitize analog input audio signals to generate input audio data;

177. See Claim 2[a].

3. 16[b] fragment the input audio data into a plurality of input speech frames, wherein the input audio data comprises foreground speech content, one or more non-content elements, and one or more speech characteristics;

178. See Claim 1[a].

4. 16[c] convert the input speech frames to low-dimensional representations of the input speech frames, wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data and the low-dimensional representations of the input speech frames omit one or more of the non-content elements;

179. See Claim 1[b].

5. 16[d] apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames;

180. See Claim 1[c].

6. 16[e] combine the target speech frames to generate output audio data, wherein the output audio data further comprises one or more portions of the foreground speech content and one or more of the speech characteristics; and

181. See Claim 1[d].

7. *16[ff] convert the output audio data to analog audio output signals before providing the analog audio output signals to an audio output device.*

182. See Claims 2[b], 2[c].

H. Claim 17: The non-transitory computer-readable medium of claim 16, wherein the non-content elements comprise one or more of background noise, microphone pops, low-fidelity audio, or audio clippings and the speech characteristics comprise one or more of pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech.

183. See Claim 16; See Claim 3.

I. Claim 20: The non-transitory computer-readable medium of claim 16, wherein the second neural network comprises one or more of a diffusion probabilistic model, a flow-based model, or a generative adversarial network-based model.

184. See Claim 16; See Claim 7.

X. OPINIONS REGARDING GROUND II: DANTREY, STRAKE, CHEN, AND QUILLEN

185. In my opinion, the additional claims in Ground 2 merely add on trivial additions regarding how the claimed neural networks are trained.

186. The additions to Claims 4-6, 11-15, and 18-19 primarily involve training each neural network with samples of the inputs/outputs that are anticipated/desired, respectively, so that such neural networks perform as trained. As a preliminary matter, it was well-known in the art, trivial, and obvious to train neural networks to subsequently operate how they are trained. That is, this concept

is obvious because it is fundamental to train neural networks to produce a desired output (from an anticipated input) by giving the network large amounts of training data samples of desired outputs (and anticipated inputs). For example, *Quillen*, as explained and mapped below, teaches such. Additionally, it has long been known that neural networks are designed to output that which they have been trained to do. EX1039-Demmin, 7:54-56 (“When designing a neural network analyzer, it is important to ensure that the training data is accurate and representative of conditions of actual use.”); EX1040-Black, 1:30-35 (“Ideally, by the end of the training process, presentation of a vector of inputs from the training data to the [artificial neural network] results in activations (outputs) at the output layer that exactly match the proper training data outputs.”).

187. And it was further known in neural network pipelined systems to train a subsequent network with a previous training output. EX1041-Guo, [0142] (“The same training data is used to generate each computer model”); EX1042-Okamoto, 3 (see §4.1 “mel-spectrograms were predicted by the seq2seq [acoustic model] with full-context label input, and the [] waveforms were synthesized by the neural vocoders trained in the AS condition with the predicted mel-spectrograms”); EX1032-Arik, 10:56-61 (“WaveNet was separately trained to be used as a vocoder treating mel-scale log-magnitude spectrograms as vocoder parameters. These

vocoder parameters are input as external conditioners to the network. The WaveNet may be trained using ground-truth mel-spectrograms and audio waveforms.”).

188. Accordingly, it was trivial and obvious to train neural networks with samples of data of the type that one wants the neural network to output based on certain inputs. In addition to the previous teachings, this is further taught by *Quillen* as explained below. Put simply, in the *Dantrey-Strake-Chen* system explained above (Ground 1), it would have been further obvious to train the neural networks to perform as intended (i.e., desired outputs based on anticipated inputs), along with a few basic implementation details.

A. Claim 4: The voice enhancement system of claim 1 wherein the one or more processors are further configured to execute the instructions to train the first neural network using input audio training data, one or more augmentations, and one or more transcripts, wherein the first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and low-dimensional representations of input audio training data speech frames.

189. In my opinion, *Dantrey-Strake-Chen* in view of *Quillen* renders obvious *the voice enhancement system of claim 1 wherein the one or more processors are further configured to execute the instructions to train the first neural network (e.g., Quillen’s neural network training process) using input audio training data (e.g., Quillen’s speech data), one or more augmentations (e.g., Quillen’s simulated noisy speech data), and one or more transcripts (e.g., Quillen’s transcript), wherein the first neural network is trained to learn a mapping between (e.g.,*

Quillen's neural network one-to-one mapping training between) *input training speech frames fragmented from the input audio training data* (e.g., training samples of *Dantrey-Strake's* frames of input speech) *and low-dimensional representations of input audio training data speech frames* (e.g., training samples of *Dantrey's* reduced noise spectrograms that have *Chen's* PCA dimensionality reduction technique applied to them), as claimed.

190. *Dantrey* discloses that its neural networks would need to be trained, and that its subtractor network would need to be trained on audio data in order to provide the de-noised feature representations. EX1004-*Dantrey*, 4:37-40 (“one or more neural networks would need to be trained on a type of this primary audio in order to be able to distinguish foreground or background noise in an audio signal.”), 4:56-58 (“a subtractor network can be trained with two outputs, including voice activity detection (VAD) and de-noised features.”). *Dantrey* does not expressly describe all of the details of how to train such neural networks. However, *Quillen* provides a well-known, exemplary way to do such training.

191. *Quillen* describes a “Method and System for Speech Enhancement” which includes “training a neural network for de-noising.” EX1007-*Quillen*, Title, Abstract. *Quillen's* training technique specifically includes *using input audio training data* (e.g., *Quillen's* speech data), *one or more augmentations* (e.g., *Quillen's* simulated noisy speech data), *and one or more transcripts* (e.g., *Quillen's*

transcript), as explained below. And *Quillen* explains that its “method 330 may train any neural network known in the art.” EX1007-Quillen, [0032].

192. Regarding *using input audio training data*, *Quillen* explains that its training “method begins by creating simulated noisy speech data from high quality *speech data*” or simply “clean speech.” EX1007-Quillen, [0002], [0017], [0028], Figs. 1, 3.

193. Regarding *using one or more augmentations*, the ’496 Patent explains that background noise is an example of an augmentation. EX1001, 6:29 (“augmentations 604 may include background noise”), Claim 5 (*augmentations ... comprise one or more of background noise*).

194. Accordingly, *Quillen’s* use of “*simulated noisy speech data*” in its training teaches *using one or more augmentations*. Ex1007-Quillen, [0002], Figs. 1, 3, Abstract, [0005], [0017], [0026], [0028], [0033]. *Quillen’s* noisy speech data includes background/environmental noise. EX1007-Quillen, [0023] (“embodiments are capable of removing a variety of different types of noise. For instance, embodiments can strip background speakers out of speech corrupted by multiple speakers, suppress complicated non-stationary noises, and remove reverberation, amongst other examples.”), [0039] (“environmental noise”).

195. Regarding *using one or more transcripts*, *Quillen* explains that its training technique includes “comparing (1) speech recognition results ... and (2) *a*

transcript of at least a portion of the high quality speech data upon which the at least a portion of the simulated noisy speech data was created.” EX1007-Quillen, [0004], [0030] (“relative to the true transcript of c, the original samples of clean speech”), [0036]-[0037], [0044].

196. Regarding the claimed *learn[ed] mapping*, *Quillen* specifically discloses “training ... *one-to-one mapping*,” for example via “Deep Normalizing Flow (DNF) training” where “DNF technology is a machine learning technique for training neural networks that carry out invertible *mappings of data*. In particular, a network is used to calculate an invertible functional mapping[.]” EX1007-Quillen, [0024], [0028] (“The neural network *learns* a maximum-likely encryption of the clean data c, *mapping* it to uncorrelated noise, conditioned on the noisy data from N. ... it can be used to *map* from the noisy condition information to a prediction of clean speech or spectral features”), [0034]-[0035] (“performing the training 332, e.g., deep normalizing flow training, *trains the neural network to determine an invertible one-to-one mapping* of high quality (clean) speech to noise”).

197. For the claimed *first neural network* in the *Dantrey-Strake-Chen* system described above (Claim 1), such one-to-one learned mapping is between samples of training samples of *Dantrey-Strake*’s frames of input speech and training samples of *Dantrey*’s reduced noise spectrograms that have *Chen*’s PCA

dimensionality reduction technique applied to them. As mentioned above, this is a fundamental, and obvious, concept in neural network training. *See supra* ¶¶178-181.

a) Motivation to Combine

198. A PHOSITA would have been motivated to incorporate *Quillen's* training techniques into the *Dantrey-Strake-Chen* system with a reasonable expectation of success. As a preliminary matter, it would have been obvious to train a neural network with examples/samples of how one wants the neural network to perform because “it is important to ensure that the training data is accurate and representative of conditions of actual use.” EX1039-Demmin, 7:54-56; *see supra* ¶178. In other words, a PHOSITA would have understood that neural networks perform how they are trained, and that is why it is important to ensure that the training data is as close to expected operating conditions as possible.

199. Further, *Dantrey* discloses that its neural networks need to be trained. EX1004-Dantrey, 4:37-40, 4:56-58. And *Quillen* discloses its training “embodiments can be used to directly enhance noisy audio recordings, resulting in clear, natural-sounding speech.” EX1007-Quillen, [0016]. Accordingly, a PHOSITA would have been motivated by these express disclosures to incorporate *Quillen's* training technique to train the *Dantrey-Strake-Chen* system.

200. Additionally, such a combination would have constituted combining prior art elements (a neural network and the training for such neural network)

according to known methods (using samples of input/output data to learn mappings with clean speech, augmentations, and transcripts) to yield predictable results (the trained neural network performs as trained). Again, it is fundamental and a predictable result that neural networks perform to match the learned mappings of their training data samples.

201. A PHOSITA would have had a reasonable expectation of success in making the proposed combination at least because it was well-known how to train a neural network based on how it is desired to perform with clean speech, augmentations, and transcripts, and *Quillen* explains its techniques “may train any neural network known in the art.” EX1007-Quillen, [0032]. A PHOSITA would have understood how to implement the training methods of *Quillen* to accomplish the functionality of the *Dantrey* system.

B. Claim 5: The voice enhancement system of claim 4, wherein the augmentations simulate one or more degraded speech characteristics and comprise one or more of background noise, masked data, microphone pops, smooth speech, or convolving speech.

202. In my opinion, *Dantrey-Strake-Chen-Quillen* renders obvious *the voice enhancement system of claim 4, wherein the augmentations simulate one or more degraded speech characteristics and comprise one or more of background noise, masked data, microphone pops, smooth speech, or convolving speech (e.g., Quillen’s simulated noisy speech data)*, as claimed. *See Claim 4.*

203. The '496 Patent explains that background noise is an example of an augmentation which simulates a degraded speech characteristic. EX1001, 6:29-34 (“*augmentations 604 may include background noise 620, masked data 622, microphone pops 624, smooth speech 626, and/or convolving speech 628, although other augmentations can also be used in other examples. **The augmentations in this example are included to simulate degraded speech characteristics.***”). Based on this disclosure, a PHOSITA would have understood that creating simulated noise simulates a degraded speech characteristic because noise degrades the speech signal.

204. Accordingly, *Quillen*'s use of “*simulated noisy speech data* from high quality speech data” teaches [*one or more*] *augmentations [which] simulate one or more degraded speech characteristics and comprise one or more of background noise* because *Quillen*'s noisy speech is simulated and includes background noise. EX1007-*Quillen*, Fig. 3, [0023] (“embodiments are capable of removing a variety of different types of noise. For instance, embodiments can strip background speakers out of speech corrupted by multiple speakers, suppress complicated non-stationary noise, and remove reverberation, amongst other examples.”), [0039] (“environmental noise”).

C. Claim 6: The voice enhancement system of claim 4, wherein the one or more processors are further configured to execute the instructions to train the second neural network using a target speech sample and the low-dimensional representation of input audio training data speech frames, wherein the second neural network is trained to use dynamic conversion to learn a mapping between each of the low-dimensional representation of input audio training data speech frames and a corresponding one of a plurality of target training speech frames.

205. In my opinion, *Dantrey-Strake-Chen-Quillen* renders obvious *the voice enhancement system of claim 4, wherein the one or more processors are further configured to execute the instructions to train the second neural network using a target speech sample (e.g., training samples of Strake's time domain frames) and the low-dimensional representation of input audio training data speech frames (e.g., training samples of Dantrey's reduced noise spectrograms that have Chen's PCA dimensionality reduction technique applied to them), wherein the second neural network is trained to use dynamic conversion (e.g., Quillen's real-time application) to learn a mapping between each of (e.g., Quillen's neural network one-to-one mapping training between) the low-dimensional representation of input audio training data speech frames (e.g., training samples of Dantrey's reduced noise spectrograms that have Chen's PCA dimensionality reduction technique applied to them) and a corresponding one of a plurality of target training speech frames (e.g., training samples of Strake's time domain frames), as claimed.*

206. As explained in the Paragraphs 195-197 and in Claim 4, it would have been obvious to train a neural network to learn mappings between samples of

anticipated inputs and desired outputs. The same reasoning and disclosures are incorporated here by reference. *See* Claim 4, Section VI.A.

207. Regarding *dynamic conversion*, the '496 Patent explains that the claimed *second neural network* can be “trained to convert ... in real-time, which may be achieved using dynamic conversion.” EX1001, 8:15-20. A PHOSITA would have understood that “dynamic conversion” as claimed in Claim 6 of the '496 Patent and in light of the specification is directed towards a real time application. For example, Claim 6 states "wherein the second neural network is trained to use dynamic conversion to learn a mapping...[.]" EX1001, Claim 6. Absent contradictory language from the specification, this refers to training a neural network to convert data dynamically, which a PHOSITA would have understood to be in real time. Indeed, the specification states "In some examples, the second neural network 210 is trained to convert each of the low-dimensional representation of input audio training data speech frames 634(1)-634(n) with the respective corresponding one of the target training speech frames 712(1)-712(n) in real-time, which may be achieved using dynamic conversion. Dynamic conversion may allow for the efficient processing of the input audio data 402, ensure that the resulting target speech of the output audio data 406 may contain the desired speech characteristics, and enable real-time voice enhancement without the need for a separate conversion step." EX1001, 8:14-25. Based on this disclosure a PHOSITA would have understood that

the neural network is trained to convert representations dynamically, or in other words, real-time.

208. Accordingly, *Quillen* teaches *dynamic conversion* per the '496 Patent because *Quillen* discloses that its training “[e]mbodiments can run in *real-time* with low latency.” EX1007-*Quillen*, [0016], [0023] (“Embodiments provide high-performance, low-latency, audio enhancement and can operate *faster than real-time*.”).

D. Claim 11

1. Claim 11[Pre]: A method for real-time voice enhancement, the method implemented by a voice enhancement system and comprising:

209. In my opinion, to the extent the preamble is limiting, *Dantrey* teaches or suggests a *method* (e.g., *Dantrey*'s method to reduce noise and enhance received audio), as claimed.

210. *Dantrey* teaches a “method” and “techniques [] to reduce noise in audio.” EX1004-*Dantrey*, Abstract, Claim 13. *See also* Claim 1[Pre].

2. Claim 11[a]: training a first neural network using input audio training data, one or more augmentations, and one or more transcripts and a second neural network using a target speech sample and a plurality of low-dimensional representation of input audio training data speech frames,

211. *See* Claims 4 and 6.

3. Claim 11[b]: applying the trained first neural network to convert input speech frames fragmented from input audio data to low-dimensional representations of the input speech frames, wherein the low-dimensional representations of the input speech frames omit one or more non-content elements of the input audio data;

212. See Claim 1[b].

4. Claim 11[c]: applying the trained second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; and

213. See Claim 1[c].

5. Claim 11[d]: combining the target speech frames to generate output audio data, wherein the output audio data further comprises one or more portions of foreground speech content of the input audio data and one or more speech characteristics of the input audio data.

214. See Claim 1[d].

E. Claim 12: The method of claim 11, wherein the trained first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and the low-dimensional representation of input audio training data speech frames

215. See Claim 11; See Claim 4.

F. Claim 13: The method of claim 11, wherein the augmentations simulate one or more degraded speech characteristics and comprise one or more of background noise, masked data, microphone pops, smooth speech, or convolving speech.

216. See Claim 11; See Claim 5.

G. Claim 14: *The method of claim 11, further comprising pre-processing the input audio data by applying one or more of a noise reduction algorithm to remove background noise from the input audio data or a filtering technique to remove high-frequency noise or pops from the input audio data.*

217. See Claim 11; See Claim 8.

H. Claim 15

1. Claim 15[a]: *The method of claim 11, further comprising: extracting one or more features from the input audio data, wherein the features comprise one or more of pitch, intonation, or formants; and*

218. See Claim 11; See Claim 9[a].

2. Claim 15[b]: *encoding the extracted features into one or more of the low-dimensional representations of the input speech frames using a dimensionality reduction technique.*

219. See Claim 11; See Claim 9[b].

I. Claim 18: *The non-transitory computer-readable medium of claim 16, wherein the instructions, when executed by the at least one processor further causes the at least one processor to train the first neural network using input audio training data, one or more augmentations, and one or more transcripts, wherein the first neural network is trained to learn a mapping between input training speech frames fragmented from the input audio training data and low-dimensional representation of input audio training data speech frames.*

220. See Claim 16; See Claim 4.

J. Claim 19: The non-transitory computer-readable medium of claim 18, wherein the instructions, when executed by the at least one processor further causes the at least one processor to train the second neural network using a target speech sample and the low-dimensional representation of input audio training data speech frames, wherein the second neural network is trained to use dynamic conversion to learn a mapping between each of the low-dimensional representation of input audio training data speech frames and a corresponding one of a plurality of target training speech frames.

221. See Claim 16; See Claim 6.

XI. OPINIONS REGARDING GROUND III: DANTREY, STRAKE, CHEN, AND HECKMANN

A. Claim 10: The voice enhancement system of claim 9, wherein the one or more processors are further configured to execute the instructions to extract the features using a hierarchical feature extraction network comprises a plurality of levels, wherein each of the levels is configured to capture a different one or more of the features and the captured different one or more of the features are compressed at each of the levels.

222. *Dantrey-Strake-Chen* in further view of *Heckmann* renders obvious the voice enhancement system of claim 9, wherein the one or more processors are further configured to execute the instructions to extract the features using a hierarchical feature extraction network comprises a plurality of levels (e.g., *Heckmann*'s hierarchical feature extraction framework with a first and a second layer), wherein each of the levels is configured to capture a different one or more of the features (e.g., *Heckmann*'s first layer that captures and compresses simple features, i.e., pitch, and the second layer that captures and compresses complex features, i.e., formants), and the captured different one or more of the features are

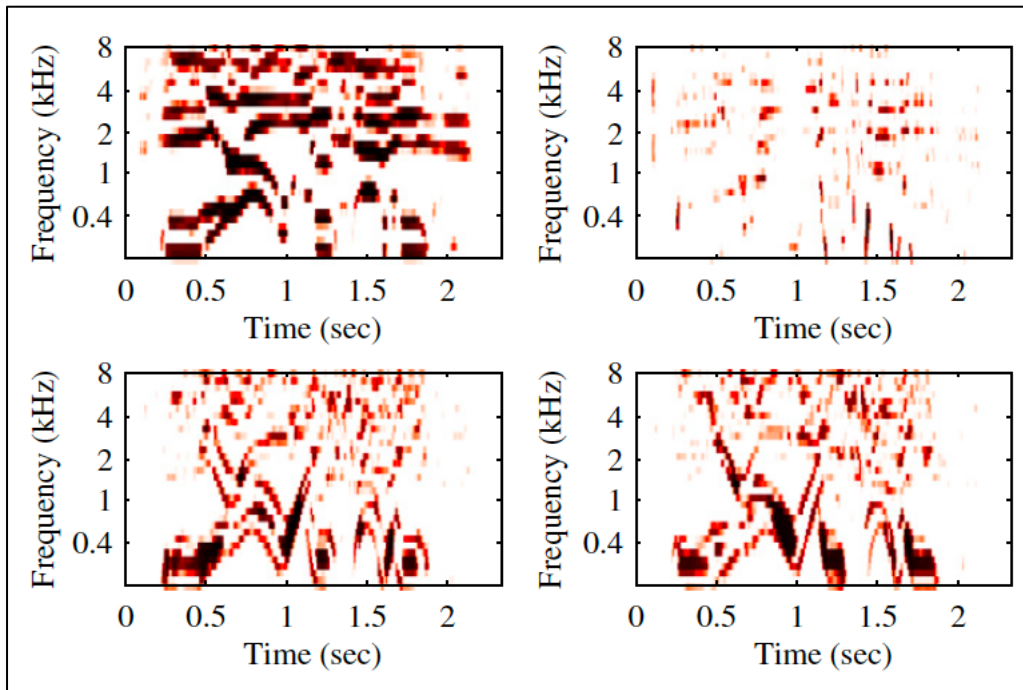
compressed at each of the levels (e.g., Heckmann's compressing simple features with a non-linear transformation and compressing complex features using NNSC), as claimed.

223. The '496 Patent explains “[i]n some examples, the low-dimensional input audio data representation 404 of the input speech may be achieved by using a hierarchical feature extraction network that extracts multiple levels of features from the input audio data 402. Each level of the network could be designed to capture different aspects of the input audio data 402, such as frequency content, *temporal dynamics*, and/or *speech characteristics*, for example. At each level of the hierarchical feature extraction network, the extracted features could be compressed into a low-dimensional input audio data representation 404 using a compression algorithm such as principal component analysis (PCA) or non-negative matrix factorization (NMF), for example.” EX1001, 7:32-44. And from Claim 9, “the features comprise one or more of *pitch*, intonation, or *formants*.”

224. Heckmann teaches using a hierarchical feature extraction framework (*hierarchical feature extraction network*) for feature extraction with a first layer (*level*) to extract simple features and a second layer (*level*) to extract complex features. EX1008-Heckmann, 4.

225. Heckmann teaches that the first layer of the extraction framework (for extracting simple features) captures simple features (*one or more features*). EX1008-

Heckmann, 4. Simple features include those that can be extracted from a single frame of sound data, such as a pitch. Heckmann depicts simple features extracted by the first layer in Fig. 6:



EX1008-Heckmann, Fig. 6. This figure depicts the features as detected by the network as frequency at a given time. Further, a PHOSITA would have understood pitch to be the perceived effect of frequency. For example, *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. Because the features are the frequencies of the speech, the features of the first layer captures pitch. *Heckmann* compresses these features using a non-linear transformation:

$$s_i^{(1)}(t, f) = H(r_i^{(1)}(t, f) - \vartheta^{(1)}),$$

EX1008-Heckmann, equ. 9. This “*non-linear compression*” uses a value of $\vartheta^{(1)} = 0.25$.” EX1008-Heckmann, 9. Equ. 9 employs the “Heaviside step function” as $H(x)$. EX1008-Heckmann, 7. A PHOSITA would have understood that the Heaviside step function is used to turn particular values on or off by outputting 0 or 1. Specifically, a negative value outputs 0, while a positive value outputs 1. Thus, when the function outputs 0, it has the effect of suppressing that term, thereby compressing the feature. Equation 9 thus shows that some features (t, f) which do not exceed a threshold are reduced to zero, i.e. removed from calculation. A PHOSITA would have understood that this non-linear transformation compresses the simple features captured by the first layer (*compression at the first level*) of the hierarchical extraction framework.

226. The output of the first layer is provided to the second layer of the hierarchical feature extraction framework, which utilizes the result to extract complex features (*capture a different one or more of the features*). EX1008-Heckmann, 4, 8. Simple features include those “covering small regions in the

spectro-temporal domain.” EX1008-Heckmann, 6. Complex features include those spanning larger time and frequency regions. EX1008-Heckmann, 4. *Heckmann* teaches that the features extracted by the second layer “represent ***complete formant configurations*** and model non-stationary patterns.” EX1008-Heckmann, 16. These complex features (*formants*) are different (*different one or more [] features*) than the simple features (e.g. pitch) extracted in the first layer.

227. *Heckmann* teaches that the second layer uses “Non-Negative Sparse Coding (NNSC)” to learn combination patterns. EX1008-Heckmann, 8. *Heckmann* teaches that NNSC is similar in function to NMF but learns complex features in the data:

NNSC differs from Non-negative Matrix Factorization (NMF) by the presence, in the cost function (12), of a sparsity enforcing term which aims at limiting the number of non-zero coefficients required for the reconstruction. Consequently, if a feature appears often in the data, it will be learned, even if it can be obtained by a combination of two or more other features. Therefore, the NNSC is expected to learn complex and global features appearing in the data.

EX1008-Heckmann, 8. A PHOSITA would have understood that NNSC is a compression technique like NMF, specifically referenced as a compression algorithm in the '496 Patent. EX1001, 7:42-44 (“a compression algorithm such as

principal component analysis (PCA) or *non-negative matrix factorization (NMF)*, for example.”). For example, *Hoyer* explains, in similar fashion to *Heckmann*, that NNSC is a combination of NMF and sparse coding. EX1047-Hoyer, 7 (“we have defined non-negative sparse coding as a combination of sparse coding with the constraints of non-negative matrix factorization.”). NMF and similar algorithms were commonly used in the art in hierarchical feature extraction. For example, *Song* uses nsNMF, another similar algorithm, in its hierarchical feature extraction. EX1023-Song, 2. Thus, the complex features captured by the second layer are compressed. Further, *Heckmann* teaches applying PCA to all of the features “to reduce the dimensionality of the features” for successful integration with an HMM. EX1008-Heckmann, 12. A PHOSITA would have understood PCA is applied to the simple and complex features to encode the features into a low-dimensional representation. *Chen* teaches ML models employing dimensionality reduction approaches such as PCA. EX1006-Chen, Title, 3:66-4:4; *see* Claims 1[b], 9. A PHOSITA would have understood that the features in the resulting combination are encoded using PCA to produce the noise-reduced spectrogram via the “noise subtractor” neural network of *Dantrey*. *See* Claim 9.

a) Motivation to Combine

228. A PHOSITA would have been motivated to incorporate *Heckman's hierarchical* feature extraction and compression functionality into the *Dantrey-Strake-Chen* system with a reasonable expectation of success.

229. For example, *Heckmann* teaches “a hierarchical framework for the extraction of spectro-temporal acoustic features. The design of the features ***targets higher robustness in dynamic environments.***” EX1008-Heckmann, Abstract. A PHOSTA would have been motivated to implement robust features, particularly in a dynamic environment where de-noising is required. A PHOSITA would have understood that a dynamic environment is one that contains various background noise along with foreground speech, and would have been motivated to use a feature extraction method that produces robust features in this type of environment.

230. Second, such a combination would have constituted applying a known technique (e.g., *Heckman's hierarchical* feature extraction) to a known device (e.g., *Dantrey's* system) ready for improvement to yield predictable results (e.g., providing feature extraction for audio processing). *KSR*, 550 U.S. 398 at 417-418. A PHOSITA would have recognized that implementing the hierarchical feature extraction taught by *Heckmann* into *Dantrey's* system increases the performance of the acoustic interference suppression circuit by extracting both complex and simple features, allowing the system to retain those features through de-noising. *Heckmann* recognizes that the by capturing both simple and complex features, recognition

performance is improved. EX1008-Heckmann, 16. For example, the second layer for feature extraction “represent[s] complete formant configurations and model non-stationary patterns” and “improves the recognition performance on clean data and in noise.” EX1008-Heckmann, 16. A PHOSITA would have understood that capturing these features would lead to similarly improved performance in noise removal. Further, the hierarchical feature extraction framework provides the additional features without greatly increasing computational complexity, as *Heckmann’s* extraction “for 1 s of speech lasts approximately 280 ms, i.e. 3.5 times faster than real time.” EX1008-Heckmann, 16. Because the hierarchical feature extraction structure determines the simple features, such as pitch, and then feeds these into the layers determining the complex features, this method is computationally efficient.

231. Finally, a PHOSITA would have had a reasonable expectation of success in implementing the hierarchical feature extraction of *Heckmann* would only involve minor software changes to the system of *Dantrey-Strake-Chen*. A PHOSITA would have understood how to implement different methods of feature extraction for use in a neural network system, including hierarchical feature extraction. Hierarchical feature extraction was well known in the art at the time of the ’496 Patent’s priority date. *See* Section V.B.2.

XII. OPINIONS REGARDING GROUNDS IV-VI: *DANTREY, STRAKE, LIU (GROUND 4), AND ADDITIONALLY QUILLEN (GROUND 5), AND ADDITIONALLY HECKMANN (GROUND 6)*

232. I first provide an overview of Grounds 4-6 because such grounds use much on the same analysis from Grounds 1-3. The only differences between Grounds 1-3 and Grounds 4-6 is that: Grounds 1-3 rely on Chen’s teachings to demonstrate unpatentability of the claimed *low-dimensional representation* limitations under Petitioner’s proposed construction; and Grounds 4-6 do not so rely on Chen, but instead rely on Liu’s teachings to demonstrate the unpatentability of the claimed *low-dimensional representation* without Petitioner’s proposed construction. For this reason, I incorporate the analysis for all claim limitations in Grounds 1-3 (except Chen’s teachings for *low-dimensional representation* under the proposed construction) by reference here. Accordingly, I only address the *low-dimensional representation* limitations here and my opinion as to why a PHOSITA would have been motivated to incorporate Liu’s teachings into the combinations described above (absent Chen).

A. Liu’s Teachings

233. Liu teaches a two neural network pipeline for speech restoration/noise removal which expressly uses a “low dimensional mel spectrogram as the intermediate-level feature” (i.e., the output of a first neural network and the input of a second neural network. EX1043-Liu, Abstract, §§ 1, 3.1-3.2. This can be seen in

Liu's Figure 1 which shows an "analysis stage" and a "synthesis stage" and the intermediate representation as the low dimensional mel spectrogram:

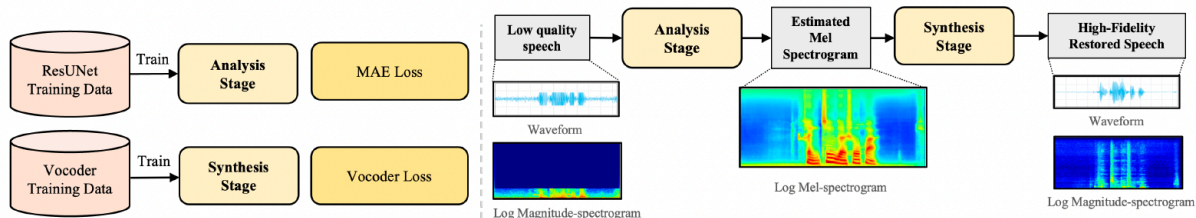


Figure 1: Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.

EX1043-Liu, Fig. 1.

234. Liu is clear that its "analysis stage" is "ResUNet." EX1043-Liu, § 3.1. ResUNet was a known neural network before the '96 Patent. EX1047-Waldner, Abstract. Also Liu's explanation of ResUNet's training data and loss function are typical features of neural networks. EX1043-Liu, Fig. 1, §§ 3.1, 4.

235. It is also clear that Liu's "synthesis stage" is "a neural vocoder" which is also a neural network because its name states so, it has training data, and it has a loss function. EX1043-Liu, Fig. 1, §§ 3.2, 4.

236. Liu teaches that its first neural network (ResUNet/analysis stage) takes in distorted speech waveforms and creates "low dimensional mel spectrogram[s]." EX1043-Liu, Fig. 1, Abstract, §§ 1, 3.1. And Liu then teaches that its second neural network receives these "low dimensional mel spectrogram[s]" and outputs clean speech waveforms. EX1043-Liu, Fig. 1, Abstract, §§ 1, 3.2.

B. Specific Claim Limitation Applications for “low-dimensional representation”

237. In this Section, I explain how Grounds 4-6 teach the specific *low-dimensional representation* limitations. As previously mentioned, I do not repeat analysis for limitations from Grounds 1-3 which do not recite *low-dimensional representations* (or where a specific ground/limitation above simply cites to another substantively similar limitation). Instead, I focus my analysis here on how Liu teaches the express claim language *low-dimensional representation*, and why a PHOSITA would have been motivated to incorporate such into the combinations presented above with a reasonable expectation of success.

1. Ground 4 (Dantrey-Strake-Liu)

a) Claim 1[b]

238. In my opinion, *Dantrey-Strake* in view of *Liu* renders obvious *conver[ting] the input speech frames* (e.g., *Dantrey-Strake*’s frames of input speech) *to low-dimensional representations of the input speech frames* (e.g., Liu’s low dimensional mel spectrograms), *wherein one or more of the fragmentation or the conversion is based on an application of a first neural network to the input audio data* (e.g. *Liu*’s low dimensional mel spectrograms are created using a first neural network such as *Dantrey*’s noise subtractor or *Liu*’s ResUNet) *and the low-dimensional representations of the input speech frames omit one or more of the non-*

content elements (e.g. *Liu's* low dimensional mel spectrograms omit background noise), as claimed.

239. As I explained above in Ground 1, in the *Dantrey-Strake* combination, frames of input speech are converted into noise reduced spectrograms via *Dantrey's* noise subtractor neural network.

240. The difference presented here in these Grounds is that *Dantrey's* noise subtractor neural network produces “low dimensional mel spectrogram” (as taught by *Liu*) as opposed to just “spectrograms” (as taught by *Dantrey*). EX1043-*Liu*, Abstract, Fig. 1, §§ 1, 3.1. In my opinion, *Liu's* teachings of a “low dimensional mel spectrogram” teaches the claimed *low-dimensional representation* because a mel spectrogram is a representation and because both are expressly “low dimensional.” EX1043-*Liu*, Abstract, Fig. 1, §§ 1, 3.1.

241. In my opinion, a PHOSITA would have been motivated to incorporate *Liu's* low dimensional mel spectrogram into the *Dantrey-Strake* system with a reasonable expectation of success for multiple reasons.

242. First, I note that *Liu's* express disclosures would have motivated such. For example, *Liu* expressly states that “[c]ompared to the conventional speech restoration methods that operate on spectrogram or waveform, VoiceFixer [*Liu*] uses the low dimensional mel spectrogram as the intermediate-level feature, which alleviates the difficulties of restoring multiple distortions simultaneously.” *Liu*, § 1.

Accordingly, a PHOSITA would have been motivated to make the proposed combination to remove multiple types of distortions simultaneously, which is one of the express aims of Liu et al.'s work on "VoiceFixer." Liu, Abstract, § 1. In other words, Liu states that using its "low dimensional mel spectrogram" would allow a noise removal system to not only remove noise but at the same time other unfavorable distortions. Liu, Abstract, § 1.

243. Second, a PHOSITA would have also understood that using a mel spectrogram with low dimensionality would have been advantageous because mel spectrograms represent "the logarithmic sensitivity to the frequency perception of audio signals which are based on the system of human hearing, having an overwhelming advantage in emphasizing audio details." EX1044-Sheng, 2. In other words, as the prior art explains, a mel spectrogram is lower dimensional than a spectrogram (*see id.* and also EX1043-Liu) and a mel spectrogram is specifically focused on human hearing (capturing the important feature information that would be particularly useful for humans to hear). Further, using a "low dimensional" intermediate representation would have motivated the combination because such were known to reduce computational complexity and improve the efficiency of neural network-based speech processing pipelines because using a vector or a representation with lower dimensions has less for a computer to process, based on my experience.

244. Third, such a combination would have constituted combining prior art elements (e.g., Dantrey's noise removal pipeline and Liu's low dimensional mel spectrogram) according to known methods (e.g., using a neural network to create a low dimensional mel spectrogram from an audio signal) to yield predictable results (e.g., lowering the dimensions of a speech representation for subsequent processing). Indeed, the predictable result of using a mel spectrogram over a non-mel spectrogram is lower dimensions (and less computational overhead) as well as focus on the human aspect of hearing. EX1044-Sheng, 2; EX1043-Liu, Abstract, §§ 1, 3.1-3.2.

245. In my opinion, a PHOSITA would have had a reasonable expectation of success in making the proposed combination because it was known how to use a neural network to create a low dimensional mel spectrogram (for example as taught by Liu). To be sure, Liu made its pretrained model available online (EX1043-Liu, n.1), and Liu explained out how to train the neural network to accomplish such. EX1043-Liu, § 4. Additionally, Dantrey already teaches a two neural network pipeline for noise removal just like Liu, so it would have been a simple modification to adjust Dantrey's intermediate output to be a mel spectrogram as opposed to just a spectrogram, particularly again because Liu provides the training data, loss function, etc. how to do so publicly. *Id.*

b) Claim 1[c]

246. In my opinion, *Dantrey-Strake-Liu* renders obvious *apply[ing] a second neural network (e.g., Dantrey’s audio generator) to the low-dimensional representations of the input speech frames (e.g., Liu’s low dimensional mel spectrograms) to generate target speech frames (e.g., Strake’s time domain frames),* as claimed.

247. As I previously explained, Liu teaches outputting a “low dimensional mel spectrogram” from its first neural network and providing such to a second neural network to ultimately obtain a clean speech waveform. Liu, Fig. 1, Abstract, §§ 1, 3.1-3.2.

248. Accordingly, the *Dantrey-Strake-Liu* combination renders Claim 1[c] obvious for the same reasons explained in Ground 1, Claim 1[c].

c) Claim 9[b]

249. In my opinion, *Dantrey-Strake-Liu* renders obvious *encod[ing] the extracted features (e.g., Dantrey’s extracted features) into one or more of the low-dimensional representations of the input speech frames (e.g., Liu’s low dimensional mel spectrogram) using a dimensionality reduction technique (e.g., Liu’s creation of a low dimensional mel spectrogram),* as claimed.

250. As explained above for Claim 9 in Ground 1, it would have been obvious to compress Dantrey’s noise reduced spectrogram (based on the Dantrey’s

extracted features) with a well-known technique such as PCA. *Supra*, Ground 1, Claim 9.

251. In these Grounds (e.g., Grounds 4-6), it would have also been obvious to use Dantrey's noise subtractor neural network to create a "low dimensional mel spectrogram" as taught by Liu (as opposed to just Dantrey's "spectrogram" which is then subsequently compressed). Liu, Abstract, Fig. 1, §§ 1, 3.1 ("working on the mel-scale can reduce the dimension of feature space and offer a more tractable restoration process."), as I just explained.

252. Therefore, the Liu teachings I explained for Claim 1 in Ground 4 also teach Claim 9[b] here.

2. Ground 5 (Dantrey-Strake-Liu-Quillen)

a) Claim 4

253. In my opinion, *Dantrey-Strake-Liu* in view of *Quillen* renders obvious *the voice enhancement system of claim 1 wherein the one or more processors are further configured to execute the instructions to train the first neural network (e.g., Quillen's neural network training process) using input audio training data (e.g., Quillen's speech data), one or more augmentations (e.g., Quillen's simulated noisy speech data), and one or more transcripts (e.g., Quillen's transcript), wherein the first neural network is trained to learn a mapping between (e.g., Quillen's neural network one-to-one mapping training between) input training speech frames*

fragmented from the input audio training data (e.g., training samples of Dantrey-Strake's frames of input speech) and low-dimensional representations of input audio training data speech frames (e.g., training samples of Liu's low dimensional mel spectrograms), as claimed.

254. In addition to the analysis provided with respect to Claim 4 in Ground 2, Liu further teaches that neural networks (such as those used in Liu's pipeline) "are usually trained on large-scale speech datasets," provides its "pre-trained model" (Liu, n.1), and describes how its training data was obtained (Liu, § 4), thus further establishing the obviousness of the claimed "training" limitations presented in Ground 2.

b) Claim 6

255. In my opinion, *Dantrey-Strake-Chen-Quillen* renders obvious *the voice enhancement system of claim 4, wherein the one or more processors are further configured to execute the instructions to train the second neural network using a target speech sample (e.g., training samples of Strake's time domain frames) and the low-dimensional representation of input audio training data speech frames (e.g., training samples of Liu's low dimensional mel spectrograms), wherein the second neural network is trained to use dynamic conversion (e.g., Quillen's real-time application) to learn a mapping between each of (e.g., Quillen's neural network one-to-one mapping training between) the low-dimensional representation of input audio*

training data speech frames (e.g., training samples of *Liu*'s low dimensional mel spectrogram) and a corresponding one of a plurality of target training speech frames (e.g., training samples of *Strake*'s time domain frames), as claimed. See Ground 2 Claim 6 and see Ground 5 Claim 4.

3. Ground 6 (Dantrey-Strake-Liu-Heckmann)

a) Claim 10

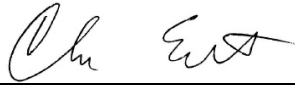
256. Claim 10 does not include any additional references to *low-dimensional representation* other than just as a function of its dependencies. Accordingly, I do not present any additional analysis for Ground 6. Ground 4 (Claims 1 and 9) and Ground 3 (Claim 10) are incorporated by reference here for this claim.

XIII. CONCLUSION

257. 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 25, 2026

By: 

Christopher Schmandt

Christopher Schmandt

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

Education

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

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

Professional Experience -- MIT

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

Architecture Machine Group, Research Associate, 1980-1984

Architecture Machine Group, Research Assistant, 1979-1980

Architecture Machine Group, Graphics Programmer, 1977-1979

Departmental Undergraduate Research Opportunities Program Coordinator, 1984-2018

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

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

Sponsored Research Activities

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

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

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

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

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

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

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

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

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

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

Professional experience -- intellectual property

IPR2026-00273

Krisp Technologies EX1003 Page 134

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 Lakshmiathy 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 nomadic 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.