

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

**PETITION FOR *INTER PARTES* REVIEW  
OF U.S. PATENT NO. 12,125,496**

**TABLE OF CONTENTS**

**I. INTRODUCTION.....1**

**II. SUMMARY OF THE '496 PATENT.....1**

A. DESCRIPTION OF THE '496 PATENT ..... 1

B. SUMMARY OF THE PROSECUTION HISTORY .....2

C. LEVEL OF SKILL OF A PERSON HAVING ORDINARY SKILL IN THE ART.....3

**III. REQUIREMENTS UNDER 37 C.F.R. § 42.104 .....3**

A. GROUNDS FOR STANDING UNDER 37 C.F.R. § 42.104(A).....3

B. IDENTIFICATION OF CHALLENGE UNDER 37 C.F.R. § 42.104(B) .....4

C. CLAIM CONSTRUCTION UNDER 37 C.F.R. § 42.104(B)(3).....4

    1. “*low-dimensional representation*” .....5

**IV. SHOWING OF ANALOGOUS, PRIOR ART .....6**

A. DANTREY .....7

B. STRAKE .....9

C. CHEN .....9

D. QUILLEN .....10

E. HECKMANN.....10

F. LIU.....11

**V. GROUND 1: CLAIMS 1-3, 7-9, 16-17, AND 20 ARE OBVIOUS OVER DANTREY, STRAKE, AND CHEN.....12**

A. GROUND OVERVIEW .....12

B. CLAIM 1 .....13

    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:.....13*

    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; .....14*

a)	Input audio data comprising foreground speech content, non-content element[s], and speech characteristic[s] .....	14
b)	Motivation to Combine .....	16
c)	Motivation to Combine .....	18
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>	20
a)	Motivation to Combine .....	24
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>	25
a)	Motivation to Combine .....	27
5.	<i>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. ....</i>	29
a)	Motivation to Combine .....	31
C.	CLAIM 2 .....	32
1.	<i>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: .....</i>	32
2.	<i>Claim 2[a]: digitize analog input audio signals obtained via the physical microphone to generate the input audio data;.....</i>	33
3.	<i>Claim 2[b]: convert the output audio data to analog audio output signals; and .....</i>	34
4.	<i>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. ....</i>	35

D.	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. ....	36
E.	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. ....	37
F.	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. ....	38
G.	CLAIM 9 .....	39
	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>	39
	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>	40
H.	CLAIM 16 .....	41
	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>	41
	2. <i>16[a] digitize analog input audio signals to generate input audio data; .....</i>	42
	3. <i>16[b] fragment the input audio data into a plurality of input speech frames, wherein the input audio data</i>	

*comprises foreground speech content, one or more non-content elements, and one or more speech characteristics; ..... 42*

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; ..... 42*

5. *16[d] apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; ..... 42*

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 ..... 43*

7. *16[f] convert the output audio data to analog audio output signals before providing the analog audio output signals to an audio output device. .... 43*

I. 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. .... 43

J. 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. .... 43

**VI. GROUND 2: CLAIMS 4-6, 11-15, AND 18-19 ARE OBVIOUS OVER DANTREY, STRAKE, CHEN, AND QUILLEN ..... 43**

A. GROUND OVERVIEW ..... 43

B. 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. ....44

1. *Motivation to Combine* ..... 48

C. 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. ....49

D. 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. ....50

E. CLAIM 11 .....52

1. *Claim 11[Pre]: A method for real-time voice enhancement, the method implemented by a voice enhancement system and comprising:* .....52

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,* .....52

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; .....52*

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.....53*

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. ....53*

F. 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 .....53

G. 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. ....53

H. 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. ....53

I. CLAIM 15 .....54

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 .....54*

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.....54*

J.	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. ....	54
K.	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. ....	54
<b>VII.</b>	<b>GROUND 3: CLAIM 10 IS OBVIOUS OVER <i>DANTREY, STRAKE, CHEN, AND HECKMANN</i></b> .....	<b>55</b>
A.	GROUND OVERVIEW .....	55
B.	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.....	55
1.	<i>Motivation to Combine</i> .....	58
<b>VIII.</b>	<b>GROUNDS 4-6: CLAIMS 1-20 ARE OBVIOUS OVER <i>DANTREY, STRAKE, AND LIU</i> (GROUND 4), AND ADDITIONALLY <i>QUILLEN</i> (GROUND 5), AND ADDITIONALLY <i>HECKMANN</i> (GROUND 6).....</b>	<b>60</b>

A.	GROUNDS OVERVIEW.....	60
B.	LIU’S TEACHINGS.....	61
C.	SPECIFIC CLAIM LIMITATION APPLICATIONS FOR “LOW-DIMENSIONAL REPRESENTATION” .....	62
1.	<i>Ground 4 (Dantrey-Strake-Liu)</i> .....	63
a)	Claim 1[b] .....	63
b)	Claim 1[c].....	66
c)	Claim 9[b] .....	66
2.	<i>Ground 5 (Dantrey-Strake-Liu-Quillen)</i> .....	67
a)	Claim 4 .....	67
b)	Claim 6 .....	68
3.	<i>Ground 6 (Dantrey-Strake-Liu-Heckmann)</i> .....	69
a)	Claim 10 .....	69
<b>IX.</b>	<b>CONCLUSION.....</b>	<b>70</b>
<b>X.</b>	<b>MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1) .....</b>	<b>71</b>
A.	REAL PARTY-IN-INTEREST .....	71
B.	RELATED MATTERS .....	71
C.	LEAD AND BACK-UP COUNSEL .....	72
D.	37 C.F.R. § 42.8(B)(4) – SERVICE INFORMATION .....	73

**TABLE OF AUTHORITIES**

CASES

*KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398 (2007) .....59  
*Phillips v. AWH Corp.*, 415 F.3d 1303 (Fed. Cir. 2005) (en banc)..... 4-5

STATUTES

35 U.S.C. § 102(a)(1) .....10, 11  
35 U.S.C. § 102(a)(2) .....8  
35 U.S.C. §§ 102(a)(1)-(2) .....9, 10

## **I. INTRODUCTION**

Petitioner, Krisp Technologies, Inc. (“Petitioner”), requests Inter-Partes Review (“IPR”) of Claims 1-20 (the “Challenged Claims”) of U.S. Patent No. 12,125,496 (“EX1001”) (the “’496 Patent”).

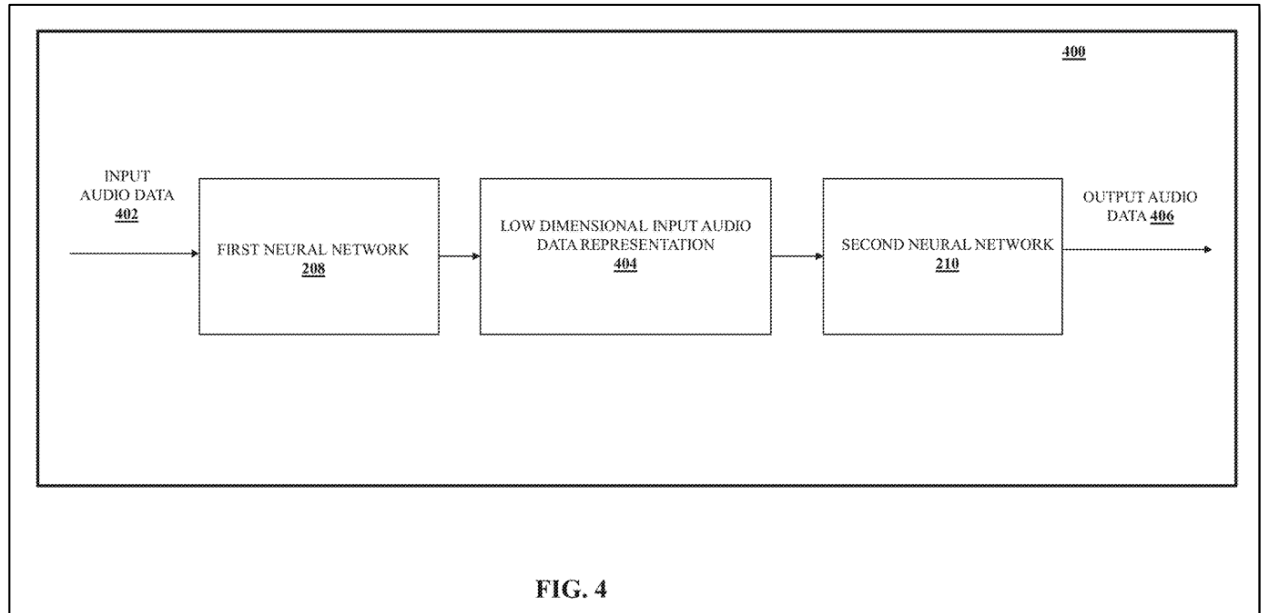
Petitioner presents two sets of Grounds which are non-redundant based on Petitioner’s proposed claim construction of the claim term “low-dimensional representation.” Petitioner proposes that “low-dimensional representation” should be construed as “a compressed representation of an input speech frame which is a result of a dimensionality reduction technique,” consistent with the specification. Grounds 1-3 demonstrate that the Challenged Claims are unpatentable under this construction, including through the application of known compression-based dimensionality reduction techniques such as PCA. Grounds 4-6 demonstrate that the Challenged Claims are unpatentable when applying the express claim language “low-dimensional representation” without such construction, as reflected in the prior art that expressly discloses representations described as “low dimensional.”

## **II. SUMMARY OF THE ’496 PATENT**

### ***A. Description of the ’496 Patent***

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.

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” to generate output audio data. EX1001, 5:56-67.

### ***B. Summary of the Prosecution History***

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 Petition and without

waiving its right to challenge priority in this or any other proceeding, Petitioner applies May 5, 2023, as the priority date for the Challenged Claims.

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.

***C. Level of Skill of a Person Having Ordinary Skill in the Art***

A PHOSITA at the time of the '496 Patent would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of audio speech signal processing, or a bachelor's degree in the same fields with at least four years of experience in the field of audio speech signal processing. Additional education or experience might substitute for the above requirements. EX1003, ¶28.

**III. REQUIREMENTS UNDER 37 C.F.R. § 42.104**

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

Petitioner certifies the '496 Patent is eligible for IPR and that the Petitioner is not barred or estopped from requesting an IPR challenging the claims of the '496 Patent.

**B. Identification of Challenge Under 37 C.F.R. § 42.104(b)**

In view of the prior art and evidence presented, the Challenged Claims of the '496 Patent are unpatentable and should be cancelled. 37 C.F.R. § 42.104(b)(1).

Further, based on the prior art references identified below, IPR of the Challenged Claims should be granted. 37 C.F.R. § 42.104(b)(2).

Ground	Claim(s)	Basis	Reference(s)
1	1-3, 7-9, 16-17, 20	§ 103	Dantrey (EX1004), Strake, (EX1005), Chen (EX1006)
2	4-6, 11-15, 18-19	§ 103	Dantrey, Strake, Chen, Quillen (EX1007)
3	10	§ 103	Dantrey, Strake, Chen, Heckmann (EX1008)
4	1-3, 7-9, 16-17, 20	§ 103	Dantrey, Strake, Liu (EX1043)
5	4-6, 11-15, 18-19	§ 103	Dantrey, Strake, Liu, Quillen
6	10	§ 103	Dantrey, Strake, Liu, Heckmann

**C. Claim Construction Under 37 C.F.R. § 42.104(b)(3)**

In this proceeding, claims are interpreted under the same standard applied by Article III courts (i.e., the Phillips standard). 37 C.F.R. § 42.100(b); *Phillips v. AWH Corp.*, 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc). For purposes of this proceeding, Petitioner only proposes a construction for one term. Grounds 1-3 demonstrate unpatentability under this construction. Grounds 4-6 demonstrate that the Challenged Claims are unpatentable even without such construction under the

express claim language, and thus regardless of whether the Board determines that an express construction is necessary.

*1. “low-dimensional representation”*

**“Low-dimensional representation”** should be construed as **“a compressed representation of an input speech frame which is a result of a dimensionality reduction technique.”** Construction here is necessary to clarify the meaning and scope of the claim term in light of the specification’s consistent description of the invention.

The specification consistently describes the claimed “low-dimensional representation” as a compressed representation of the speech signal produced through a dimensionality reduction technique. EX1001, 7:22-54. While the specification also states that the representation has “lower dimensionality than that of the input audio data” (*id.* at 4:44-46), that statement merely reflects the result of the compression process and does not itself define the term. Construing the term to mean only a representation having fewer dimensions than the input would improperly read out the ’496 Patent’s repeated description of compression through a dimensionality reduction technique. *See Phillips v. AWH Corp.*, 415 F.3d 1303, 1315–17 (Fed. Cir. 2005) (claim terms interpreted in light of the specification).

#### IV. SHOWING OF ANALOGOUS, PRIOR ART

The prior art is analogous to the claimed invention of the '496 Patent. The prior art is from the same field of endeavor of the '496 Patent, namely speech audio processing. EX1001, Abstract, 1:11-13, 2:42-45; EX1004-Dantrey, Abstract, 2:64-66, 3:60-65, 4:45-53, 5:4-10; EX1005-Strake, Title, Abstract, [02], [019]; EX1006-Chen, Abstract, 1:11-13, 1:13-17, 2:42-45, 3:7-18; EX1007-Quillen, Abstract, [0001]-[0002], [0016], [0018]; EX1008-Heckmann, Abstract, Introduction; EX1043-Liu, Title, Abstract, § 1; EX1003, ¶¶43, 54, 58, 64, 68, 73, 77, 81.

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: EX1001, 2:42-45, 2:51-56, 4:38-41, Fig. 4, Claims 1, 11, 16; EX1004-Dantrey, Abstract, 2:64-66, 4:22-27, 4:30-32; EX1043-Liu, §§ 1, 3.1-3.2, Fig. 1; EX1003, ¶¶44-50, 55, 59, 82.
- (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: 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; EX1004-Dantrey, 3:54-62, 4:20-40; EX1005-Strake, [037]-[042], EX1006-Chen,

Abstract, 2:44-57, 3:7-18, 4:38-48, 5:64-65, 6:21-64, Fig. 3; EX1043-Liu, §§ 1, 3.1-3.2, Fig. 1; EX1003, ¶¶44-50, 55, 59, 65, 69, 82.

(C) 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; EX1006-Chen, 3:66-4:4; EX1043-Liu, §§ 1, 3.1-3.2, Fig. 1; EX1003, ¶¶44-50, 55, 70, 82.

(D) 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; EX1007-Quillen, Abstract, [0003]-[0006], [0022], [0024]; EX1043-Liu, § 4; EX1003, ¶¶44-50, 55, 74, 82.

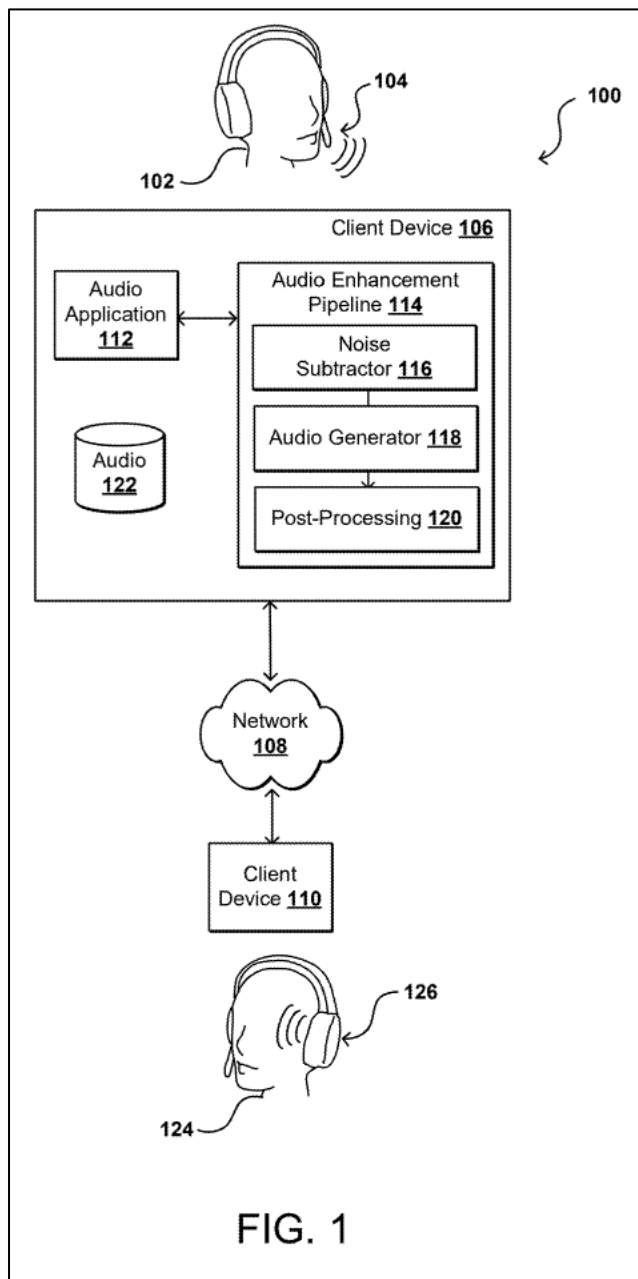
(E) Problems related to feature extraction, particularly when processing speech in noisy environments. EX1001, 1:19-26, 1:54-64, 2:9-15, 7:16-22, 7:32-39, Claims 9, 10, 15; EX1008-Heckmann, Abstract, Introduction, 6-9; EX1043-Liu, §§ 1, 3.1-3.2, Fig. 1; EX1003, ¶¶44-50, 55, 78, 82.

**A. *Dantrey***

U.S. Patent No. 12,412,590 (“*Dantrey*”) was filed December 19, 2019 and issued September 9, 2025. *Dantrey* therefore, is prior art to the ’496 Patent under at

least 35 U.S.C. § 102(a)(2). *Dantrey* was not cited or considered during prosecution of the '496 Patent. *See generally* EX1001, EX1002.

*Dantrey* teaches apparatuses, systems, and techniques for reducing noise in audio. EX1004-Dantrey, Abstract, Title. Fig. 1 illustrates a system that processes audio to remove noise. EX1004-Dantrey, 2:64-66.



EX1004-Dantrey, Fig. 1. *Dantrey* teaches digital audio (captured from a human speaker through a microphone and then digitized) is input to a first neural network noise subtractor 116 which outputs a noise reduced spectrogram to a second neural network audio generator 118, where this second neural network outputs clean speech audio. EX1004-Dantrey, 3:54-5:19, Fig. 1, claim 7.

**B. *Strake***

International Application Publication No. 2020/199990 (“*Strake*”) was filed on March 24, 2020, and published on October 8, 2020. *Strake* therefore, qualifies as prior art to the ’496 Patent under at least 35 U.S.C. §§ 102(a)(1)-(2). EX1005-*Strake*, (22), (43). *Strake* was not cited or considered during the prosecution of the ’496 Patent. *See generally* EX1001, 1002.

*Strake* teaches a “speech processing system” that employs neural networks to remove background interference. EX1005-*Strake*, [019], [029]-[030].

*Strake* teaches processing the audio data in frames throughout its system. EX1005-*Strake*, [037], [042].

**C. *Chen***

U.S. Patent No. 10,867,616 (“*Chen*”) was filed on October 10, 2019 and issued December 15, 2020. *Chen* therefore, is prior art to the ’496 Patent under at least 35 U.S.C. §§ 102(a)(1)-(2). *Chen* was not cited or considered during prosecution of the ’496 Patent. *See generally* EX1001, EX1002.

*Chen* teaches “solutions for eliminating undesired audio artifacts, such as background noises, on an audio channel.” EX1006-Chen, Abstract.

*Chen* also teaches applying the well-known PCA dimensionality reduction technique by a machine learning model. EX1006-Chen, 3:66-4:4.

***D. Quillen***

U.S. Patent Application Publication No. 2021/0241780 (“*Quillen*”) was filed on October 31, 2020 and published on August 5, 2021. *Quillen* therefore, is prior art to the ’496 Patent under at least 35 U.S.C. §§ 102(a)(1)-(2). *Quillen* was not cited or considered during prosecution of the ’496 Patent. *See generally* EX1001, EX1002.

*Quillen* teaches specific details for how to train neural networks which remove noise. EX1007-Quillen, Abstract.

***E. Heckmann***

“A Hierarchical Framework for Spectro-Temporal Feature Extraction” to Heckmann et al. (“*Heckmann*”) was publicly available at least as early as May 2011 and therefore qualifies as prior art under 35 U.S.C. § 102(a)(1). EX1049-Munford-Heckmann, ¶¶66-69 (declaration of Jane Munford establishing public availability of *Heckmann*). *Heckmann* was not cited or considered during the prosecution of the ’496 Patent. *See generally* EX1001, 1002.

*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.

**F. Liu**

“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 and therefore qualifies as prior art under 35 U.S.C. § 102(a)(1). EX1045-Ching, 1-2, 11-12, 48-54 (declaration of Mina Ching, Records Request Processor at the Internet Archive); EX1050-Munford-Liu, ¶¶10-14 (declaration of Jane Munford establishing public availability of *Liu*). *Liu* was not cited or considered during prosecution of the ’496 Patent. *See generally* EX1001, 1002.

*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:

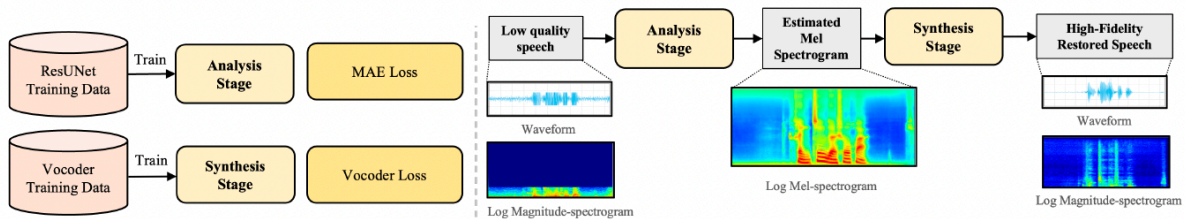


Figure 1: Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.

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

**V. GROUND 1: CLAIMS 1-3, 7-9, 16-17, AND 20 ARE OBVIOUS OVER DANTREY, STRAKE, AND CHEN**

**A. *Ground Overview***

The '496 patent—which went straight to allowance—broadly claims audio noise removal via a two neural network pipeline. This concept was well-known in the art. For example, *Dantrey* teaches a two neural network pipeline that takes in noisy audio data and outputs clean audio data, as claimed. *Dantrey* just does not expressly spell out two straight-forward and obvious concepts in such a pipeline: (1) that the data is processed in frames; and (2) that the neural network applies a compression technique.

However, these concepts were also well-known in the art. Specifically, in two additional references that remove audio noise via neural network pipelines: (1) *Strake* teaches the claimed framing functionality; and (2) *Chen* teaches the claimed neural network compression technique. As explained below, it would have been obvious to process data in frames through such a pipeline because speech signals are non-stationary but exhibit quasi-stationary behavior at shorter durations. And it would have also been obvious to compress data with the neural network because the prior art compression techniques advantageously reduce computing overhead.

**B. 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:***

To the extent the preamble is limiting, *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.<sup>1</sup>

The '496 Patent states that 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.

---

<sup>1</sup> Claim language is *italicized*. Citations to a claim limitation and/or section number are to the limitation's mapping in this Petition and are incorporated by reference, including motivations to combine and showings of reasonable expectation of success, which themselves include the benefits described for the respective motivations to combine. All emphases added unless otherwise noted.

*Dantrey* also discloses “a system for enhancing audio” via “noise removal.”

*Dantrey*, Title, Abstract, 1:27-28, Fig. 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.

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;***

*Dantrey* in view of *Strake* renders obvious *fragment[ing]* (e.g., *Strake*'s framing functionality) 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 such as pitch frequency, pitch harmonic structure, formant structure, spectral envelope, spectral phase), as claimed.

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

The '496 Patent explains that *non-content elements* include *background noise* and that *speech characteristics* include *pitch, intonation, melody, stress, articulation, annunciation, voice identity, or unintelligible speech*. EX1001, Claim

3. Accordingly, *foreground speech content* in the context of the Challenged Claims refers to prominent speech content from a user (what is being said) in obtained audio data because the *speech characteristics* describe how the speech is being said and *non-content elements* refer to the noise. EX1003, ¶¶91-94 (discussing the claim language and background of the technology and explaining that audio data comprises both the prominent content of speech as well as characteristics of speech such as pitch, stress, formant structure, articulation, etc., in addition to noise which may be in the signal).

*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.

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. EX1003, ¶94 (discussing multiple background references). 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.* EX1003, ¶94 (explaining that if one hears human speech then one hears both the prominent content and the speech characteristics).

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.

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**

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. EX1003, ¶¶97-102.

*Dantrey* and *Strake's* express disclosures would have motivated such combination. *Id.* *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.” *Id.*; 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, and thus advantageous to include. EX1003, ¶100.

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). EX1003, ¶101.

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.-related information in captured speech and because such would have only required minor modifications in the signal processing of the captured speech. EX1003, ¶102.

(1) *Fragmenting input audio data into frames*

*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. EX1003, ¶103 (citing EX1027-Peer and explaining how 1024 samples divided by a sample rate of 16,000 samples per second equals in a frame size of 64 ms).

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.

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].

**c) Motivation to Combine**

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. EX1003, ¶106-110.

For example, *Dantrey* describes its system as a “pipeline” and describes exemplary sample sizes and rates of data for input into such pipeline. *Id.*; 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 ... [t]he justification for such a segmentation is that speech signals are non-stationary and exhibit quasi-stationary behavior at the shorter durations.” EX1003, ¶107 (quoting EX1029-Zhu). 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. *Id.* And a PHOSITA would have understood that continuous audio streams are advantageously broken up into frames before subsequent processing. *Id.*

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). EX1003, ¶109.

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 such would have only required minor modifications in programming. EX1003, ¶110.

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;***

*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.

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, Claim 9.

As explained above, the term *low-dimensional representation* should be construed as “a compressed representation of an input speech frame which is a result of a dimensionality reduction technique.” *Supra* Section III.C.1. Under this construction and in accordance with the specification, creating a *low-dimensional representation* by extracting features and applying a technique such as PCA satisfies this construction. EX1001, 7:22-27.

*Dantrey* teaches its “digital audio signal” from the captured speech is fed into 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; EX1003, ¶115 (comparing *Dantrey*’s claim 7 to claim 12).

*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”); EX1003, ¶¶115-120.

*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. EX1003, ¶¶117-119 (discussing *Dantrey*, 4:41-5:7).

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

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.

*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. PCA was a very well-known dimensionality reduction technique in the prior art, and PCA was also well-known to be applied to a variety of different data formats, specifically including spectrograms and MFCCs. EX1003, ¶122 (citing EX1008-Heckmann, 737, Fig.1, EX1014-Zheng, 2-4, EX1031-Wells, 13:36-42, EX1015-Wang, 3 (see §4.3), EX1012-Martinek, 18 (see §3.3.3).

Accordingly, the *Dantrey-Strake-Chen* combination teaches Claim 1[b] because *Dantrey’s* “noise subtractor 116” neural network applies *Chen’s* ML PCA dimensionality reduction technique to *Dantrey’s* noise reduced spectrogram representations which are based on *Dantrey’s* MFCC extracted features (which teaches the proposed construction of *low-dimensional representations* because the PCA compressed spectrograms are compressed representations of an input speech frames which are a result of a dimensionality reduction technique) before sending such PCA-compressed representations 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. EX1003, ¶123-124; EX1004-Dantrey, 3:54-64, 4:41-5:7, Fig. 1, claim 7.

*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**

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. EX1003, ¶¶125-128.

For 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, for example by saving on computational expense. EX1003, ¶126 (quoting EX1015-Wang, 1 (see Abstract, Introduction)). Applying a dimensionality reduction technique such as PCA saves the most important data characteristics while reducing the complexity by reducing redundant data coding information. *Id.*

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). EX1003, ¶127. A PHOSITA would have recognized that using PCA to encode the features allows for faster subsequent processing while still maintaining the most important information. EX1003, ¶127; 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 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 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. EX1003, ¶128; EX1006-Chen, 3:66-4:4; EX1014-Zheng, 2; 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***

*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.

*Dantrey* teaches that its “**audio generator**” (which receives the output from *Dantrey*’s noise subtractor as frames of PCA-reduced spectrograms with reduced noise in the present combination) “**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.

*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*. EX1003, ¶131.

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 (where the latter is the subject of Claim 1[d] discussed *infra*).

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**

A PHOSITA would have been motivated to incorporate *Strake*'s frame-based inverse transform technique (which produces time frame domains) into the *Dantrey-Strake-Chen* system explained above with a reasonable expectation of success. EX1003, ¶¶134-138.

For example, and as previously mentioned, speech processing pipelines such as *Dantrey* advantageously use frames to keep the processing pipeline appropriately filled and because “speech signals are non-stationary and exhibit quasi-stationary behavior at the shorter durations.” EX1003, ¶135 (quoting EX1029-Zhu). Accordingly, a PHOSITA would have been motivated to process speech data in *Dantrey’s* audio generator also in frames. *Id.*

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. EX1003, ¶136; EX1016-Wu, (“To further reduce the delay of voice conversion, frame-based approaches capable of converting spectral parameters frame by frame are more desirable”).

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). EX1003, ¶137.

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. EX1003, ¶138.

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.***

*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.

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.

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].

*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***”).

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

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; EX1005-*Strake*, [031] (“The second neural

network circuit 103 is configured to restore one or more speech components of the clean speech signal.”).

**a) Motivation to Combine**

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. EX1003, ¶¶145-148.

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. EX1003, ¶146; EX1017-Sargsyan, 9:15-20 (“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.”). Further, a PHOSITA would have known that using OLA is advantageous because it smooths the frames boundaries thus reducing the possibility of artifacts. *Id.*

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). EX1003, ¶147.

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. EX1003, ¶148.

**C. 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:***

*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.*

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

*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;***

*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.

*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”). EX1003, ¶153.

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 *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; EX1003, ¶154 (citing EX1034-*Kee*, [0027]). 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. EX1003, ¶154.

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

*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.

*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.

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. EX1003, ¶157 (citing EX1004-Dantrey, 3:10-15, 3:44-49, 4:11-14, 69:59, 95:21-41). 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*). EX1003, ¶157 (citing 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. EX1003, ¶150 (citing EX1034-Kee). 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. EX1003, ¶157.

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.***

*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.

*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.

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 [a] second person 124 through an appropriate speaker 126 or playback mechanism.” EX1004-*Dantrey*, 4:12-17.

**D. 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.**

*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.

See Claim 1[a].

**E. 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.**

*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.*

*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.*

**F. 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.**

*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 echoes and reverberation), as claimed.

*Dantrey* 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.

Additionally, *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 been motivated to incorporate *Strake's* pre-processing to remove echoes and reverberation into the *Dantrey-Strake-Chen* system described above with a reasonable expectation of success. EX1003, ¶168. 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. *Id.*; 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. EX1003, ¶168. 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.” *Id.*

**G. 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**

*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.

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.*”); EX1003, ¶170.

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]; EX1003, ¶171.

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.***

*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.*

As explained above, in the *Dantrey-Strake-Chen* system, a noise reduced spectrogram is produced via *Dantrey's* "noise subtractor" neural network based on the features extracted via MFCC. EX1003, ¶173 (discussing *Dantrey*, 4:41-5:7, Fig. 1); *see* Claims 1[a]-1[b], 9[a]. *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; EX1003, ¶¶173 (citing EX1014-*Zheng*, 2-4, EX1031-*Wells*, 13:36-42, EX1015-*Wang*, 3 (*see* §4.3)).

#### ***H. 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:***

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].

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;***

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;***

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;***

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;***

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*

See Claim 1[d].

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

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

- I. *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.*

See Claim 16; See Claim 3.

- J. *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.*

See Claim 16; See Claim 7.

## **VI. GROUND 2: CLAIMS 4-6, 11-15, AND 18-19 ARE OBVIOUS OVER DANTREY, STRAKE, CHEN, AND QUILLEN**

### **A. Ground Overview**

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 Mr. Schmandt explains, 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). EX1003, ¶¶185-188 (further discussing and citing EX1039-Demmin, 7:54-56, EX1040-Black, 1:30-35, EX1041-Guo, [0142], EX1042-Okamoto, 3 (see §4.1), EX1032-Arik, 10:56-61).

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.

***B. 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.***

*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.

*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.

*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].

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.

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*).

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”).

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].

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”).

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 in the Ground 2 Overview, this

is a fundamental, and obvious, concept in neural network training. *See* Section VI.A.; EX1003, ¶¶197, 185-188.

***1. Motivation to Combine***

A PHOSITA would have been motivated to incorporate *Quillen's* training techniques into the *Dantrey-Strake-Chen* system with a reasonable expectation of success. EX1003, ¶¶198-201. 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.” *See* Section VI.A.; EX1003, ¶185-188.

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. EX1003, ¶199.

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). EX1003, ¶200.

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.” EX1003, ¶201.

***C. 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.***

*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.

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.”*); EX1003, ¶203.

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”).

**D. 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.**

*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.

As explained in the Ground 2 Overview 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* Section VI.A, Claim 4.

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; EX1003, ¶207 (discussing *dynamic conversion* as claimed).

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*.”).

**E. Claim 11**

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

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.

*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,**

*See* Claims 4-6.

- 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;**

*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***

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.***

See Claim 1[d].

- F. ***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***

See Claim 11; See Claim 4.

- G. ***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.***

See Claim 11; See Claim 5.

- H. ***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.***

See Claim 11; See Claim 8.

**I. 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**

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.**

See Claim 11; See Claim 9[b].

- J. 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.**

See Claim 16; See Claim 4.

- K. 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.**

See Claim 16; See Claim 6.

**VII. GROUND 3: CLAIM 10 IS OBVIOUS OVER *DANTREY, STRAKE, CHEN, AND HECKMANN***

***A. Ground Overview***

Claim 10 adds one additional obvious concept to the Challenged Claim set, extracting features with a hierarchical feature extraction network containing a plurality of levels and performing compression at each level. As explained below this is further obvious because hierarchical feature extraction was a known technique for extracting audio features.

***B. 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.***

*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 such as pitch, and Heckmann's second layer that captures and compresses complex features, such as 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.

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.”

*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.

*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. EX1003, ¶225 (explaining that *Heckmann* depicts the first layer features as frequency at a given time, and that a PHOSITA would have understood pitch to be the perceived effect of frequency). *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; EX1003, ¶225. 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. EX1003, ¶225.

*Heckmann* next teaches a second layer of the hierarchical feature extraction framework which extracts complex features (*capture a different one or more of the features*). EX1008-Heckmann, 4, 8; EX1003, ¶226. 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.

*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. 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. EX1003, ¶227; EX1001, 7:42-44 (“a compression algorithm such as principal component analysis (PCA) or *non-negative matrix factorization (NMF)*, for example.”). 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 compress such. EX1003, ¶227 (discussing *Chen*); see Claims 1[b], 9.

### 1. *Motivation to Combine*

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. EX1003, ¶¶228-231.

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. Accordingly,

a PHOSTA would have been motivated to implement *Heckmann's* process to achieve higher robustness in a dynamic environments where de-noising is required, such as the *Dantrey-Strake-Chen* environment. EX1003, ¶229.

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 robust feature extraction for audio processing). *KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398, 417-418 (2007). 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. EX1003, ¶230; 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. 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.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because implementing the hierarchical feature extraction of *Heckmann* would only involve minor software changes to the system of *Dantrey-Strake-Chen*. EX1003, ¶231. Hierarchical feature extraction was well known in the art at the time of the '496 Patent's priority date. EX1003, ¶¶231, 35.

**VIII. GROUNDS 4-6: CLAIMS 1-20 ARE OBVIOUS OVER *DANTREY, STRAKE, AND LIU* (GROUND 4), AND ADDITIONALLY *QUILLEN* (GROUND 5), AND ADDITIONALLY *HECKMANN* (GROUND 6)**

***A. Grounds Overview***

Grounds 4-6 provide the same claim limitation analysis for claims 1-20 as Grounds 1-3 (which are incorporated by reference here) except for Chen's teaching of *low-dimensional representation* based on Petitioner's proposed construction. Chen is not used as a prior art reference in Grounds 4-6 and is not relied on to teach the claimed *low-dimensional representation*.

Grounds 4-6 demonstrate that the Challenged Claims are unpatentable under the express language *low-dimensional representation*, and thus regardless of whether the Board determines claim construction is necessary for this proceeding. In particular, the prior art teaches intermediate representations that are expressly described as "low dimensional" within the context of speech processing systems.

These Grounds do not rely on any requirement that the claimed *low-dimensional representation* be produced by a type of compression-based

dimensionality reduction technique (e.g., PCA) as required Petitioner’s proposed construction in view of the specification. Rather, they demonstrate that the claimed limitations are met where the prior art expressly discloses representations characterized as “low dimensional.”

Accordingly, Petitioner shows how the *low-dimensional representation* limitations are taught by *Liu* (and why a PHOSITA would have been motivated to incorporate *Liu*’s teachings of a *low-dimensional representation* into the combinations presented in Grounds 1-3). The analysis for all other limitations remains the same (and is incorporated by reference to these Grounds), except as specifically addressed below.

### B. *Liu*’s Teachings

*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:

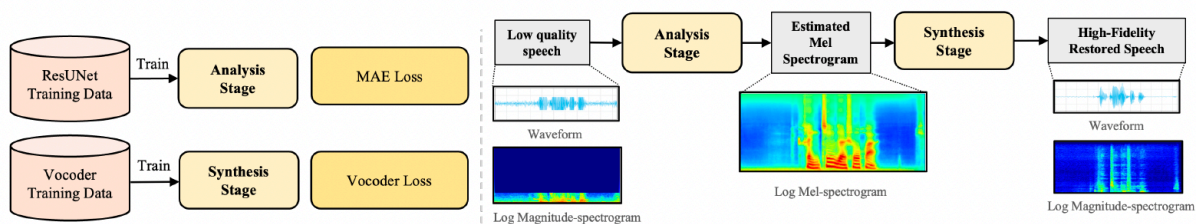


Figure 1: Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.

EX1043-Liu, Fig. 1, Abstract, §§ 1, 3.1-3.2.

*Liu* applies a first neural network (e.g., *Liu*'s analysis stage, ResUNet) to an input waveform of distorted/degraded speech to obtain the "low dimensional mel spectrogram." *Id.* *Liu* then applies a second neural network (e.g., *Liu*'s synthesis stage, a neural vocoder) to the "low dimensional mel spectrogram" to ultimately obtain a clean speech waveform. *Id.*; EX1003, 234-235 (explaining that ResUNet and neural vocoders were typical examples of neural networks before the '496 Patent). *Liu* thus teaches a speech processing pipeline that utilizes an intermediate representation expressly described as a "low dimensional mel spectrogram." EX1043-Liu, Abstract, § 3.1

***C. Specific Claim Limitation Applications for "low-dimensional representation"***

Petitioner provides a Ground-by-Ground, limitation-by-limitation analysis showing how *Liu*'s "low dimensional mel spectrogram" teachings render the Challenged Claims obvious without any express construction. For conciseness, Petitioner only provides additional discussion for limitations using the term *low-dimensional representation* (which do not themselves solely cite above for their respective analyses in Grounds 1-3), as the other limitations remain obvious as explained in Grounds 1-3 and are incorporated by reference here.

**1. Ground 4 (Dantrey-Strake-Liu)**

**a) Claim 1[b]**

*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.

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

In Grounds 4-6, *Dantrey*'s noise subtractor neural network does the same thing (i.e., subtracts noise) but instead of simply converting the input audio frames to "spectrograms," the combination is further modified by *Liu* to convert the input audio frames to "**low dimensional mel spectrogram[s]**," as expressly taught by *Liu*. EX1043-Liu, Abstract, Fig. 1, §§ 1, 3.1. While *Liu* describes its intermediate

representation as “low dimensional,” Petitioner does not rely on *Liu* as disclosing a type of compression-based dimensionality reduction technique (e.g., PCA) as required Petitioner’s proposed construction in view of the specification. Rather, *Liu* is cited in these Grounds to demonstrate that, applying the express claim language, the prior art teaches representations expressly characterized as “low dimensional.”

*(1) Motivation to Combine*

A PHOSITA would have been motivated to incorporate *Liu*’s low dimensional mel spectrogram technique into the *Dantrey-Strake* system with a reasonable expectation of success. EX1003, ¶¶241-245.

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.” EX1043-*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.” EX1003, 242; EX1043-*Liu*, Abstract, § 1.

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.” EX1003, 243 (citing EX1044-Sheng, 2). Additionally, 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. EX1003, 243.

Further, 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 and focusing on the human perception of speech). EX1003, 244.

A PHOSITA would have had a reasonable expectation of success in making the proposed combination. EX1003, 245. For example, it was well-known how to use a neural network to convert an audio signal into a low dimensional mel spectrogram, as *Liu* teaches. EX1003, 245; EX1043-Liu, Fig. 1, §§ 1, 3.1. And because *Dantrey* also teaches a two neural network pipeline system for noise removal and speech enhancement/restoration, a PHOSTIA would have been able to readily adjust *Dantrey* to obtain a low dimensional mel spectrogram with *Dantrey*’s first neural network (as opposed to the “spectrogram” taught by *Dantrey*), as

expressly taught by *Liu*. EX1003, 245. A PHOSITA would have further enjoyed a reasonable expectation of success in doing so because *Liu* provided its training teachings along with training/model/loss data and configuration files publicly online. EX1003, 245; EX1043-Liu, Abstract, n.1-n.5.

**b) Claim 1[c]**

*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.*

As 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. EX1043-Liu, Fig. 1, Abstract, §§ 1, 3.1-3.2.

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]**

*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.

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.

In these Grounds (Ground 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). EX1043-*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."); EX1003, 251.

Accordingly, the *Liu* teachings explained for Claim 1 in Ground 4 also teach Claim 9[b].

Accordingly, Ground 4 demonstrates that the Challenged Claims are unpatentable even without Petitioner's proposed construction for *low-dimensional representation*.

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

### **a) Claim 4**

*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.*

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**

*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 Ground 5 Claim 4.

Accordingly, Ground 5 demonstrates that the Challenged Claims are unpatentable even without Petitioner's proposed construction for *low-dimensional representation*.

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

**a) Claim 10**

Claim 10 does not include additional references to *low-dimensional representation* other than through its dependencies.

Accordingly, Ground 6 demonstrates that the Challenged Claims are unpatentable even without Petitioner's proposed construction for *low-dimensional representation*.

**IX. CONCLUSION**

For the foregoing reasons, Petitioner respectfully requests *inter partes* review of the Challenged Claims.

Respectfully submitted,

ERISE IP, P.A.

BY: /s/ Eric A. Buresh  
Eric A. Buresh, Reg. No. 50,394  
eric.buresh @eriseip.com  
7015 College Blvd., Suite 700  
Overland Park, KS 66211  
P: (913) 777-5600  
F: (913) 777-5601

COUNSEL FOR PETITIONER

**X. MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)**

**A. *Real Party-In-Interest***

Petitioner is the real parties-in-interest. 37 C.F.R. § 42.8(b)(1).

**B. *Related Matters***

Pursuant to 37 C.F.R. § 42.8(b)(2), Petitioner is aware of the following matters involving the '496 Patent or related patents:

- *Sanas.ai Inc. v. Krisp Technologies, Inc.*, Case No. 3:25-cv-05666-RS (N.D. Cal.)
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00272 (PTAB), challenging U.S. Patent No. 11,948,550
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00274 (PTAB), challenging U.S. Patent No. 12,131,745
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, IPR2026-00275 (PTAB), challenging U.S. Patent No. 11,715,457
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, PGR2026-00032 (PTAB), challenging U.S. Patent No. 12,412,561
- *Krisp Technologies, Inc. v Sanas.ai Inc.*, PGR2026-00033 (PTAB), challenging U.S. Patent No. 12,417,756

**C. Lead and Back-Up Counsel**

Petitioner provide the following designation and service information for lead and back-up counsel. 37 C.F.R. § 42.8(b)(3) and (b)(4).

<b>Lead Counsel</b>	<b>Back-Up Counsel</b>
<p>Eric A. Buresh (Reg. No. 50,394)  eric.buresh@eriseip.com  PTAB@eriseip.com</p> <p><u>Postal and Hand-Delivery Address:</u>  ERISE IP, P.A.  7015 College Blvd., Suite 700  Overland Park, Kansas 66211  Telephone: (913) 777-5600  Fax: (913) 777-5601</p>	<p>Chris R. Schmidt (Reg. No. 63,982)  chris.schmidt@eriseip.com</p> <p>Nick R. Apel (Reg. No. 84,017)  nick.apel@eriseip.com</p> <p>Nathan S. Johnson (Reg. No. 83,850)  nathan.johnson@eriseip.com</p> <p><u>Postal and Hand-Delivery Address:</u>  ERISE IP, P.A.  7015 College Blvd., Suite 700  Overland Park, Kansas 66211  Telephone: (913) 777-5600  Fax: (913) 777-5601</p> <p>Sten Larson (Reg. No. 81,518)  sten.larson@eriseip.com</p> <p><u>Postal and Hand-Delivery Address:</u>  ERISE IP, P.A.  717 17th Street, Suite 1400  Denver, Colorado 80202  Telephone: (913) 777-5600  Fax: (913) 777-5601</p>

***D. 37 C.F.R. § 42.8(b)(4) – Service Information***

Please address all correspondence to the lead and back-up counsel as shown above. Petitioner consents to electronic service by e-mail at the e-mail addresses provided above.

**CLAIM LISTING APPENDIX**  
**U.S. Patent No. 12,125,496 for Claims 1-20**

<b>Claim Designation</b>	<b>Claim Language</b>
Claim 1(P)	A voice enhancement system, comprising memory having instructions stored thereon and one or more processors coupled to the memory and configured to execute the instructions to:
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;
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;
Claim 1(c)	apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames; and
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.
Claim 2(P)	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:
Claim 2(a)	digitize analog input audio signals obtained via the physical microphone to generate the input audio data;
Claim 2(b)	convert the output audio data to analog audio output signals; and
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.

<b>Claim Designation</b>	<b>Claim Language</b>
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.
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 representation of input audio training data speech frames.
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.
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.
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.
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.

<b>Claim Designation</b>	<b>Claim Language</b>
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
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.
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.
Claim 11(P)	A method for real-time voice enhancement, the method implemented by a voice enhancement system and comprising:
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,
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;
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
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.

<b>Claim Designation</b>	<b>Claim Language</b>
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.
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.
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.
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
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.
Claim 16 (P)	A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:
Claim 16(a)	digitize analog input audio signals to generate input audio data;
Claim 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;
Claim 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;
Claim 16(d)	apply a second neural network to the low-dimensional representations of the input speech frames to generate target speech frames;

<b>Claim Designation</b>	<b>Claim Language</b>
Claim 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
Claim 16(f)	convert the output audio data to analog audio output signals before providing the analog audio output signals to an audio output device.
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.
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.
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.
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.

**APPENDIX OF EXHIBITS**

<b>Exhibit 1001</b>	U.S. Patent No. 12,125,496 (“496 Patent”)
<b>Exhibit 1002</b>	Prosecution History of the 12,125,496 Patent
<b>Exhibit 1003</b>	Declaration of Christopher Schmandt
<b>Exhibit 1004</b>	U.S. Patent No. 12,412,590 to Dantrey et al. (“Dantrey”)
<b>Exhibit 1005</b>	WIPO Publication No. 2020/199990 to Strake et al. (“Strake”)
<b>Exhibit 1006</b>	U.S. Patent No. 10,867,616 to Chen et al. (“Chen”)
<b>Exhibit 1007</b>	U.S. Patent Application Publication No. 2021/0241780 to Quillen (“Quillen”)
<b>Exhibit 1008</b>	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”)
<b>Exhibit 1009</b>	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”)
<b>Exhibit 1010</b>	Christopher Schmandt, <i>Voice Communication with Computers Conversational Systems</i> , Van Nostrand Reinhold, (1994) (“Schmandt”)
<b>Exhibit 1011</b>	U.S. Patent Application Publication No. 2024/0098218 to Nguyen et al. (“Nguyen”)
<b>Exhibit 1012</b>	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. <a href="https://doi.org/10.3390/s21196343">https://doi.org/10.3390/s21196343</a> (“Martinek”)
<b>Exhibit 1013</b>	Casey O’Callaghan, <i>Pitch</i> , <a href="http://caseyocallaghan.com/research/papers/Pitch.pdf">http://caseyocallaghan.com/research/papers/Pitch.pdf</a> (“O’Callaghan”)
<b>Exhibit 1014</b>	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”)
<b>Exhibit 1015</b>	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”)

<b>Exhibit 1016</b>	Shuhua Gao, Xiaoling Wu, Cheng Xiang, and Dongyan Huang, Development of a Computationally Efficient Voice Conversion System on Mobile Phones, <a href="https://doi.org/10.1017/ATSIP.2018.23">https://doi.org/10.1017/ATSIP.2018.23</a> (2018) (“ <i>Wu</i> ”)
<b>Exhibit 1017</b>	U.S. Patent Application Publication No. 2020/0066296 to Sargsyan et al. (“ <i>Sargsyan</i> ”)
<b>Exhibit 1018</b>	Su-Hyun Han , Ko Woon Kim, SangYun Kim , Young Chul Youn, <i>Artificial Neural Network: Understanding the Basic Concepts without Mathematics</i> , Dement Neurocognitive Disord. 2018 Sep;17(3):83-89 (“ <i>Han</i> ”)
<b>Exhibit 1019</b>	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> ”)
<b>Exhibit 1020</b>	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> ”)
<b>Exhibit 1021</b>	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> ”)
<b>Exhibit 1022</b>	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> ”)
<b>Exhibit 1023</b>	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> ”)
<b>Exhibit 1024</b>	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> ”)
<b>Exhibit 1025</b>	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> ”)
<b>Exhibit 1026</b>	PCT Publication No. 2022/168102 to Keshet et al. (“ <i>Keshet</i> ”)
<b>Exhibit 1027</b>	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> ”)

<b>Exhibit 1028</b>	U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“ <i>Prabhavalkar</i> ”)
<b>Exhibit 1029</b>	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> ”)
<b>Exhibit 1030</b>	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> ”)
<b>Exhibit 1031</b>	U.S. Patent No. 7,328,153 to Wells et al. (“ <i>Wells</i> ”)
<b>Exhibit 1032</b>	U.S. Patent No. 10,796,686 to Arik et al. (“ <i>Arik</i> ”)
<b>Exhibit 1033</b>	U.S. Patent Application Publication No. 2003/0088408 to Thyssen et al. (“ <i>Thyssen</i> ”)
<b>Exhibit 1034</b>	U.S. Patent Application Publication No. 2005/0066209 to Kee et al. (“ <i>Kee</i> ”)
<b>Exhibit 1035</b>	Abdullah Zaini Alsheibi, <i>Unsupervised Learning Algorithm for Noise Suppression and Speech Enhancement Applications</i> , Electronic Theses and Dissertations. 2168. <a href="https://digitalcommons.du.edu/etd/2168">https://digitalcommons.du.edu/etd/2168</a> (2023) (“ <i>Alsheibi</i> ”)
<b>Exhibit 1036</b>	U.S. Patent Application Publication No. 2015/0371655 to Gao (“ <i>Gao</i> ”)
<b>Exhibit 1037</b>	U.S. Patent No. 10,561,361 to Quatieri et al. (“ <i>Quatieri</i> ”)
<b>Exhibit 1038</b>	U.S. Patent No. 9,195,649 to Neuhauser et al. (“ <i>Neuhauser</i> ”)
<b>Exhibit 1039</b>	European Patent Application Publication No. 0756172 to Demmin et al. (“ <i>Demmin</i> ”)
<b>Exhibit 1040</b>	U.S. Patent No. 6,269,351 to Black (“ <i>Black</i> ”)
<b>Exhibit 1041</b>	U.S. Patent Application Publication No. 2017/0193066 to Zhu et al. (“ <i>Zhu</i> ”)
<b>Exhibit 1042</b>	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> ”)
<b>Exhibit 1043</b>	Haohe Liu et al., <i>VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration</i> , Interspeech 2022 (“ <i>Liu</i> ”)
<b>Exhibit 1044</b>	Leyuan Sheng et al., <i>Reducing sover-smoothness in speech synthesis using Generative Adversarial Networks</i> , IEEE 2018 (“ <i>Sheng</i> ”)
<b>Exhibit 1045</b>	Declaration of Mina Ching, Records Request Processor at the Internet Archive (“ <i>Ching</i> ”)

<b>Exhibit 1046</b>	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> ”)
<b>Exhibit 1047</b>	Patrik O. Hoyer, <i>Non-Negative Sparse Coding</i> , Neural Networks Research Centre Helsinki University of Technology (“ <i>Hoyer</i> ”)
<b>Exhibit 1048</b>	Haohe Liu et al., <i>VoiceFixer: Toward General Speech Restoration with Neural Vocoder</i> , October 5, 2021 (“ <i>VoiceFixer</i> ”)
<b>Exhibit 1049</b>	Declaration of June Munford dated March 18, 2026 regarding public availability of <i>Heckmann</i> (“ <i>Munford-Heckmann</i> ”)
<b>Exhibit 1050</b>	Declaration of June Munford dated March 24, 2026 regarding public availability of <i>Liu</i> (“ <i>Munford-Liu</i> ”)

**CERTIFICATION OF WORD COUNT**

The undersigned certifies pursuant to 37 C.F.R. §42.24 that the foregoing Petition for *Inter Partes* Review, excluding any table of contents, mandatory notices under 37 C.F.R. §42.8, certificates of service or word count, or appendix of exhibits, contains 13,584 words according to the word-processing program used to prepare this document (Microsoft Word).

Dated: March 26, 2026

ERISE IP, P.A.

BY: /s/ Eric A. Buresh  
Eric A. Buresh, Reg. No. 50,394  
eric.buresh@eriseip.com  
7015 College Blvd., Suite 700  
Overland Park, KS 66211  
P: (913) 777-5600  
F: (913) 777-5601

*COUNSEL FOR PETITIONER*

**CERTIFICATE OF SERVICE ON PATENT OWNER  
UNDER 37 C.F.R. § 42.105**

Pursuant to 37 C.F.R. §§ 42.6(e) and 42.105(b), the undersigned certifies that on March 21, 2024, a complete and entire copy of this Petition for *Inter Partes* Review and Exhibits were provided via Federal Express to the Patent Owner by serving the correspondence address of record for the '496 Patent:

Troutman Pepper Locke LLP  
80 Linden Oaks  
Suite 320  
Rochester, NY 14625

Further, a courtesy copy of this Petition for *Inter Partes* Review was sent via email to Patent Owner's litigation counsel:

Michael Ng ([Michael.Ng@kobrekim.com](mailto:Michael.Ng@kobrekim.com))  
Daniel Zabeer ([Daniel.Zaheer@kobrekim.com](mailto:Daniel.Zaheer@kobrekim.com))  
Jessica Fender ([Jessica.Fender@kobrekim.com](mailto:Jessica.Fender@kobrekim.com))  
Victoria Fordin ([Victoria.Fordin@kobrekim.com](mailto:Victoria.Fordin@kobrekim.com))  
Zachary Ritz ([Zachary.Ritz@kobrekim.com](mailto:Zachary.Ritz@kobrekim.com))

Respectfully submitted,

ERISE IP, P.A.

BY: /s/ Eric A. Buresh  
Eric A. Buresh, Reg. No. 50,394  
[eric.buresh@eriseip.com](mailto:eric.buresh@eriseip.com)  
7015 College Blvd., Suite 700  
Overland Park, KS 66211  
P: (913) 777-5600  
F: (913) 777-5601

*COUNSEL FOR PETITIONER*