UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.,
Petitioner

v.

SANAS.AI INC.,
Patent Owner

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

PETITION FOR POST-GRANT REVIEW OF U.S. PATENT NO. 12,417,756

TABLE OF CONTENTS

I. INTRODUCTION 1

II. SUMMARY OF THE '756 PATENT 1

A. DESCRIPTION OF THE '756 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. § 42.204 3

A. GROUNDS FOR STANDING UNDER 37 C.F.R. § 42.204(A) 3
B. IDENTIFICATION OF CHALLENGE UNDER 37 C.F.R. § 42.204(B) 4
C. CLAIM CONSTRUCTION UNDER 37 C.F.R. § 42.204(B)(3) 4

IV. SHOWING OF ANALOGOUS, PRIOR ART 5

A. OVERVIEW OF *NGUYEN* 6
B. OVERVIEW OF *DIRAC* 7
C. OVERVIEW OF *NARAYANAN* 9
D. OVERVIEW OF *KAMRUZZAMAN* 10
E. OVERVIEW OF *KEE* 10
F. OVERVIEW OF WANG 11

V. GROUND 1: NGUYEN IN VIEW OF KEE, DIRAC, AND NARAYANAN RENDERS OBVIOUS CLAIMS 1-3, AND 6 OF THE '756 PATENT 12

A. CLAIM 1 12
1. *Claim 1[Pre]: "A speech processing system, comprising an audio interface coupled to a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to:"* *12*
a) *Nguyen's* Teachings 13
b) *Kee's* Teachings 16
c) Motivation to Combine 17
2. *Claim 1[a]: "apply one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent*

features of first input speech associated with a first accent of a first user;" *20*
a) *Nguyen's* Teachings 20
b) *Dirac's* Teachings 22
c) Motivation to Combine 24
3. Claim 1[b]: "analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns;" *27*
a) *Nguyen's* Teachings 28
b) *Narayanan's* Teachings 30
c) Motivation to Combine 33
4. Claim 1[c]: "synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent; and" *37*
5. Claim 1[d]: "provide to the audio interface output audio data for output via the audio output device, wherein the output audio data is generated based on the modified version of the second input speech." *40*
B. CLAIM 2 41
1. Claim 2: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics." *41*
C. CLAIM 3 42
1. Claim 3: "The speech processing system of claim 1, wherein the accent features comprise one or more pitch contours, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the

other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent." *42*

D. CLAIM 6 45

1. Claim 6: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device." *45*

VI. GROUND 2: NGUYEN IN VIEW OF KEE, DIRAC, NARAYANAN, AND KAMRUZZAMAN RENDERS OBVIOUS CLAIM 4 OF THE '756 PATENT 46

A. CLAIM 4 46

1. Claim 4: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint." *46*

a) *Kamruzzaman's* Teachings 47

b) Motivation to Combine 49

VII. GROUND 3: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND WANG RENDERS OBVIOUS CLAIM 5 OF THE '756 PATENT 51

A. CLAIM 5 51

1. Claim 5: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics." *51*

a) *Wang's* Teachings 52

b) Motivation to Combine 55

VIII. GROUND 4: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND KAMRUZZAMAN RENDERS OBVIOUS CLAIMS 7-11 AND 13 OF THE '756 PATENT.....57

A. CLAIM 7 57

1. Claim 7[Pre]: "A method for real-time accent mimicking, the method implemented by a speech processing system and comprising:".....57

2. Claim 7[a]: "applying one or more trained machine learning models to first input audio data to extract accent features of first input speech associated with a first accent of a first user;".....58

3. Claim 7[b]: "analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise a unique fingerprint of the voice of the second user;".....58

4. Claim 7[c]: "synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and".....59

5. Claim 7[d]: "providing output audio data generated based on the modified version of the second input speech.".....59

B. CLAIM 8 59

1. Claim 8: "The method of claim 7, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.".....59

C. CLAIM 9 59

1. Claim 9: "The method of claim 7, further comprising extracting from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.".....59

D. CLAIM 10 59

1. Claim 10: "The method of claim 7, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch

contours comprise variations in pitch throughout the first input speech. The intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent." 59

E. CLAIM 11 60

1. Claim 11: "The method of claim 7, further comprising applying a mel frequency cepstral coefficient (MFCC) analysis to extract the unique fingerprint of the voice of the second user." 60

F. CLAIM 13 60

1. Claim 13: "The method of claim 7, further comprising receiving the second input audio data from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device." 60

IX. GROUND 5: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND WANG RENDERS OBVIOUS CLAIMS 14-17 AND 19-20 OF THE '756 PATENT 60

A. CLAIM 14 60

1. Claim 14[Pre]: "A non-transitory computer-readable medium comprising instructions that, when executed by at least one processor, cause the at least one processor to:" 60

2. Claim 14[a]: "apply one or more trained machine learning models to first input audio data obtained via a microphone to extract accent features of first input speech associated with a first accent of a first user;" 61

3. Claim 14[b]: "analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise speaker-specific voice characteristics;" 61

4. Claim 14[c]: "synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and" 61

5. Claim 14[d]: "provide for output via an audio output device output audio data generated based on the modified version of the second input speech." *62*
B. CLAIM 15 62
1. Claim 15: "The non-transitory computer-readable medium of claim 14, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent." *62*
C. CLAIM 16 62
1. Claim 16: "The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics." *62*
D. CLAIM 17 62
1. Claim 17: "The non-transitory computer-readable medium of claim 14, wherein the accent features comprise one or more pitch contour, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent." *62*
E. CLAIM 19 63
1. Claim 19: "The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor to apply a speaker identity encoding technique to encode the speaker-specific voice characteristics." *63*
F. CLAIM 20 63
1. Claim 20: "The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device." *63*

X. GROUND 6: NGUYEN IN VIEW OF DIRAC, NARAYANAN, KAMRUZZAMAN, AND WANG RENDERS OBVIOUS CLAIMS 12 AND 18 OF THE '756 PATENT 63

A. CLAIM 12 63
1. Claim 12: "The method of claim 7, further comprising applying a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics." 63
B. CLAIM 18 64
1. Claim 18: "The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint." 64

XI. CONCLUSION 64

XII. MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1) 65

A. REAL PARTY-IN-INTEREST 65
B. RELATED MATTERS 65
C. LEAD AND BACK-UP COUNSEL 66
D. 37 C.F.R. § 42.8(B)(4) – SERVICE INFORMATION 67

TABLE OF AUTHORITIES

CASES

In re Translogic Technology, Inc., 504 F.3d 1249 (Fed. Cir 2007) 18
KSR Int'l Co. v. Teleflex Inc., 550 U.S. 398 (2007) 18, 19, 26
Phillips v. AWH Corp., 415 F.3d 1303 (Fed. Cir. 2005) (en banc) 4

STATUTES

35 U.S.C. § 102(a)(1) 6, 7, 9, 10, 11
35 U.S.C. § 102(a)(2) 6, 7, 9, 11

I. INTRODUCTION

Petitioner, Krisp Technologies, Inc. ("Petitioner") requests Post-Grant Review ("PGR") of Claims 1-20 (the "Challenged Claims") of U.S. Patent No. 12,417,756 (the "'756 Patent").

II. SUMMARY OF THE '756 PATENT

A. Description of the '756 Patent

The '756 Patent is generally directed to audio analysis and real-time accent mimicking. EX1001, 1:10-12. To accomplish the accent mimicking, the patent proposes a speech processing system which performs the following steps:

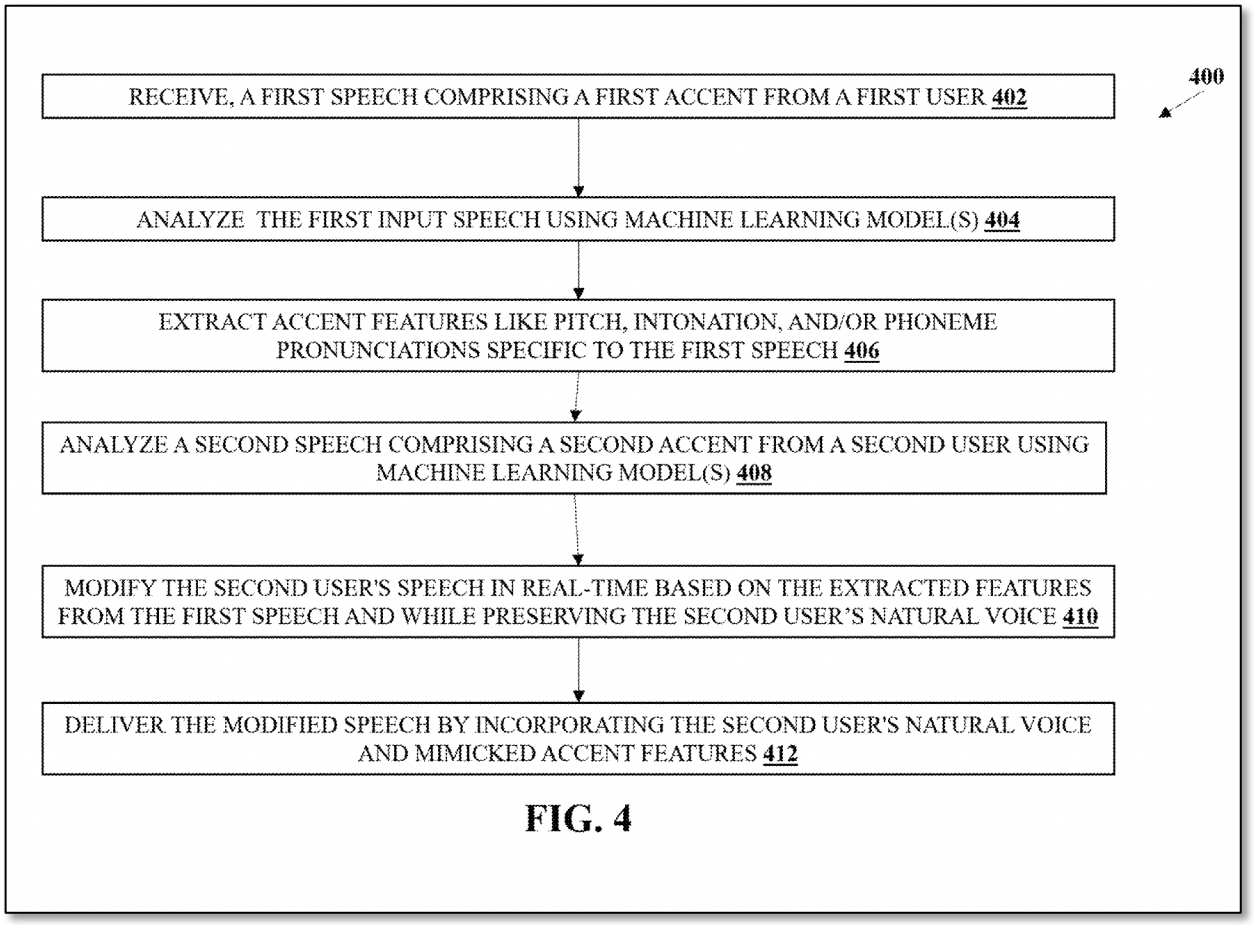


FIG. 4

EX1001, Fig. 4.

B. Summary of the Prosecution History

The Application that resulted in the '756 Patent was filed on January 17, 2025, and purports to claim priority to provisional patent application No. 63/678,180, filed on August 1, 2024. EX1001, (22), (60). For purposes of this Petition and without waiving its right to challenge priority in this or any other proceeding, Petitioner applies August 1, 2024, as the priority date for the Challenged Claims.

Applicant received a non-final rejection on March 13, 2025, in which the claims were rejected as anticipated and obvious in view of the prior art. EX1002, 82-85. In response, the Applicant amended the claims to narrow the scope of "generated characteristics." EX1002, 89-100.

The Examiner found the claims patentable in view of these amendments:

> Prior art of record discloses a similar method, but fails to teach the claims in combination with analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns and synthesizing a modified version of the second input speech by modifying the obtained second input audio data based on the generated

> characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.

EX1002, 132.

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

A PHOSITA at the time of the '756 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 speech processing and machine learning, or a bachelor's degree in the same fields with at least four years of experience in the field of speech processing. Additional education or experience might substitute for the above requirements. EX1003, ¶27.[1]

III. REQUIREMENTS UNDER 37 C.F.R. § 42.204

A. Grounds for Standing Under 37 C.F.R. § 42.204(a)

Petitioner certifies the '756 Patent is eligible for post-grant review and that the Petitioner is not barred or estopped from requesting a PGR challenging the claims of the '756 Patent. 37 C.F.R. § 42.204(a). Further, this Petition is filed within nine months of the date of the grant of the '756 Patent (September 16, 2025). 37 C.F.R. § 42.202.

[1] All citations to EX1003 are to paragraph numbers in the Declaration of Christopher Schmandt.

B. Identification of Challenge Under 37 C.F.R. § 42.204(b)

In view of the prior art and evidence presented, the Challenged Claims of the '756 Patent are unpatentable and should be canceled. 37 C.F.R. § 42.204(b)(1). Further, based on the prior art references identified below, PGR of the Challenged Claims should be granted. 37 C.F.R. § 42.204(b)(2).

Claim(s)	Basis[2]	Reference(s)
1-3, 6	§ 103	*Nguyen* (EX1004), *Kee* (EX1012), *Dirac* (EX1005), *Narayanan* (EX1006)
4	§ 103	*Nguyen*, *Kee*, *Dirac*, *Narayanan*, *Kamruzzaman* (EX1011)
5	§ 103	*Nguyen*, *Kee*, *Dirac*, *Narayanan*, *Wang* (EX1007)
7-11, 13	§ 103	*Nguyen*, *Dirac*, *Narayanan*, *Kamruzzaman*
14-17, 19-20	§ 103	*Nguyen*, *Dirac*, *Narayanan*, *Wang*
12, 18	§ 103	*Nguyen*, *Dirac*, *Narayanan*, *Kamruzzaman*, *Wang*

C. Claim Construction Under 37 C.F.R. § 42.204(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. § 42.200(b); *Phillips v. AWH Corp.*, 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc). For purposes of this proceeding, Petitioner applies the plain and ordinary meaning of all claim terms as understood by a PHOSITA, unless specifically noted below.

[2] The AIA version of 35 U.S.C. § 103 applies.

Petitioner proposes that "***characteristics specific to a natural voice of the second user***" should be construed as "**speaker-specific attributes that are globally representative of the second user's voice and that are used to maintain the second user's voice throughout synthesis**." The specification consistently describes the claimed "***characteristics specific to a natural voice of the second user***" as speaker-specific information of the second user that is used globally to maintain the natural sound of the second user throughout the speech modification process. EX1001, 4:59-63, 5:1-15, 5:64-6:3, 6:19-23.

IV. SHOWING OF ANALOGOUS, PRIOR ART

The prior art is analogous to the claimed invention of the '756 Patent. The prior art is from the same field of endeavor as the '756 Patent, namely speech audio processing. EX1001, Abstract, 1:10-12, 2:14-22; EX1004-Nguyen, Abstract, [0075], [0098], [0117]; EX1005-Dirac, Abstract, 1:5-20, 2:19-22, 3:12-17; EX1006-Narayanan, Title, Abstract, [0004], [0017]; EX1011-Kamruzzaman, Title, Abstract, 2-3; EX1012-Kee, [0027]; EX1007-Wang, Abstract, [0028], [0030], [0047]; EX1003, ¶¶49, 53, 56, 59, 62, 65, 68, 71.

The prior art is also reasonably pertinent to at least one problem facing the '756 Patent inventors, namely at least one of:

(A) Retaining the voice characteristics of a speaker while processing speech audio. EX1001, Abstract, 1:45-55, 5:11-15, 6:32-40, 7:44-48; EX1004-

Nguyen, Abstract, [0013], [0016], [0072], [0073], [0076]; EX1006-Narayanan, Abstract, [0018], [0023]-[0024], [0035], [0075], EX1007-Wang, Abstract, [0030], [0036], [0047], [0049]; EX1003, ¶¶63, 72.

(B) Performing real-time accent conversion. EX1001, Title, 1:10-12, 1:35-44, 1:49-55; EX1004-Nguyen, [0012]-[0013], [0099]; EX1005-Dirac, 3:7-9, 7:61-8:15; EX1003, ¶¶57, 60.

(C) Extracting vocal features that are unique to a speaker. EX1001, 4:19-29, 5:1-5, 7:58-67, 8:34-47; EX1011-Kamruzzaman, Abstract, 1-2, 5; EX1003, ¶66.

(D) Inputting and outputting audio data. EX1001, 3:4-7; EX1012-Kee, [0027]; EX1003, ¶69.

A. Overview of *Nguyen*

U.S. Patent Application Publication No. 2024/0098218 ("*Nguyen*") was filed on January 30, 2023, and published on March 21, 2024. *Nguyen* claims priority to provisional application No. 63/406,942 filed on September 15, 2022. *Nguyen*, therefore, is prior art to the '756 Patent under 35 U.S.C. §§ 102(a)(1)-(2). *Nguyen* was not cited or considered during the prosecution of the '756 Patent. *See generally* EX1001, EX1002.

Nguyen teaches "accent conversion for virtual conferences[.]" EX1004-Nguyen, Title, [0018]. *Nguyen's* accent conversion is accomplished by one or more

accent conversion ("AC") models. The AC models "receive audio streams from the virtual conference and each generates an output audio stream having the corresponding target accent." EX1004-Nguyen, [0070].

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

B. Overview of *Dirac*

U.S. Patent No. 10,163,451 ("*Dirac*") was filed on December 21, 2016, and issued on December 25, 2018. *Dirac*, therefore, is prior art to the '756 Patent under 35 U.S.C. §§ 102(a)(1)-(2). *Dirac* was not cited or considered during the prosecution of the '756 Patent. *See generally* EX1001, EX1002.

Dirac teaches "[t]echniques for accent translation." EX1005-Dirac, Abstract. Namely, "a first party and a second party may participate in an audio communication session." EX1005-Dirac, 7:63-64. The communications session takes place over various networks or on a phone call. EX1005-Dirac, 7:65-8:7. The accent translation

techniques "may be used in other scenarios, such as to translate the accent of a live speaker or presenter to an audience." EX1005-Dirac, 8:8-11.

"[S]peech spoken by the first party may be captured by audio capture components 621, such as a microphone." EX1005-Dirac, 8:19-21. The speech is then provided as input to the accent translation components. EX1005-Dirac, 8:21-23. Fig. 7 provides an example of the accent translation system taught by *Dirac*. EX1005-Dirac, 1:48-50.

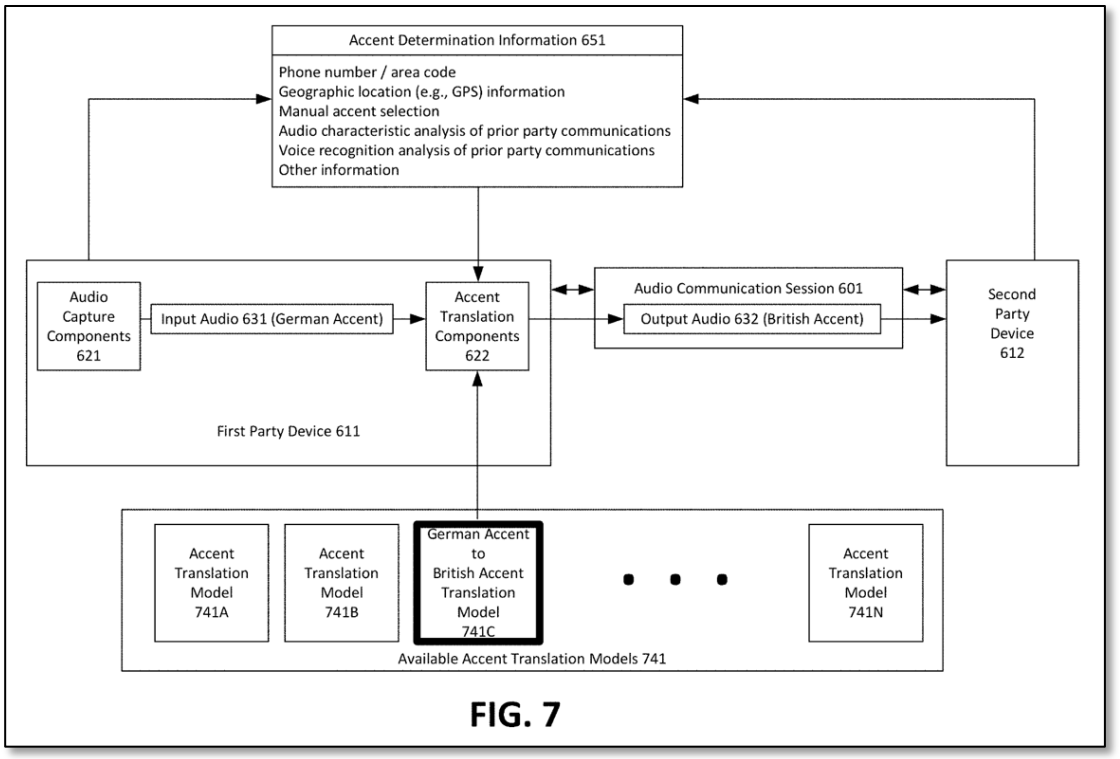


FIG. 7

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

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

C. Overview of *Narayanan*

U.S. Patent Application No. 2022/0417659 ("*Narayanan*") was filed on June 23, 2021, and was published on December 29, 2022. *Narayanan*, therefore, is prior art to the '756 Patent under 35 U.S.C. § 102(a)(1)-(2). *Narayanan* was not cited or considered during the prosecution of the '756 Patent. *See generally* EX1001, EX1002.

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

> For example, the voice profile module 106 may determine the voice profile based on the spoken audio content in the content portion that indicates the incorrect word (as opposed to background audio content in the content portion). The voice profile may be determined based on audio, vocal, and/or linguistic characteristics of the spoken audio content, such as audio spectral patterns, vocal frequency, vocal pitch, speaking speed, intonation, loudness, amplitude, speech patterns, speech cadence, the number and/or duration of utterances, the number and/or duration of breaks between utterances.

EX1006-Narayanan, [0023].

D. Overview of *Kamruzzaman*

"*Speaker Identification using MFCC-Domain Support Vector Machine*" by Kamruzzaman et al. ("*Kamruzzaman*"), published in the International Journal of Electrical and Power Engineering in 2007, was publicly available at least as early as September 20, 2007, and therefore, qualifies as prior art to the '756 Patent under at least 35 U.S.C. § 102(a)(1). EX1036-Munford, ¶¶62-66 (declaration of June Munford establishing public availability of *Kamruzzaman*). *Kamruzzaman* was not cited or considered during the prosecution of the '756 Patent. *See generally* EX1001, EX1002.

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

E. Overview of *Kee*

U.S. Patent Application Publication No. 2005/0066209 ("*Kee*") was filed on June 17, 2004, and published March 24, 2005. *Kee*, therefore, is prior art to the '756

Patent under 35 U.S.C. §§ 102(a)(1)-(2). *Kee* was not cited or considered during the prosecution of the ’756 Patent. *See generally* EX1001, EX1002.

Kee teaches a portable electronic device. EX1012-Kee, Title. The system contains a main processor that runs applications on an operating system. EX1012-Kee, Abstract. The system contains an audio interface as well as a digital-to-analog converter and an analog-to-digital converter. EX1012-Kee, [0002].

F. Overview of Wang

U.S. Patent Application No. 2024/0274122 (“*Wang*”) was filed on March 30, 2023, and published August 15, 2024. *Wang* claims priority to provisional application No. 63/444,312 filed on February 9, 2023. *Wang*, therefore, is prior art to the ’756 Patent under at least 35 U.S.C. § 102(a)(2). *Wang* was not cited or considered during the prosecution of the ’756 Patent. *See generally* EX1001, EX1002.

Wang teaches a speech translation system. EX1007-Wang, Abstract. *Wang* discloses using a speaker encoder as part of the speech translation system:

> The expressive speech translation system 100 may include a ***speaker encoder 130*** for generating s***peaker embedding data 135***. The speaker embedding data 135 represents ***voice identity characteristics*** that differ from speaker to speaker such as ***fundamental frequency (e.g., pitch), timbre, cadence,*** etc. In some implementations, the expressive speech translation system 100 may use the speaker encoder 130 to generated speaker

> embedding data 135b using the input audio data; thus, generating a speaker embedding that represents a speaker of the source speech.

EX1007-Wang, [0030].

V. GROUND 1: NGUYEN IN VIEW OF KEE, DIRAC, AND NARAYANAN RENDERS OBVIOUS CLAIMS 1-3, AND 6 OF THE '756 PATENT

A. Claim 1

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

To the extent the preamble is limiting, *Nguyen* in view of *Kee* renders obvious *a speech processing system*[3] (e.g., *Nguyen's* accent conversion system)*, comprising an audio interface* (e.g., *Kee's* audio interface) *coupled to a microphone* (e.g., *Nguyen's* microphone) *and an audio output device* (e.g., *Nguyen's* speaker or headphones)*, memory* (e.g., *Nguyen's* memory) *having instructions stored thereon*

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

(e.g., *Nguyen's* processor executable instructions)*, and one or more processors* (e.g., *Nguyen's* processor) *coupled to the memory* (e.g., *Nguyen's* memory) *and the audio interface* (e.g., *Kee's* audio interface) *and configured to execute the instructions*, as claimed.

a) *Nguyen's* Teachings

Nguyen discloses a "system 300 for accent conversion for virtual conferences." EX1004-Nguyen, [0063], Title, Abstract. In the system, a video conferencing application located on client devices is used to perform accent conversion. EX1004-Nguyen, [0072]. The client devices are "able to interact with each other to conduct the meeting, such as by viewing video feeds and hearing audio feeds from other participants, and by capturing and transmitting video and audio of themselves." EX1004-Nguyen, [0064].

The client devices also utilize a microphone for audio input and a speaker or headphones for audio output. EX1004-Nguyen, [0072], [0012], [0115].

Fig. 3B shows a client device on which "Accent Conversion" (AC) is performed by way of the video conferencing application:

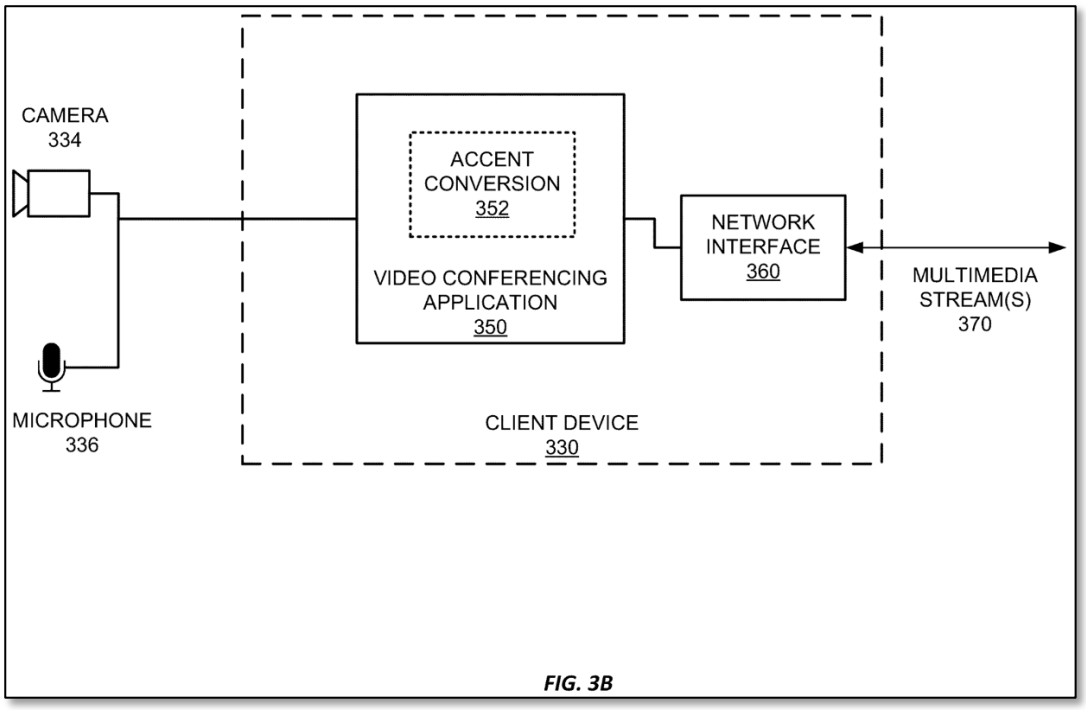
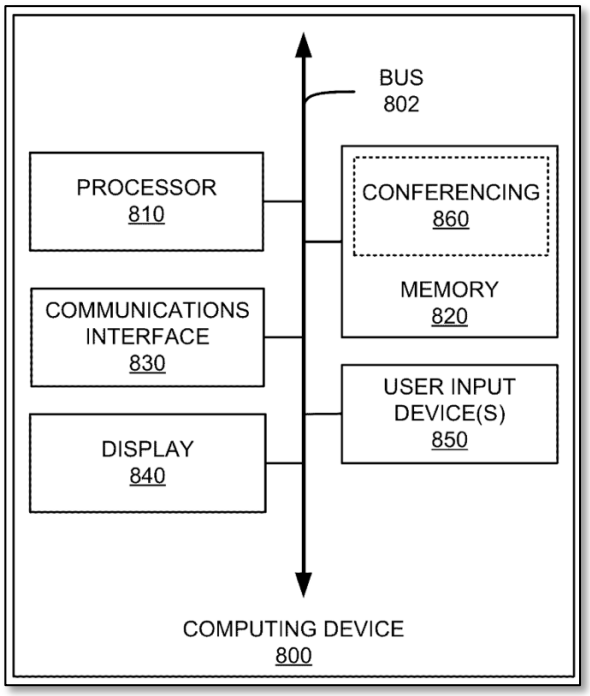


FIG. 3B

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

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

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



EX1004-Nguyen, Fig. 8.

Therefore, *Nguyen* discloses *a speech processing system, comprising ... a microphone and an audio output device, memory having instructions stored thereon, and one or more processors coupled to the memory and the audio interface and configured to execute the instructions to*.

However, *Nguyen* does not expressly disclose *an audio interface.*

b) *Kee's* Teachings

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

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

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

EX1012-Kee, [0027].

The *Nguyen-Kee* combination teaches *an audio interface coupled to a microphone and an audio output device* and *one or more processors coupled to the memory and the audio interface*. *Nguyen's* "processor 810 [] is in communication with the memory 820 and ***other components of the computing device*** 800 using one or more communications buses," where such "other components" in the *Nguyen-Kee* combination include *Nguyen*'s microphone, *Nguyen's* speaker/headphone, and *Kee's* audio interface. EX1004-Nguyen, [0072], [0012], [0115], [0117], Fig. 8; EX1012-Kee, [0027].

c) Motivation to Combine

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

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

First, *Nguyen* explicitly contemplates receiving input audio for processing, but does not provide details on how this functionality is implemented. EX1003, ¶84 (discussing *Nguyen* at [0071]-[0072]). A PHOSITA would have known that for a computing device to process audio data, the audio data first must be converted to digital form. EX1003, ¶84. *Kee* provides a detailed description of an audio interface with analog-to-digital conversion capabilities. EX1012-Kee, [0027]. Thus, a PHOSITA would have been motivated to look to a reference such as *Kee* as a guide providing one option (an audio interface) for converting analog that input audio to digital form. EX1003, ¶84; *KSR Int'l Co. v. Teleflex Inc*., 550 U.S. 398, 421 (2007) (motivation exists "[w]hen there is a design need . . . and there are a finite number of identified, predictable solutions" within a PHOSITA's "technical grasp"); *see also In re Translogic Technology, Inc*., 504 F.3d 1249, 1262 (Fed. Cir 2007) ("[I]n looking for a multiplexer circuit for the individual 2:1 multiplexers disclosed in Gorai, a person of ordinary skill in the art would have solved this design need by 'pursu[ing] known options within his or her technical grasp.'"). Further, a PHOSITA would have understood that the use of an audio interface would have provided *Nguyen's* system with not only the means to receive audio data and digitize it, but

the means to convert processed or generated audio data back into analog form so that the processed audio is able to be transmitted to the output device. EX1003, ¶85. Moreover, the use of an audio interface was well-known in the art long before the '756 Patent. EX1003, ¶86.

Second, such a combination would have constituted combining prior art elements (e.g., *Kee's* audio interface and *Nguyen's* accent conversion system) according to known methods according to known methods to yield predictable results (e.g., enabling *Nguyen's* accent conversion system to convert analog microphone signals to digital signals for subsequent processing). *KSR Int'l Co. v. Teleflex Inc*., 550 U.S. 398, 417-418 (2007); EX1003, ¶87. The result of this combination, an accent conversion system capable of analog-to-digital and digital-to-analog conversion, facilitates the functionality described by *Nguyen*. A PHOSITA would have understood that computer audio processing is performed on digital data, but that audio is received and output in analog. Thus, *Kee's* audio interface effectuates *Nguyen's* accent conversion system by providing a means to convert audio data between analog and digital formats. EX1003, ¶87.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because both *Nguyen* and *Kee* teach similar electronic devices that use processors, and the implementation of audio interfaces (and particularly analog-to-digital and digital-to-analog converters) was well known to

PHOSITAs. EX1004-Nguyen, [0071]-[0072], Fig. 8; EX1012-Kee, Title, [0002]. Thus, a PHOSITA would have understood that *Nguyen's* system is capable of incorporating *Kee's* audio interface. EX1003, ¶88. A PHOSITA would have understood that only minor modifications to the hardware and software would have been required to implement the audio interface, as taught by *Kee*, into the accent conversion system, as taught by *Nguyen.* EX1003, ¶88.

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

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

a) *Nguyen's* Teachings

Nguyen teaches performing accent conversion ("AC") on communications between at least two users. EX1004-Nguyen, [0016], [0068]-[0070], [0072]. As established above in Claim 1[Pre], the *Nguyen-Kee* combination system renders

obvious the use of an audio interface. *Nguyen* teaches that "[t]he video conferencing application 350 receives audio and video data from a microphone 336 and a camera 334, respectively, connected to the client device 330." EX1004-*Nguyen*, [0071]. A PHOSITA would have understood that the microphone and audio interface, which convert the signals from analog to digital, would have worked together to allow the device and video conferencing application to receive audio data. EX1003, ¶91.

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

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

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

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

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

b) *Dirac's* Teachings

Dirac teaches an accent translation system for translating a party's speech in a first accent to a second accent. Specifically, *Dirac* teaches that the "accent translation system may determine that the first party speaks with the first accent[.]" EX1005-Dirac, 3:12-14. "[T]he accent translation system may also determine that the second party speaks with the second accent" and "may then translate the first party's speech from the first accent to the second accent, for example by applying a respective accent translation model for translating speech from the first accent to the second accent." EX1005-Dirac, 3:26-35.

Dirac teaches *extract[ing] accent features* because *Dirac* obtains audio characteristics from one party's speech to analyze and detect which accent is spoken.

Specifically, *Dirac* discloses that "an audio characteristic analysis of audio communications by the first party may be used to ***determine the first party's accent***." EX1005-Dirac, 8:53-55. To determine the accent, "audio communications from the first party may be analyzed to obtain audio characteristics (e.g., pitch, tone, stress, melody, etc.) of the first party's accent." EX1005-Dirac, 8:59-61. Audio characteristics analysis techniques, such as fast Fourier transform (FFT) and mel-frequency cepstral coefficients (MFCC), can be used to obtain the audio characteristics. EX1005-Dirac, 8:62-64. "[T]he audio characteristics of the first party's accent may then be compared to audio characteristics of various accent sample sets … to determine an accent set and an accent to which the first party's accent substantially correlates." EX1005-Dirac, 8:64-9:2.

Dirac teaches *appl[ing] one or more trained machine learning models* to perform the claimed *extract[ion]* because *Dirac* discloses: (1) that its translation model is determined "by one or more artificial neural networks" (EX1005-Dirac, 22:9-11); and (2) that "one or more artificial neural networks … may efficiently parse and analyze large amounts of input data" including "new audio samples" (EX1005-Dirac, 7:52-60). EX1003, ¶98 (explaining how *Dirac* teaches using one or more neural networks in its accent translation system to analyze the input audio communications to obtain the audio characteristics of the first party's accent and

determine which accent model it correlates with); EX1005-Dirac, 8:58-67 (describing analysis of audio to determine a party's accent).

Thus, *Dirac* teaches *appl[ing] one or more trained machine learning models to first input audio data … to extract accent features of first input speech associated with a first accent of a first user*, where the *first input speech associated with the first accent of a first user* is the speech from *Nguyen*'s exemplary first user with a target accent.

c) Motivation to Combine

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

A PHOSITA would have been motivated to incorporate *Dirac's* accent detection and feature extraction process into *Nguyen-Kee's* accent conversion system with a reasonable expectation of success for several reasons. EX1003, ¶101.

First, *Nguyen* explicitly contemplates performing accent detection on an audio stream but does not provide details on how this functionality is implemented. EX1003, ¶102 (discussing *Nguyen* at [0090]). Because *Nguyen's* accent conversion requires determining the accent of a participant to apply the correct accent

conversion model, a PHOSITA would have been motivated to look to other references, such as *Dirac*, for detailed processes to do so. EX1003, ¶103. *Dirac* is one example that provides a detailed description of such an accent detection process, specifically for the speech of a participant. EX1003, ¶104; EX1005-Dirac, 8:53-9:2, 9:32-38. Thus, a PHOSITA would have been motivated to make the proposed combination, i.e., to look to *Dirac* as a guide for implementing accent detection in *Nguyen's* device. EX1003, ¶104.

Further, as established above, *Dirac* teaches the use of a machine learning model to analyze the input audio data. EX1005-Dirac, 7:52-56. A PHOSITA would have appreciated that machine learning models are advantageously used for accent detection because machine learning models are able to make accurate predictions, such as the accent of speakers, from large amounts of input data, such as an input audio stream during live conferences. EX1003, ¶105. For example, a PHOSITA would have understood that training and utilizing a machine learning model would have improved the accent detection process by creating a robust and accurate prediction model. EX1003, ¶105.

Third, such a combination would have constituted applying a known technique (e.g., *Dirac's* accent recognition) to a known device (e.g., *Nguyen's* accent conversion system) ready for improvement to yield predictable results (e.g., improved, robust, and accurate accent recognition for subsequent accent

conversion). *KSR Int'l Co. v. Teleflex Inc*., 550 U.S. 398, 417-418 (2007). Additionally, accent detection using a neural network was well known in the art well before the '756 Patent. EX1003, ¶106.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because *Nguyen* and *Dirac* both teach the use of accent detection in an accent conversion process. EX1003, ¶107; EX1004-Nguyen, [0095]; EX1005-Dirac, 8:53-61. Further, both teach the use of neural networks in their respective accent conversion systems. EX1004-Nguyen, [0013]; EX1005-Dirac, 7:52-56. Thus, a PHOSITA would have understood that *Nguyen's* system is capable of incorporating *Dirac*'s neural network to implement accent characteristic extraction and accent detection. EX1003, ¶107. A PHOSITA would have understood that the combination would have only required minor modifications to the software of *Nguyen*. EX1003, ¶107.

In the *Nguyen-Kee-Dirac* combination, *Nguyen's* system receives an audio stream through the microphone and *Kee's* audio interface during a virtual conference. The accent conversion functionality of the client device, as taught by *Nguyen,* receives audio in the form of speech from a first participant via the microphone, which is then digitized by the client device's audio interface, as taught by *Kee*. The accent conversion functionality of the client device receives audio in the form of speech from a second participant via a network. The accent features of

the first participant (a *first user*) and the second participant are extracted utilizing the neural network(s) of *Dirac*. The neural network(s) compare these characteristics to sample sets of accents accessible by the model to determine which accent correlates with that of the first participant and which correlates to the second participant's accent, as taught by *Dirac*. This allows the system of *Nguyen* to select the relevant AC model to perform a conversion from the accent of the second participant to that of the first participant.

In summary, *Nguyen-Kee-Dirac* teaches applying one or more neural networks to the speech of a first participant having a target accent to extract the accent features of the first participant's accent to determine the accent. Thus, *Nguyen-Kee-Dirac* teaches *appl[ing] one or more trained machine learning models to first input audio data obtained via the microphone and the audio interface to extract accent features of first input speech associated with a first accent of a first user.*

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

Nguyen in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *analyz[ing] obtained second input audio data associated with second input speech associated*

with a second accent of a second user (e.g., *Nguyen's* received audio from its second participant with a source accent) *to generate characteristics specific to a natural voice of the second user* (e.g., *Narayanan's* generating a voice profile)*, wherein the generated characteristics correspond to vocal traits that are distinct to the second user* (e.g., *Narayanan's* voice profile contains characteristics of the speaker's specific speech) *and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns* (e.g., *Narayanan's* voice characteristics, including intonation), as claimed.

As explained above, Petitioner proposes that *characteristics specific to a natural voice of the second user* should be construed as "speaker-specific attributes that are globally representative of the second user's voice and that are used to maintain the second user's voice throughout synthesis." *Supra* Section III.C.

Petitioner notes that the '756 Patent uses "extract" and "generate" interchangeably to describe the process of obtaining *characteristics specific to a natural voice of the second user*. *Compare* EX1001, 7:44-48, 8:56-64; *with* EX1001, 5:66-6:1, 6:19-23.

a) *Nguyen's* Teachings

Nguyen teaches that after the accents of the first and second participants have been identified, the first participant at a client device is provided an indication that accent conversion for the second participant's audio stream is available, so that the

second participant's accent is converted to that of the first participant. EX1004-Nguyen, [0095] (describing that "the client device 330 itself could make such a determination [(accent mismatch)] and provide the suggestion to the user" that accent conversion is available). If the first participant accepts, the client software in the first participant's client device performs accent conversion on the second participant's speech. EX1004-Nguyen, [0095] ("If the user accepts the suggestion 544, … the client software 350 may enable its accent conversion functionality 352 to perform accent conversion.").

Nguyen generally teaches that the speech input into the accent conversion system comprises of two parts besides content—the accent and the identity characteristics. *See* EX1004-Nguyen, [0013], [0016], [0072]-[0073]; EX1003, ¶114. *Nguyen* teaches that its system retains identity characteristics through accent conversion but doesn't expressly teach that the identity characteristics are *generate[d]*, as claimed. EX1004-Nguyen, [0072]. *Nguyen* discusses that identity characteristics are those "such as timbre and pitch[.]" EX1004-Nguyen, [0073]; EX1001, 4:19-22 (the claimed *characteristics* include "prosodic features (e.g., pitch and timbre)"). A PHOSITA would have understood that identity characteristics identify a unique speaker and are therefore specific to a natural voice of a user and distinct to the user. EX1003, ¶114. Further, timbre and pitch are prosodic features. EX1001, 4:19-22.

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

b) *Narayanan's* Teachings

Narayanan teaches "[s]ystems, methods, and devices relating to audio correction." EX1006-Narayanan, Abstract. *Narayanan* discloses a method and system for real-time audio correction for mispronounced or incorrect words. EX1006-Narayanan, [0001], [0004]. This method includes the creation of a voice profile to retain the characteristics of the speaker through the audio correction process. EX1006-Narayanan, [0070]-[0075], Fig. 5. *Narayanan* explains its "***voice profile may be determined a single time*** and applied" across various portions of speech. EX1006-Narayanan, [0035], Fig. 2.

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

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

> The voice profile may be ***generally representative of the speaker's speech*** such that the voice profile may be used to generate spoken audio content. ***The voice profile may indicate various characteristics of the speaker's speech, such as audio, vocal, and/or linguistics characteristics***. Example audio, vocal, and/or linguistic characteristics may include audio spectral patterns, vocal frequency, ***vocal pitch***, speaking speed, ***intonation***, loudness, amplitude, speech patterns, speech cadence, the number and/or duration of utterances, the number and/or duration of breaks between utterances, etc. Accordingly, the voice profile may be determined based on similar

> characteristics of the spoken audio content and/or the first one or more words spoken by the speaker.

EX1006-Narayanan, [0075]. *Narayanan* specifies that these characteristics include intonation, which a PHOSITA would have recognized as an *intonation pattern*. EX1006-Narayanan, [0075]; EX1003, ¶118. A PHOSITA would have understood that the characteristics included in the voice profile, such as intonation and vocal pitch, are *specific to the natural voice of the [] user*. EX1003, ¶119.

A PHOSITA would have understood that the characteristics contained within the voice profile are *generated* from the spoken audio data. EX1003, ¶120. Thus, *Narayanan* discloses *analyz[ing] obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics correspond to vocal traits that are distinct to the second user and comprise one or more of a voice quality or one or more phonetic patterns, prosodic features, articulation styles, or intonation patterns*.

And *Narayanan's* teachings also teach the claimed *characteristics specific to a natural voice of the second user* under Petitioner's proposed construction of "speaker-specific attributes that are globally representative of the second user's voice and that are used to maintain the second user's voice throughout synthesis" because *Narayanan* teaches that its "voice profile may be determined a single time

and applied" across various portions of speech, and its voice profile is "generally representative of the speaker's speech." EX1006-Narayanan, [0035], Fig. 2, [0075]; EX1003, ¶120 (further citing EX1007-Wang and explaining how a PHOSITA would have understood how speaker specific information is stored and used throughout the speech modification process). Although *Narayanan* allows for profile updates, those updates refine a persistent profile rather than replacing it with frame-level extraction. *Id.* Accordingly, *Narayanan's* teachings of deriving a voice profile that is determined once and applied across multiple portions of speech demonstrates that *Narayanan's* voice profile characteristics are not frame-specific, but instead globally representative of the speaker's voice. *Id.*

c) Motivation to Combine

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

A PHOSITA would have been motivated to incorporate *Narayanan's* voice profile generation and functionality into *Nguyen-Kee-Dirac's* accent conversion system with a reasonable expectation of success, for several reasons. EX1003, ¶122.

First, *Nguyen* specifically states that to alleviate communication issues, its system converts one accent to another accent while retaining the identity

characteristics of the original speaker. EX1004-Nguyen, [0013]. However, *Nguyen* does not provide a comprehensive list of its identity characteristics to preserve. EX1003, ¶123. Accordingly, a PHOSITA would have been encouraged and motivated to looked to other references, such as *Narayanan* and its use of voice profiles for details regarding additional identity characteristics as candidates for preservation of the speaker's identity. EX1003, ¶123. Further, the use of voice profiles in speech processing was well known in the art before the '756 Patent. EX1003, ¶123.

A PHOSITA would have also been motivated to make the proposed combination because such a combination would have improved *Nguyen's* system by allowing it to retain vocal characteristics, beyond pitch and timbre, for each participant, by way of the voice profile, during accent conversion. EX1003, ¶124; EX1006-Narayanan, [0018]. This would have allowed the output of the accent conversion system to retain more identity characteristics of the second participant, improving the output of the accent conversion system by generating output audio that is closer to the voice of the second participant. EX1003, ¶124. A PHOSITA would have appreciated that such a combination would have furthered *Nguyen*'s stated goal of changing a speaker's accent "but not their voice[.]" EX1004-Nguyen, [0016]; *see* EX1004-Nguyen, [0017]; EX1003, ¶124.

Additionally, *Narayanan* teaches that its voice profile is stored and later updated based on further speech of a speaker. EX1006-Narayanan, [0047]. A PHOSITA would have understood that this would have provided an additional advantage to *Nguyen*'s system, allowing the voice profile and subsequent speech modification to improve over time. EX1003, ¶125. A PHOSITA would have understood that the more speech data that is analyzed to generate speech characteristics, the more accurate the voice characteristics will be in representing a user's voice. EX1003, ¶125; EX1005-Dirac, 7:56-60. As a result, an accent conversion or speech modification performed on the user's speech will retain more accurate characteristics of the user's voice, improving the similarity of the output to the user's voice. EX1003, ¶125.

Further, such a combination would have constituted applying a known technique (e.g., *Narayanan's* generating voice profiles for speakers) to a known device (e.g., *Nguyen's* accent conversion system) ready for improvement to yield predictable results (e.g., using voice profiles to retain the specific identity characteristics of a speaker for representation of the speaker's true voice during accent conversion). *KSR*, 550 U.S. at 417-418.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because *Nguyen* already teaches retaining identity characteristics of the speaker during accent conversion, while *Narayanan* provides

further implementation details of generating a voice profile containing the speaker characteristics to retain those characteristics in speech modification. EX1003, ¶127; EX1004-Nguyen, [0013]; EX1006-Narayanan, [0004], [0075]. A PHOSITA would have understood that only minor changes to the software of *Nguyen* are required to generate Narayanan's voice profiles to retain the identity characteristics of the speaker. EX1003, ¶127. Particularly, because *Nguyen* already teaches the use of profiles that include accent information, a PHOSITA would have understood how to modify these profiles to be *Narayanan's* voice profiles. EX1004-Nguyen, [0090], [0095], [0108]; EX1003, ¶127.

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

data to generate characteristics corresponding to the vocal traits of a second user contained in a voice profile.

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

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

As discussed above, the combination of *Nguyen*, *Kee*, and *Dirac* teaches extracting accent features of a first input audio data associated with a first accent of a first participant as part of an accent detection process. *See* Claim 1[a]. The further combination of *Nguyen*, *Kee*, *Dirac*, and *Narayanan* teaches generating characteristics, stored in a voice profile, corresponding to vocal traits of a second

participant by analyzing a second input audio. *See* Claim 1[b]. As explained below, the combination further teaches generating *a modified version of the second input speech* that is based on *the extracted accent features* of the *first user* and *the generated characteristics* of the *second user*.

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

A PHOSITA would have understood that an example of this functionality, consistent with the example provided in the mappings for Claims 1[a] and 1[b], is as

follows: A first and second participant speak during a video conference. EX1003, ¶133; EX1004-Nguyen, [0090], [0095]; *see* Claim 1[a]. The accent conversion system of *Nguyen* detects the participants' accents based on the accent characteristics of the participants' respective speech. EX1003, ¶133; EX1004-Nguyen, [0090], [0095]; *see* Claim 1[a]. The accent conversion model is then selected by the accent conversion system based on the detected accent of each participant. EX1003, ¶133; EX1004-Nguyen, [0108], [0065]. The audio stream of the second participant is modified by the accent conversion model to produce a second audio stream containing the identity characteristics of the second participant but with the accent of the first participant. EX1003, ¶133; EX1004-Nguyen, [0072], [0013], [0016]; *see* Claim 1[b].

The only differences between *Nguyen* and the *Nguyen-Kee-Dirac-Narayanan* combination are that the accent detection is performed using the method of *Dirac,* and identity characteristics are maintained by creating a voice profile as taught by *Narayanan*. Accordingly, a PHOSITA would have understood that in the combination, the second audio stream would have been *based on* the generated voice profile (containing the identity characteristics) and the extracted accent features, which are included within the AC model as the output accent. EX1003, ¶134. Specifically, because the AC model is selected based on the extracted accent features, the output of the AC model is *based on* the extracted accent features.

EX1003, ¶134. Additionally, the output audio stream has speech with identity characteristics that are retained based on the identity characteristics within the voice profile and mimics the target accent. EX1003, ¶134. Thus, *Nguyen-Kee-Dirac-Narayanan* teaches *synthesiz[ing] a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent*.

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

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

As established above, *Nguyen* teaches that a second audio stream containing modified speech is generated by the system. *See* Claim 1[c]. *Nguyen* further teaches that after accent conversion "[t]he generated output audio stream(s) can then be provided to the requesting participants instead of the original unaltered audio stream." EX1004-Nguyen, [0070]. *Nguyen* teaches "transmit[ting] the second audio

stream to speakers or headphones to be output to the user." EX1004-Nguyen, [0115]. As established in Claim 1[Pre], the *Nguyen-Kee* combination teaches an audio interface that has analog-to-digital and digital-to-analog conversion functionalities. *See* Claim 1[Pre]. A PHOSITA would have understood that in such a combination, the processed audio stream would have been provided to the audio interface to be converted to analog, allowing it to be output to the user. EX1003, ¶137. Further, a PHOSITA would have understood that the generated second audio stream (*output audio data*) contains the modified speech having the target accent and therefore *is generated based on the modified version of the second input speech*. EX1003, ¶137.

B. Claim 2

1. Claim 2: "The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics."

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

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

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

C. Claim 3

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

Nguyen in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *wherein the accent features* (e.g., *Dirac’s* audio characteristics of an accent) *comprise one or more pitch contours* (e.g., *Dirac’s* melody)*, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the*

first input speech (e.g., *Dirac's* melody throughout the speech)*, the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent*, as claimed.

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

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

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

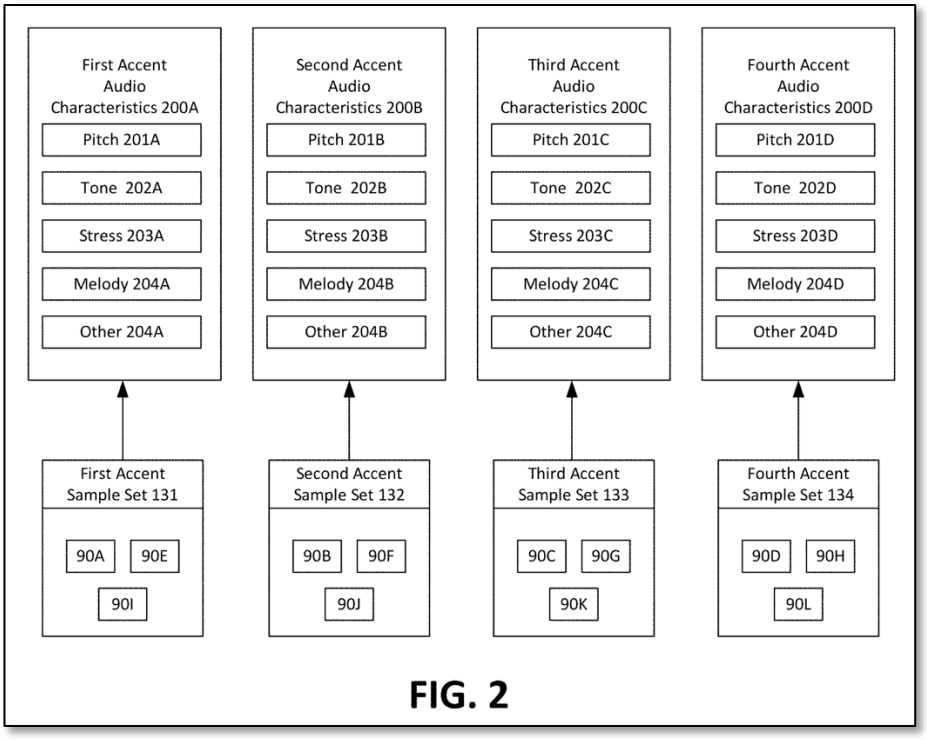


FIG. 2

EX1005-Dirac, Fig. 2. As shown, these characteristics include melody. *Id.* A PHOSITA would have understood that melody is the rise and fall of pitch in a communication. EX1003, ¶146. Accordingly, *Nguyen* in view of *Kee*, *Dirac*, and *Narayanan* renders obvious *wherein the accent features comprise one or more pitch contours … and the pitch contours comprise variations in pitch throughout the first input speech.*

D. Claim 6

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

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

Nguyen teaches its system may be a single client device 330 (*speech processing system*). EX1004-Nguyen, Fig. 3B, [0071]-[0072]. *Nguyen* further teaches accent conversion "functionality 352 can operate on an audio stream received from the microphone 336 or it can ***operate on one or more audio streams received from other participants in a virtual conference***." EX1004-Nguyen, [0072]. A PHOSITA would have understood that the audio stream received from other participants, such as a second participant, would have been *captured at* a

microphone of said participant's device. EX1003, ¶149. *Nguyen* discloses that each client device transmits video and audio streams to other client devices through the virtual conference provider 310 via network 320 and receives streams from other client devices through the virtual conference provider via the network. EX1004-Nguyen, [0071], Fig. 3A. These other client devices are *remote* from the first client device. EX1003, ¶149; EX1004-Nguyen, [0072]. And because the second input audio data (the audio which is converted to a target accent) is received by the first client device from another participant's remote device, *Nguyen* therefore teaches *receiv[ing] the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device*. EX1003, ¶150.

VI. GROUND 2: NGUYEN IN VIEW OF KEE, DIRAC, NARAYANAN, AND KAMRUZZAMAN RENDERS OBVIOUS CLAIM 4 OF THE '756 PATENT

A. Claim 4

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

Nguyen in view of *Kee, Dirac, Narayanan*, and *Kamruzzaman* renders obvious *wherein the one or more processors* (e.g., *Nguyen's* processors) *are further*

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

The '756 Patent explains that "[a]n MFCC analysis extracts a unique fingerprint." EX1001, 7:62.

a) *Kamruzzaman's* Teachings

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

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

EX1011-Kamruzzaman, 1-2.

Fig. 1 is "[a] block diagram of the structure of an MFCC processor" which a PHOSITA would have understood outlines the process for performing an MFCC analysis. EX1011-Kamruzzaman, 2; EX1003, ¶155.

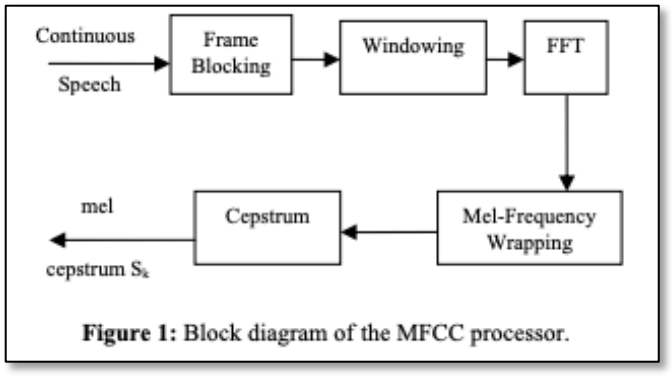


Figure 1: Block diagram of the MFCC processor.

EX1011-Kamruzzaman, Fig. 1.

Kamruzzaman discloses that by performing the steps of Fig. 1, "a set of mel-frequency cepstrum coefficients is computed." EX1011-Kamruzzaman, 2. "This set of coefficients is called an acoustic vector," which is "used to represent and recognize the voice characteristic of the speaker." EX1011-Kamruzzaman, 2. A PHOSITA would have understood that the acoustic vector representation of the speaker's voice characteristics contains speaker identifying information. EX1003, ¶156; *see* EX1001, 7:62.

Further, because *Kamruzzaman* describes using a Support Vector Machine (SVM) which uses the variability from its MFCCs, *Kamaruzzaman* teaches *(MFCC) analysis to extract a unique fingerprint*, as claimed. EX1011-Kamruzzaman, 1-2; EX1003, ¶156 (explaining how a PHOSITA would have understood

Kamruzzaman's MFCC technique to capture speaker-specific information in order to construct a global support vector, such as for use in *Narayanan*'s voice profile).

b) Motivation to Combine

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

A PHOSITA would have been motivated to incorporate *Kamruzzaman's* MFCC analysis into the *Nguyen-Kee-Dirac-Narayanan* system's voice profile with a reasonable expectation of success for several reasons. EX1003, ¶158.

First, *Nguyen* specifically states that to alleviate communication issues, the system converts one accent to another accent while retaining the identity characteristics of the original speaker. EX1004-Nguyen, [0013]. However, *Nguyen* does not provide a detailed explanation of what characteristics are identity characteristics, nor does *Nguyen* provide a list of characteristics to preserve. EX1003, ¶159. While *Narayanan* provides characteristics, such as audio spectral patterns and intonation, to be included in a speaker's voice profile, *Narayanan* does not provide specifics for how the various traits are determined and represented. EX1003, ¶159. Accordingly, and based on these references' disclosures, a PHOSITA would have been motivated to look to an additional reference, such as

Kamruzzaman, for specific implementation details of obtaining features/characteristics from a speech signal to incorporate into the voice profile. EX1011-Kamruzzaman, 1. *Kamruzzaman* is one example which provides that, and explains how MFCCs represent the voice characteristics of a speaker. EX1011-Kamruzzaman, 2. Thus, a PHOSITA would have looked to *Kamruzzaman* for specifics on how features are determined and represented. EX1011-Kamruzzaman, 2; EX1003, ¶159. Use of MFCCs to represent the voice characteristics of a speaker was also well known in the art at the time of the '756 Patent. EX1003, ¶159.

Second, such a combination would have constituted applying a known technique (e.g., *Kamruzzaman's* use of MFCCs) to a known device (e.g., *Nguyen-Narayanan's* voice profiles) ready for improvement to yield predictable results (e.g., extracting improved identity characteristics). *KSR*, 550 U.S. at 417-418. Because MFCCs are based on the human ear's ability to hear certain frequencies, a PHOSITA would have understood that providing the voice profile with data that represents the voice characteristics of the speaker based on humans' frequency perception would have improved the performance of the accent conversion model. EX1003, ¶160; EX1011-Kamruzzaman, 2. Retaining voice characteristics perceptible to a human would have improved the resulting output audio with a modified accent to sound more like the original speaker. EX1003, ¶160.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because the *Nguyen-Kee-Dirac-Narayanan* combination already teaches generating characteristics stored in the voice profile, and *Kamruzzaman* simply provides an additional, and well-known, method to generate characteristics, which a PHOSITA would have readily been able to implement. EX1003, ¶161. A PHOSITA would have understood that it would have required only minor changes to the software of the *Nguyen-Kee-Dirac-Narayanan* combination to implement an MFCC analysis to generate characteristics for the voice profile. EX1003, ¶161.

VII. GROUND 3: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND WANG RENDERS OBVIOUS CLAIM 5 OF THE '756 PATENT

A. Claim 5

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

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

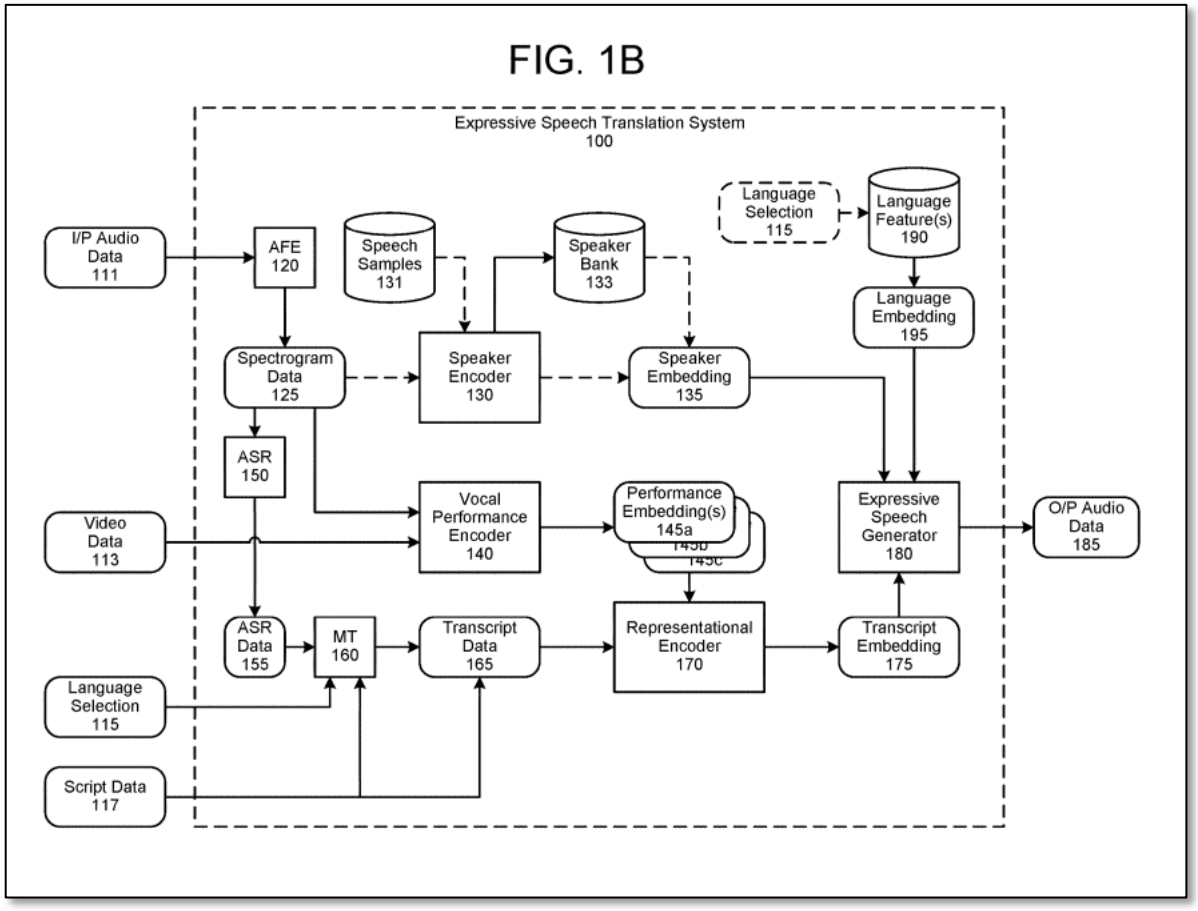
Wang's speaker encoder to generate speaker embedding data)*, wherein the generated characteristics comprise the speaker-specific voice characteristics* (e.g., *Narayanan's* voice profile comprises the speaker embedding data generated from applying *Wang's* speaker encoder to *Narayanan's* characteristics of the speaker's specific speech), as claimed.

The '756 Patent explains that the speaker-specific voice characteristics may include pitch, timbre, and/or rhythm. EX1001, 6:35-36. *Narayanan* teaches that the voice characteristics stored in the voice profile "may include audio spectral patterns, vocal frequency, ***vocal pitch***, speaking speed, ***intonation***, loudness, amplitude, ***speech patterns, speech cadence***, the number and/or duration of utterances, the number and/or duration of breaks between utterances, etc." Thus, *Narayanan's* characteristics of the speaker's specific speech stored in the voice profile are *speaker-specific voice characteristics* and are part of the *generated characteristics*. *Narayanan* does not expressly teach *apply[ing] a speaker identity encoding technique to encode speaker-specific voice characteristics*.

a) *Wang's* Teachings

Wang teaches "[a]n expressive speech translation system [that] may process source speech in a source language and output synthesized speech in a target language while retaining vocal performance characteristics such as intonation,

emphasis, rhythm, style, and/or emotion." EX1007-Wang, Abstract. Fig. 1B depicts the expressive speech translation system.



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

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

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

EX1007-Wang, [0030]. The voice identity characteristics of *Wang* overlap with the characteristics described in *Narayanan*. EX1003, ¶167 (comparing EX1007-Wang, [0030] to EX1006-Narayanan, [0075]). Like the characteristics of *Narayanan*, which are "generally representative of the speaker's speech[,]" the voice identity characteristics of *Wang*, which "differ from speaker to speaker," are vocal traits that are distinct to a speaker and represent speaker-specific voice characteristics. EX1003, ¶167; *See* Claim 1[b].

Wang also teaches "[t]he expressive speech translation system 100 may include a speaker bank 133 that stores speaker embedding data 135a corresponding to different voice identities." EX1007-Wang, [0030].

Finally, *Wang* teaches that "[u]sing the speaker embedding data 135 … an expressive speech generator 180 may synthesize speech that reflects the speech characteristics of the target language while reproducing the original speaker's expressiveness and preserving their voice identity characteristics." EX1007-Wang, [0047].

b) Motivation to Combine

In the *Nguyen-Kee-Dirac-Narayanan-Wang* combination, the system encodes the voice identity characteristics of the second speaker taught by *Narayanan* into a speaker embedding, as taught by *Wang*. In such a combination, the speaker embedding data is included in the voice profile of the second user, as taught by *Narayanan.*

A PHOSITA would have been motivated to incorporate *Wang's* technique of encoding the speaker identity characteristics into a speaker embedding into the *Nguyen-Kee-Dirac-Narayanan* system's voice profile with a reasonable expectation of success for several reasons. EX1003, ¶172.

First, while *Narayanan* provides specific identity characteristics as candidates for preservation, a PHOSITA would have understood that in order for the identity characteristics to be usable by a machine learning model, the identity characteristics must be first encoded in a usable medium. EX1003, ¶173. Accordingly, a PHOSITA would have looked to a reference, such as *Wang*, which provides a method for

converting the identity characteristics into embeddings usable by a machine learning model. EX1003, ¶173. *Wang* teaches an encoder that encodes the identity characteristics into a speaker embedding that is usable by the processing system to preserve those characteristics when synthesizing data. EX1007-Wang, [0030], [0049].

Second, such a combination would have constituted applying a known technique (e.g., *Wang's* technique for encoding identity characteristics into a speaker embedding) to a known method (e.g., *Nguyen-Kee-Dirac-Narayanan's* identity characteristics contained in the voice profiles) ready for improvement to yield predictable results (e.g., *Nguyen-Kee-Dirac-Narayanan's* accent translation system would improve in its ability to process and preserve the identity characteristics contained in the voice profile via their encoding into a speaker embedding). *KSR*, 550 U.S. at 417-418. Further, using encoding to convert data to a form usable by machine learning models was well known in the art at the time of the '756 Patent. EX1003, ¶174.

Finally, a PHOSITA would have had a reasonable expectation of success in making the combination because the *Nguyen-Kee-Dirac-Narayanan* combination already teaches generating and preserving characteristics stored in the voice profile using a machine learning model, and *Wang* simply provides a well-known way to prepare the characteristics for processing by a machine learning model. EX1003,

¶175. A PHOSITA would have understood that it would have required only minor changes to the software of the *Nguyen-Kee-Dirac-Narayanan* combination to encode the identity characteristics into a speaker embedding, as taught by *Wang*, and store the speaker embedding in a voice profile. EX1003, ¶175.

VIII. GROUND 4: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND KAMRUZZAMAN RENDERS OBVIOUS CLAIMS 7-11 AND 13 OF THE '756 PATENT

Ground 4 relies on the same prior art references as Ground 2, except that the claims of Ground 4 do not recite an "audio interface," and therefore do not rely on *Kee*.

A. Claim 7

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

To the extent the preamble is limiting, *Nguyen* discloses, or at least renders obvious, *a method for real-time accent mimicking* (e.g., *Nguyen's* accent conversion method)*, the method implemented by a speech processing system* (e.g., *Nguyen's* system for accent conversion for virtual conferences), as claimed.

As established in Claim 1[Pre], *Nguyen* discloses a speech processing system. *See* Claim 1[Pre]. *Nguyen* teaches a *method* for accent conversion to convert a second participant's accent to a first participant's accent (e.g., *accent mimicking*) for virtual conferences that is implemented by the *speech processing system*. EX1004-Nguyen, Title, Abstract; *see* Claim 1[Pre]. *Nguyen* also teaches that during a

conference, participants communicate with each other and the virtual conference provider. EX1004-Nguyen, [0025], [0027]. This communication happens over a "***real-time*** media server." EX1004-Nguyen, [0052]-[0054]. *Nguyen* further teaches that the accent conversion process receives "audio streams from the virtual conference and each generates an output audio stream having the corresponding target accent." EX1004-Nguyen, [0070]. *Nguyen* teaches that "[t]he generated output audio stream(s) can then be provided to the requesting participants instead of the original unaltered audio stream." EX1004-Nguyen, [0070]. Because the system is performing the accent conversion during a virtual conference, over a real-time media server, a PHOSITA would have understood that *Nguyen's* accent conversion method occurs in *real-time*. EX1003, ¶178.

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

See Claim 1[a].

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

See Claim 4.

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

See Claim 1[c].

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

See Claim 1[d].

B. Claim 8

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

See Claims 1[c] and 7.

C. Claim 9

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

See Claims 2 and 7.

D. Claim 10

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

See Claims 3 and 7.

E. Claim 11

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

See Claims 4 and 7.

F. Claim 13

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

See Claims 6 and 7.

IX. GROUND 5: NGUYEN IN VIEW OF DIRAC, NARAYANAN, AND WANG RENDERS OBVIOUS CLAIMS 14-17 AND 19-20 OF THE '756 PATENT

Ground 5 relies on the same prior art references as Ground 3, except that the claims of Ground 5 do not recite an "audio interface," and therefore do not rely on *Kee*.

A. Claim 14

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

To the extent the preamble is limiting, *Nguyen* discloses, or at least renders obvious, *a non-transitory computer-readable medium* (e.g., *Nguyen's* non-transitory computer readable media) *comprising instructions* (e.g., *Nguyen's* processor-executable instructions stored on the media) *that, when executed by at least one*

processor (e.g., *Nguyen's* processor which executes the stored instructions)*, cause the at least one processor* (e.g., *Nguyen's* processor that performs a function according to the instructions) *to*, as claimed.

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

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

See Claim 1[a].

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

See Claim 5.

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

See Claim 1[c].

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

See Claim 1[d].

B. Claim 15

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

See Claims 1[c] and 14.

C. Claim 16

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

See Claims 2 and 14.

D. Claim 17

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

See Claims 3 and 14.

E. Claim 19

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

See Claims 5 and 14.

F. Claim 20

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

See Claims 6 and 14.

X. GROUND 6: NGUYEN IN VIEW OF DIRAC, NARAYANAN, KAMRUZZAMAN, AND WANG RENDERS OBVIOUS CLAIMS 12 AND 18 OF THE '756 PATENT

A. Claim 12

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

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

same reasons and with the same reasonable expectation of success provided in Claims 4 and 5.

B. Claim 18

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

See Claims 4, 12, and 14.

XI. CONCLUSION

For the foregoing reasons, Petitioner respectfully requests *Post Grant* 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

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

A. Real Party-In-Interest

Petitioner is the real party-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 '756 Patent or related patents:

- *Sanas.ai Inc. v. Krisp Technologies, Inc.*, Case No. 3:25-cv-05666-TSH (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.*, IPR2026-00273 (PTAB), challenging U.S. Patent No. 12,125,496
- *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

C. Lead and Back-Up Counsel

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

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

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. 12,417,756 for Claims 1-20

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

Claim Designation	Claim Language
Claim 3	The speech processing system of claim 1, wherein the accent features comprise one or more pitch contours, other intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the other intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.
Claim 4	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.
Claim 5	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to apply a speaker identity encoding technique to encode speaker-specific voice characteristics, wherein the generated characteristics comprise the speaker-specific voice characteristics.
Claim 6	The speech processing system of claim 1, wherein the one or more processors are further configured to execute the instructions to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.
Claim 7(P)	A method for real-time accent mimicking, the method implemented by a speech processing system and comprising:
Claim 7(a)	applying one or more trained machine learning models to first input audio data to extract accent features of first input speech associated with a first accent of a first user;
Claim 7(b)	analyzing obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise a unique fingerprint of the voice of the second user;

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

Claim Designation	Claim Language
Claim 14(b)	analyze obtained second input audio data associated with second input speech associated with a second accent of a second user to generate characteristics specific to a natural voice of the second user, wherein the generated characteristics comprise speaker-specific voice characteristics;
Claim 14(c)	synthesize a modified version of the second input speech by modifying the obtained second input audio data based on the generated characteristics and the extracted accent features; and
Claim 14(d)	provide for output via an audio output device output audio data generated based on the modified version of the second input speech.
Claim 15	The non-transitory computer-readable medium of claim 14, wherein the modified version of the second input speech preserves aspects of the natural voice of the second user and mimics the first accent.
Claim 16	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to extract from the first input audio data one or more prosodic features, linguistic features, or global speaker characteristics.
Claim 17	The non-transitory computer-readable medium of claim 14, wherein the accent features comprise one or more pitch contours, intonation patterns, or phoneme pronunciations and the pitch contours comprise variations in pitch throughout the first input speech, the intonation patterns comprise the rise and fall of pitch at the ends of phrases or sentences, or the phoneme pronunciations comprise a unique production of phonemes in the first accent.
Claim 18	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a mel frequency cepstral coefficient (MFCC) analysis to extract a unique fingerprint of the voice of the second user, wherein the generated characteristics comprise the unique fingerprint.

Claim Designation	Claim Language
Claim 19	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to apply a speaker identity encoding technique to encode the speaker-specific voice characteristics.
Claim 20	The non-transitory computer-readable medium of claim 14, wherein the instructions, when executed by the at least one processor further causes the at least one processor to receive the second input audio data via one or more communication networks and from a user computing device that is remote from the speech processing system, wherein the second input audio data is captured at the user computing device.

APPENDIX OF EXHIBITS

Exhibit 1001	U.S. Patent No. 12,417,756 ("'756 Patent")
Exhibit 1002	Prosecution History of U.S. Patent No. 12,417,756 ("'756 Prosecution History")
Exhibit 1003	Declaration of Christopher Schmandt
Exhibit 1004	U.S. Patent Application Publication No. 2024/0098218 to Nguyen et al. ("*Nguyen*")
Exhibit 1005	U.S. Patent No. 10,163,451 to Dirac et al. ("*Dirac*")
Exhibit 1006	U.S. Patent Application Publication No. 2022/0417659 to Narayanan et al. ("*Narayanan*")
Exhibit 1007	U.S. Patent Application Publication No. 2024/0274122 to Wang et al. ("*Wang*")
Exhibit 1008	B.H. Juang and Lawrence Rabiner, *Automatic Speech Recognition – A Brief History of the Technology Development*, 2004 ("*Juang*")
Exhibit 1009	Li Deng, *Deep learning: from speech recognition to language and multimodal processing*, APSIPA Transactions on Singal and Information Processing, Vol. 5, E. 1, 2016 ("*Deng*")
Exhibit 1010	Vlado Delic, Zoran Peric, Milan Secujski, Niksa Jakovljevic, Jelena Nikolic, Dragisa Miskovic, Nikola Simic, Sinisa Suzic, and Tijana Delic, *Speech Technology Progress Based on New Machine Learning Paradigm*, Computational Intelligence and Neuroscience, 2019 ("*Delic*")
Exhibit 1011	S.M. Kamruzzaman, A.N.M. Rezaul Karim, Saiful Islam, and Emdadul Haque, *Speaker Identification using MFCC-Domain Support Vector Machine*, International Journal of Electrical and Power Engineering, 2007 ("*Kamruzzaman*")
Exhibit 1012	U.S. Patent Application Publication No. 2005/0066209 to Kee et al. ("*Kee*")
Exhibit 1013	C. O'Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallaghan.com/research/papers/Pitch.pdf) September 30, 2015. ("*O'Callaghan*")
Exhibit 1014	Berrak Sisman, Mingyang Zhang and Haizhou Li, *Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27, (2019) ("*Sisman*")
Exhibit 1015	Ivan Chow and Steven Brown, *A Musical Approach to Speech Melody*, Frontiers in Psychology, Vol. 9, 2018 ("*Chow*")

Exhibit 1016	Johan 't Hart, Rene Collier, and Antonie Cohen, *A Perceptual Study of Intonation*, Press Syndicate of the University of Cambridge, ISBN-0521366437, 1990 ("*Hart*")
Exhibit 1017	Elise Piazza, Marius Iordan, and Casy Lew-Williams, *Mothers Consistently Alter their Unique Vocal Fingerprints when Communicating with Infants*, Current Biology, Vol 27, Issue 20, 2017 ("*Piazza*")
Exhibit 1018	U.S. Patent No. 10,614,826 to Huffman et al. ("*Huffman*")
Exhibit 1019	U.S. Patent Application Publication No. 2024/0304200 to Kashet et al. ("*Kashet*")
Exhibit 1020	Constantinos Xydeas, *Differential Encoding Techniques Applied to Speech Signals*, 1978, ("*Xydeas*")
Exhibit 1021	Jaison Joshy and Koj Sambyo, *A Comparison and Contrast of the Various Feature Extraction Techniques in Speaker Recognition*, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 9, No. 11, 2016 ("*Joshy*")
Exhibit 1022	Gaurav Giroti, Tejas Nakhate, Mahesh Laddha, and Prof. Manoj Sarve, *Person Identification through Voice using MFCC and Multi-class SVM*, International Research Journal of Engineering and Technology, Vo. 5, Issue 5, 2018 ("*Giroti*")
Exhibit 1023	Cathy van Eck, *Between Air and Electricity*, Bloomsburry Academic, ISBN-978151327609 PDF, 2017 ("*Eck*")
Exhibit 1024	Ashok Ambardar, *Analog and Digital Signal Processing*, Second Edition, Brooks/Cole Publishing Company, ISBN-9780534954093, 1999 ("*Ambardar*")
Exhibit 1025	Mehrayar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, *Foundations of Machine Learning second edition*, The MIT Press, 2018 ("*Mohri*")
Exhibit 1026	Su-Hyun Han, Ko Woon Kim, Sang Yun Kim and Young Chul Youn, *Artificial Neural Network: Understanding the Basic Concepts without Mathematics*, Dement Neurocogn Disord, 2018 Sep;17(3):83-89 ("*Han*")
Exhibit 1027	Jason Brownlee, *Difference Between Algorithm and Model in Machine Learning* (August 19, 2020), https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/ ("*Brownlee*")
Exhibit 1028	Maurya Vijayaramachandran, and Siddique Afraaz N, *Impact of Hidden Layer in Artificial Neural Networks*, IOSR Journal of

	Engineering, Vol. 10, Issue 11, November 2020, Series -I, 33-38 ("*Siddique*")
Exhibit 1029	Maad M. Mijwel, Adam Esen and Aysar Shamil, *Overview of Neural Networks*, Babylonian Journal of Machine Learning, 1:2, April 2019 ("*Mijwel*")
Exhibit 1030	U.S. Patent No. 8,564,544 to Jobs et al. ("*Jobs*")
Exhibit 1031	Qingchen Zhang, Laurence Yang, Zhikui Chen, and Peng Li, *A Survey on Deep Learning for Big Data*, Information Fusion, Vol. 42, 2018 ("*Zhang*")
Exhibit 1032	Josef Michalek and Jan Vanek, *An Open-Source GPU-Accelerated Feature Extraction Tool*, 12th International Conference on Signal Processing, 2014 ("*Michalek*")
Exhibit 1033	Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schluter, Shuo-yiin Chang, and Tara Sainath, *Deep Learning for Audio Signal Processing*, IEEE Journal of Selected Topics in Signal Processing, Vol. 13, No. 2, 2019 ("*Purwins*")
Exhibit 1034	U.S. Patent Application Publication No. 2021/0082402 to Sun et al. ("*Sun*")
Exhibit 1035	John Holmes and Whendy Holmes, *Speech Synthesis and Recognition*, Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003 ("*Holmes*")
Exhibit 1036	Declaration of June Munford

CERTIFICATION OF WORD COUNT

The undersigned certifies pursuant to 37 C.F.R. §42.24 that the foregoing Petition for *Post Grant* 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 12,490 words according to the word-processing program used to prepare this document (Microsoft Word).

Dated: March 30, 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 30, 2024, a complete and entire copy of this Petition for *Post Grant* Review and Exhibits were provided via Federal Express to the Patent Owner by serving the correspondence address of record for the '756 Patent:

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

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

Michael Ng (Michael.Ng@kobrekim.com)
Daniel Zabeer (Daniel.Zaheer@kobrekim.com)
Jessica Fender (Jessica.Fender@kobrekim.com)
Victoria Fordin (Victoria.Fordin@kobrekim.com)
Zachary Ritz (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
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601

COUNSEL FOR PETITIONER