

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. PGR2026-00032

U.S. Patent No. 12,412,561

DECLARATION OF CHRISTOPHER SCHMANDT

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS.....	10
A.	Educational Background and Professional Experience	10
II.	METHODOLOGY AND MATERIALS CONSIDERED	14
III.	OVERVIEW AND LEGAL STANDARDS	17
A.	Person of Ordinary Skill in the Art.....	18
B.	Obviousness	19
C.	Analogous Art	24
D.	Claim Construction.....	25
IV.	LEVEL OF A PERSON OF ORDINARY SKILL.....	25
V.	OVERVIEW OF TECHNOLOGY	26
A.	Audio Chunks.....	26
B.	Acoustic Features	27
C.	Linguistic Features.....	32
D.	Prior Art Accent Conversion Systems	37
VI.	OVERVIEW OF THE '561 PATENT.....	43
A.	Summary of the '561 Patent.....	43
B.	Field of Endeavor of the '561 Patent.....	44
C.	Problem Purportedly Solved by the Inventors of the '561 Patent.....	45
D.	File History of '561 Patent	46
VII.	OVERVIEW OF THE PRIOR ART	46

A.	Overview of <i>Feinauer</i>	46
B.	Overview of <i>Fan</i>	49
C.	Overview of <i>Prabhavalkar</i>	53
D.	Overview of <i>Sargsyan</i>	55
E.	Overview of <i>Kalinli</i>	56
F.	Overview of <i>Watts</i>	57
VIII.	SUMMARY OF UNPATENTABILITY	57
IX.	OPINIONS REGARDING GROUND 1: CLAIMS 1-2, 6-7, 11-12, and 16-18 ARE OBVIOUS OVER <i>Feinauer</i> , <i>Prabhavalkar</i> , and <i>Fan</i>	58
A.	Claim 1	58
1.	Claim 1(Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”	58
2.	Claim 1(a)(1): “extracting, from a chunk of a stream of chunks of an input audio signal”	62
3.	Claim 1(a)(2): “extracting, ... acoustic features”	71
4.	Claim 1(a)(3): “extracting...linguistic features, wherein the linguistic features have a reduced accent;”	77
5.	Claim 1(b): “synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and”	84
6.	Claim 1(c): “generating, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal.” ..	101
B.	Claim 2	116
1.	Claim 2: “The method of claim 1, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the	

input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.” 116

C. Claim 6 118

1. Claim 6 (Pre): “A computing system, comprising an input device, a vocoder, memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:” 118
2. Claim 6 (a): “extract, from a chunk of a stream of chunks of an input audio signal obtained via the input device, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;” 119
3. Claim 6 (b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;” 119
4. Claim 6 (c): “and generate, via the vocoder and based on the spectrum representation, an output chunk of an output audio signal.” 119

D. Claim 7 119

1. Claim 7: “The computing system for real-time correction of claim 6, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.” 119

E. Claim 11 120

1.	Claim 11(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”	120
2.	Claim 11(a): “extract, from a chunk of a stream of chunks of an input audio signal, linguistic features and acoustic features, wherein the linguistic features have a reduced accent;”	120
3.	Claim 11(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”	120
4.	Claim 11(c): “generate, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal;” ..	121
5.	Claim 11(d): “and provide the output audio signal for output via an auditory channel of an output device.”	121
F.	Claim 12	122
1.	Claim 12: “The non-transitory computer-readable storage medium of claim 11, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”	122
G.	Claim 16	122
1.	Claim 16 (Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”	122
2.	Claim 16 (a): “dividing a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks	

	including a pre-defined number of frames of the speech audio signal;”	122
3.	Claim 16 (b): “extracting, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”	123
4.	Claim 16 (c): “obtaining a speaker embedding for a human speaker; and”	131
5.	Claim 16 (f): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”	131
H.	Claim 17	131
1.	Claim 17(Pre): “A computing system, comprising memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”	131
2.	Claim 17(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”	131
3.	Claim 17(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”	132
4.	Claim 17(c): “obtain a speaker embedding for a human speaker; and generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”	132
I.	Claim 18	132
1.	Claim 18(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when	

executed by one or more processors, cause the processors to:”
.....132

2. Claim 18(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;” 133
3. Claim 18(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;” 133
4. Claim 18(c): “obtain a speaker embedding for a human speaker; and” 133
5. Claim 18(d): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.” 133

X. OPINIONS REGARDING GROUND 2: CLAIMS 3, 8, 13 ARE OBVIOUS OVER *Feinauer, Prabhavalkar, Fan, and Sargsyan* 134

A. Claim 3 134

1. Claim 3: “The method of claim 1, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.” 134

B. Claim 8 139

1. Claim 8: “The computing system for real-time correction of claim 6, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.” 139

C. Claim 13 140

1. Claim 13: “The non-transitory computer-readable storage medium of claim 11, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized

with a second sample rate different from the first sample rate.”
.....140

XI. OPINIONS REGARDING GROUND 3: CLAIMS 4, 9, and 14 ARE OBVIOUS OVER *Feinauer, Prabhavalkar, Fan, and Kalinli* 140

A. Claim 4 140

2. Claim 4: “The method of claim 1, further comprising processing, in parallel via at least two processing units of the computing system, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.” 140
2. Claim 9: “The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.” 144
3. Claim 14: “The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.” 144

XII. OPINIONS REGARDING GROUND 4: CLAIMS 5, 10, 15 ARE OBVIOUS OVER *Feinauer, Prabhavalkar, Fan, and Watts* 144

A. Claim 5 144

1. Claim 5: “The method of claim 1, further comprising: extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, and generating the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and providing, as an input to at least one of the modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at

least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.” 144

B. Claim 10 156

1. Claim 10: “The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to: extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.” 156

C. Claim 15 156

1. Claim 15: “The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to: extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.” 156

XIII. CONCLUSION.....158

I, Christopher Schmandt, declare as follows:

I. INTRODUCTION AND QUALIFICATIONS

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

A. Educational Background and Professional Experience

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

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

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

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

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

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

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

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

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

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

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

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

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

II. METHODOLOGY AND MATERIALS CONSIDERED

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

13. For this work, I have been asked to review U.S. Patent No. 12,412,561 (“the ’561 Patent”) (Ex. 1001) including the specification and claims, and the ’561 Patent’s prosecution history (“’561 File History”) (Ex. 1002). In developing my opinions relating to the ’561 Patent, I have considered the materials cited or discussed herein, including those itemized in the Exhibit Table below.

Exhibit 1001	U.S. Patent No. 12,412,561 (“’561 Patent”)
Exhibit 1002	Prosecution History of U.S. Patent No. 12,412,561 (“’561 FH”)
Exhibit 1003	Declaration of Chris Schmandt
Exhibit 1004	U.S. Patent Application Publication No. 2020/0193971 to Feinauer et al. (“ <i>Feinauer</i> ”)

Exhibit 1005	U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“ <i>Prabhavalkar</i> ”)
Exhibit 1006	U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>Fan</i> ”)
Exhibit 1007	U.S. Patent Application Publication No. 2020/0066296 to Sargsyan et al. (“ <i>Sargsyan</i> ”)
Exhibit 1008	U.S. Patent No. 7,319,959 to Watts et al. (“ <i>Watts</i> ”)
Exhibit 1009	U.S. Patent Application Publicaiton No. 2012/0116756 to Kalinli (“ <i>Kalinli</i> ”)
Exhibit 1010	<i>Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning</i> to Zhao et al. (“ <i>Accentron</i> ”)
Exhibit 1011	<i>Intentionally left blank</i>
Exhibit 1012	S. Ding, et al., Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition (“ <i>Ding</i> ”)
Exhibit 1013	Priority Chinese Patent Publication No. 202010996501 to Fan et al. and Certified English Translation
Exhibit 1014	Excerpts of the Prosecution History for the U.S. Patent Application Publication No. 2023/0223006 to Fan et al. (“ <i>FanHistory</i> ”)
Exhibit 1015	Priority PCT 2022/057759 to Fan et al. and Certified English Translation.
Exhibit 1016	Munford Librarian Declaration
Exhibit 1017	C. Xydeas, Differential Encoding Techniques Applied to Speech Signals, 1978 (“ <i>Xydeas</i> ”)
Exhibit 1018	U.S. Patent Application Publication No. 2013/0246064 to Wasserblat et al. (“ <i>Wasserblat</i> ”)
Exhibit 1019	U.S. Patent Application Publication No. 2021/0272571 to Balasubramaniam et al. (“ <i>Balasubramaniam</i> ”)
Exhibit 1020	U.S. Patent Application Publication No. 2020/0294492 to Sypniewski et al. (“ <i>Sypniewski</i> ”)
Exhibit 1021	C. Schmandt, <i>Voice Communication with Computers</i> , Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)
Exhibit 1022	C. O’Callaghan, Pitch, (https://web.archive.org/web/20150930164248/http://caseyocallaghan.com/research/papers/Pitch.pdf) September 30, 2015 (“ <i>O’Callaghan</i> ”)
Exhibit 1023	B. Sisman et al., <i>Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion</i> ,

	IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, June 2019, 1085-1097 (“ <i>Sisman</i> ”)
Exhibit 1024	J. Holmes, <i>Speech Synthesis and Recognition</i> , Second Edition, ISBN 0-7484-0856-8 (hc.)—ISBN 0-7484-0857-6 (pbk.), Taylor & Francis e-Library, 2003 (“ <i>Holmes</i> ”)
Exhibit 1025	H. Fletcher, Loudness, Its Definition, Measurement and Calculation, Bell Telephone Laboratories, October 1933, (“ <i>Fletcher</i> ”)
Exhibit 1026	S. Ding et al., Personal VAD: Speaker-Conditioned Voice Activity Detection, arXiv, April 8, 2020 (“ <i>Chang</i> ”)
Exhibit 1027	C. Wiltshire, <i>Chapter 13 English Sounds in Context: The Pronunciation of Phonemes and Morphemes</i> , in <i>Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners</i> , 303-324, 2019 (“ <i>Wiltshire</i> ”)
Exhibit 1028	N. Kazanina, et al., <i>Phonemes: Lexical access and beyond</i> , Psychon Bull Rev 25, 560–585, September 5, 2017 (“ <i>Kazanina</i> ”)
Exhibit 1029	D. Jurafsky et al., <i>Speech and Language Processing</i> , Chapter 27, October 16, 2019 (“ <i>Jurafsky</i> ”)
Exhibit 1030	L. Sun, <i>Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training</i> , IEEE on Multimedia and Expo, August 29, 2016 (“ <i>Sun</i> ”)
Exhibit 1031	G. Zhao et. al, <i>Converting Foreign Accent Speech Without a Reference</i> , IEEE/ACM Transactions On Audio, Speech, and Language Processing, Vol. 29, 2021, 2367-2381 (“ <i>Zhao</i> ”)
Exhibit 1032	U.S. Patent Application Publication No. 2022/0122579 to Biadsky et al. (“ <i>Biadsky</i> ”)
Exhibit 1033	U.S. Patent Appl. No. 2018/0174595 to Dirac et al. (“ <i>Dirac</i> ”)
Exhibit 1034	L. Roberts, <i>Understanding the Mel Spectrogram</i> , (https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53), March 5, 2020 (“ <i>Roberts</i> ”)
Exhibit 1035	N. Umeda, <i>Constant duration in American English</i> , The Journal of the Acoustical Society of America, March 1, 1977 (“ <i>Umeda</i> ”)
Exhibit 1036	U.S. Patent Application No. 2010/0274554 to Orr et al. (“ <i>Orr</i> ”)
Exhibit 1037	U.S. Patent No. 11,410,658 to Ma et al. (“ <i>Ma</i> ”)
Exhibit 1038	U.S. Patent Application No. 2019/0392798 to Danjyo et al. (“ <i>Danjyo</i> ”)

14. I have considered these materials from the viewpoint of a PHOSITA as of the priority date of the '561 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the '561 Patent is January 10, 2022. I note that my opinions provided in this Declaration are made from the perspective of a PHOSITA as of this priority date of the '561 Patent, unless expressly stated otherwise. To the extent that I use any verb tense in this Declaration that is present tense (e.g., “a PHOSITA would understand” instead of “a PHOSITA would have understood”), such verb tense should be understood to be my opinion as of the '561 Patent’s priority date (again, unless expressly stated otherwise). I merely use the present verb tense for ease of reading.

III. OVERVIEW AND LEGAL STANDARDS

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

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

A. Person of Ordinary Skill in the Art

16. I understand that the '561 Patent, the record of proceedings at the Patent Office (which I understand is called the “File History” or “Prosecution History”), and the teachings of the prior art are evaluated from the perspective of a person of ordinary skill in the art (“PHOSITA”). I understand that the factors considered in determining the ordinary level of skill in the art include: (i) the levels of education and experience of persons working in the field; (ii) the types of problems encountered in the field; and (iii) the sophistication of the technology. I may also consider, if available, the education level of the inventor, prior art solutions to the problems encountered in the art, and the rapidity with which innovations are made in the relevant art.

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

B. Obviousness

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C. Analogous Art

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

D. Claim Construction

26. I understand that the claim terms in this PGR will be construed according to their plain and ordinary meaning as understood in light of the claim language, the patent's description, and the prosecution history viewed from the perspective of a PHOSITA. I further understand that where a patent defines claim language, the definition in the patent controls, even if there are other definitions that might be understood by those working in the art.

IV. LEVEL OF A PERSON OF ORDINARY SKILL

27. Based on my review and analysis of the '561 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a PHOSITA in the field of the '561 Patent at the time of the earliest possible priority date (January 10, 2022) would have been knowledgeable regarding the field of speech audio processing. In my experience working in this field, most workers of ordinary skill in the art as of the earliest possible priority date of January 10, 2022, would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of speech processing and machine learning. *See, e.g.*, the '561 Patent (Ex. 1001) at 1:23-3:3 (describing the "Background" and "Summary"). A person with less relevant practical experience but with additional education may also qualify as a PHOSITA in the field of the '561 Patent provided the additional education focused on speech

processing technology. When I refer to the understanding of a PHOSITA, I am referring to the understanding of such a person as of January 10, 2022.

V. OVERVIEW OF TECHNOLOGY

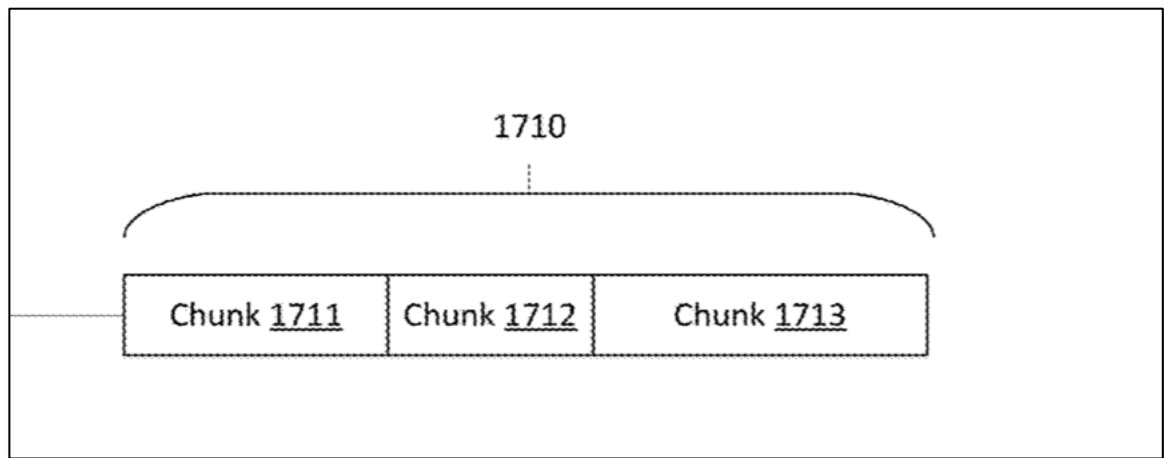
28. I was asked to briefly summarize the background of the technology from the standpoint of a PHOSITA prior to January 10, 2022.

A. Audio Chunks

29. It was well known to process audio signals utterances in smaller portions, especially to process audio in real-time. For example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” in order to maintain “low real time delay” EX1018-Wasserblat, Abstract, [0025]. U.S. Patent Appl. No. 2021/0272571 to Balasubramaniam et al. (“*Balasubramaniam*”) teaches waveforms “of each speech utterance may be split into chunks of 200 ms with 10 ms overlapping frames.” EX1019-Balasubramaniam, [0310]. This overlap is referred to as “context[.]” EX1019-Balasubramaniam, [0312].

30. As another example, U.S. Patent Appl. No. 2020/0294492 to Sypniewski et al. (“*Sypniewski*”) teaches a front end module 201 that splits raw audio data into a series of frames. EX1020-Sypniewski, [0059], Fig. 2. *Sypniewski* provides an example in which a 7-10 second long portion of input audio “may be segmented on relative silence such that the input audio comprises an entire

utterance.” EX1020-Sypniewski, [0057], [0110]. *Sypniewski* characterizes these 7-10 second portions of audio as “chunks.” EX1020-Sypniewski, [0162]. Fig. 17 below illustrates an example of audio 1710 split into chunks 1711-1713.



EX1020-Sypniewski, Fig. 17. (excerpt) *Sypniewski* notes “[a]ny frame duration, spacing, and overlap may be used as appropriate for any given implementation as determined by one skilled in the art.” EX1020-Sypniewski, [0059].

B. Acoustic Features

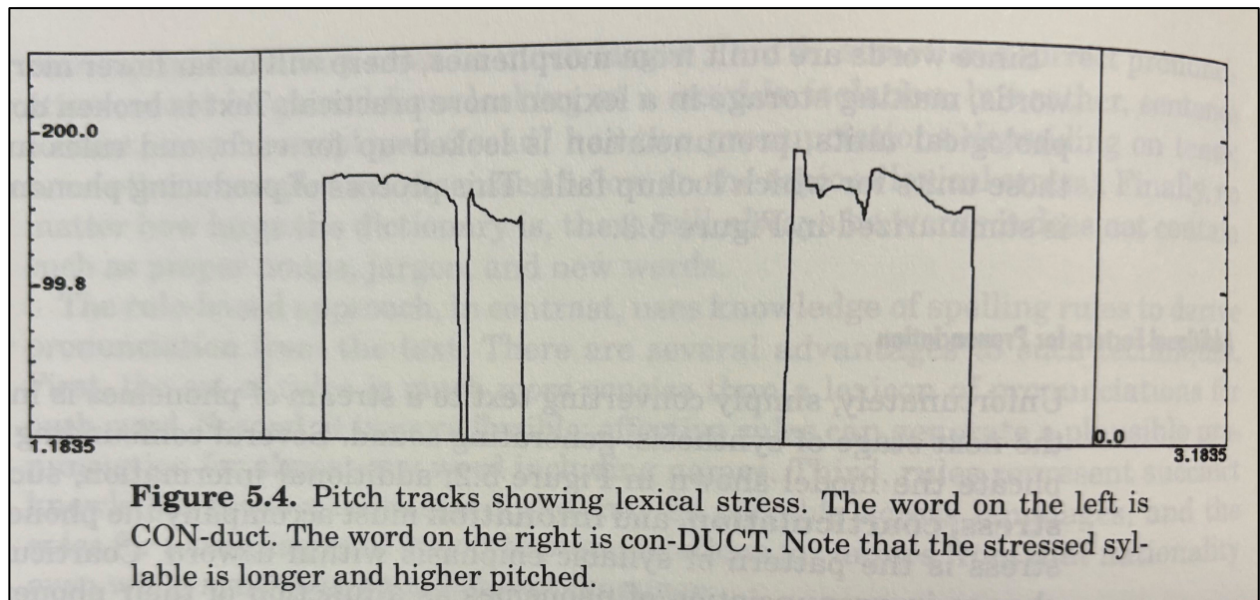
31. It was well known that acoustics are a measured property of sound such as, but not limited to, a characteristic of speech analyzed in physical terms. Further, it was well known in the prior art that acoustic features may include several features, including but not limited to stress, accent, volume, energy, tone, pitch, main frequency F0, and a voice activity detector as discussed below.

32. It was common in the prior art that several of these features were also referred to as “style” and/or “prosody”. *Voice Communication with Computers* to

Schmandt (“*Schmandt*”) teaches that prosody “refers to the spoken style of discourse independent of lexical content[.]” EX1021-Schmandt, 301-302. *Schmandt* further teaches that stress, pitch, duration, rhythm, meter, and intonation are all examples of prosody speech style. EX1021-Schmandt, p. 17, 301-302.

a. Stress or Accent

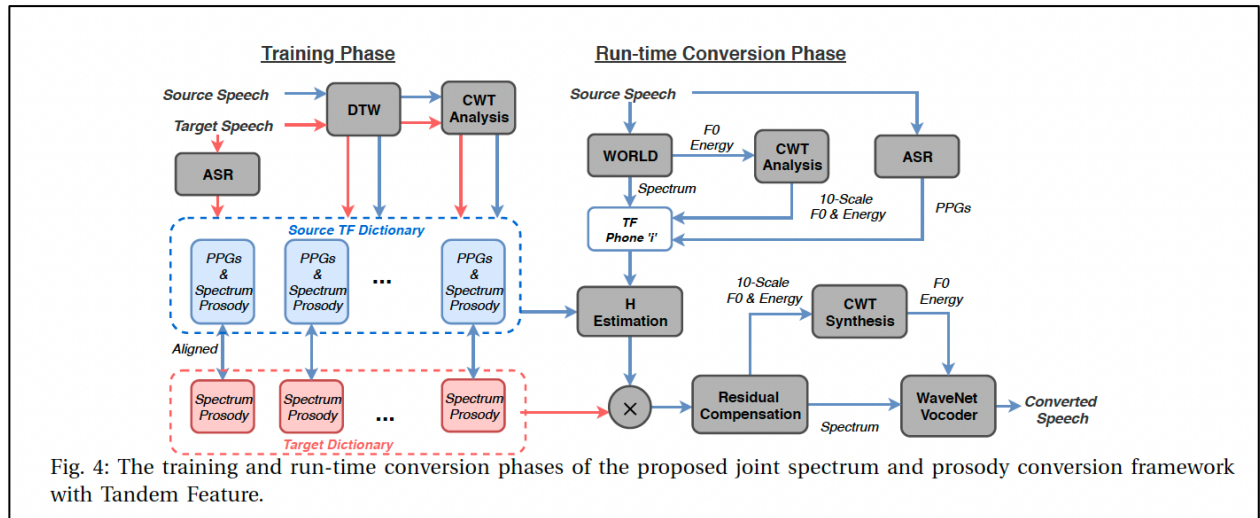
33. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches lexical “stress is part of a complex interplay between the intonation and rhythm of a whole sentence...Stress covers many phenomena, all realized as variations in rhythm and meter; lexical stress is what differentiates syllabic “accent” as opposed to emphasizing one word over another in a sentence. Lexical stress is important for understanding spoken language; incorrectly assigned stress may make a word unintelligible. For some English words the assignment of stress differentiates noun and verb forms, e.g., ‘convert’ and ‘conflict’...Stress...is not directly phonemic. In addition to producing the proper phonemes, the syllables containing those phonemes also must be marked for stress. This is done by assigning pitch and duration for each voiced phoneme during the conversion to phonemes.” EX1021-Schmandt, 88.



EX1021-Schmandt, Fig. 5.4

b. Pitch or Fundamental Frequency (F0)

34. *Pitch* to O’Callaghan (“O’Callaghan”) teaches pitch “of a sound is identical with its fundamental frequency; that is, the pitch of a periodic sound is the greatest whole-number frequency by which the frequency of each of its sinusoidal components is divisible without remainder.” EX1022-O’Callaghan, 4. In other words, “the fundamental frequency of a complex periodic tone determines its perceived pitch.” EX1022-O’Callaghan, 4. As another example, EX1023, *Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion* to Sisman et al. (“Sisman”) teaches “[a]t run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour.” EX1023-Sisman, 6-7.

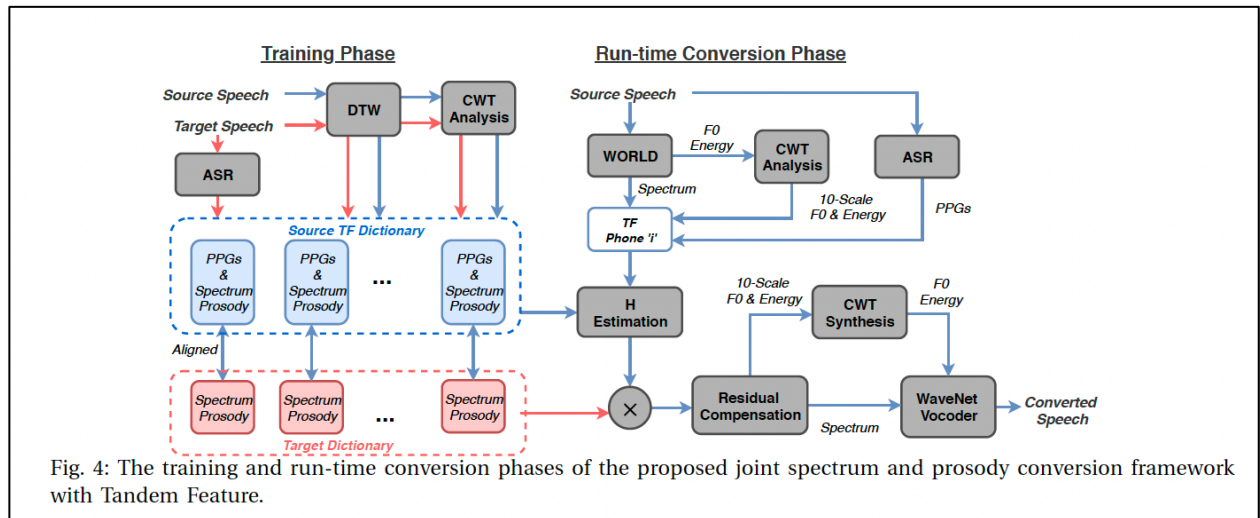


EX1023-Sisman, Fig. 4. Pitch or fundamental frequency is “a crucial prosodic feature in speech.” EX1023-Sisman 3rhc. *Speech Synthesis and Recognition* to Holmes et al. (“Holmes”) teaches that pitch may change by “the pattern of stressed syllables...that are influenced by the phonetic detail of an utterance.” EX1024-Holmes, 104-105.

c. Energy or Volume

35. It was common to measure energy and volume in signals. To calculate loudness, the energy of a signal is measured over time. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“Fletcher”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point.” EX1025-Fletcher, 83. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. As another example,

Sisman teaches “[a]t run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour.” EX1023-Sisman, 6-7.



EX1023-Sisman, Fig. 4.

d. Voice Activity Detection (VAD)

36. *Personal VAD: Speaker-Conditioned Voice Activity Detection* to Chang et al. (“*Chang*”) teaches “a system to detect the voice activity of a target speaker at the frame level.” EX1026-Chang, Abstract. *Chang* teaches to find the “voice activity of a target speaker in real time.” EX1026-Chang, 11hc. The voice activity determines if audio contains “speech/non-speech.” EX1026-Chang, 11hc.

37. It was well known to perform VAD so that the “performance of downstream components by discarding non-speech signals, but also significantly reduces the overall computational cost due to its relatively small size.” EX1026-

Chang 1rhc. Detecting VAD “reduce[s] the computational cost and battery consumption.” EX1026-Chang, Abstract.

C. Linguistic Features

38. Linguistic features were commonly known in the prior art to include an expression of a phoneme. A phoneme is the smallest unit of sound that distinguishes meaning in a language. EX1027-Wiltshire, 304.

39. *Phonemes: Lexical access and beyond* to Kazanina et al. (“Kazanina”) teaches phonemes have two essential properties: “they are ‘segment-sized’ (the size of a consonant or a vowel) and abstract (a single phoneme may [] have different acoustic realisations).” EX1028-Kazanina, 560. *English Sounds in Context: The Pronunciation of Phonemes and Morphemes* to Wiltshire (“Wiltshire”) teaches changing one phoneme changes the meaning of a word. EX1027-Wiltshire, 560. For example, “the words *pat* and *bat* are identical except for the initial sounds, which are therefore responsible for indicating the difference in meaning between the two words.” EX1027-Wiltshire, 304. Such pairs of words are called “minimal pairs,” because they only differ by a single sound but have completely different meanings. EX1027-Wiltshire, 304. Thus, phonemes are extremely helpful in differentiating minimal pairs because “the phoneme is the smallest potential unit of difference between similar words recognizable as different to the native [speaker].” EX1027-Wiltshire, 304.

40. As such, “two words that are distinct in meaning differ in phonological form, with a minimal difference being exactly one phoneme within the same position in the word (e.g. /kæt/ ‘cat’ vs. /mæt/ ‘mat’). EX1028-Kazanina, 561. Additionally, different words with distinct meanings can employ the same set of phonemes but in different orders (e.g. cat /kæt/ vs. act /ækt/ vs. tack /tæk/). EX1028-Kazanina, 561.

41. Different phonetic alphabet systems have been designed to represent phonemes. As discussed in *Speech and Language Processing*, Chapter 27 “Phonetics” to Jurafsky et al. (“*Jurafsky*”), one phonetic alphabet is the International Phonetic Alphabet (IPA), which is an evolving standard originally developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. EX1029-Jurafsky, 2. Another phonetic alphabet is the ARPAbet, which uses “ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters.” EX1021-Schmandt, 16. ARPAbet was developed “for the convenience of computer manipulation and representation of speech using ASCII-printable characters.” EX1021-Schmandt, 16.

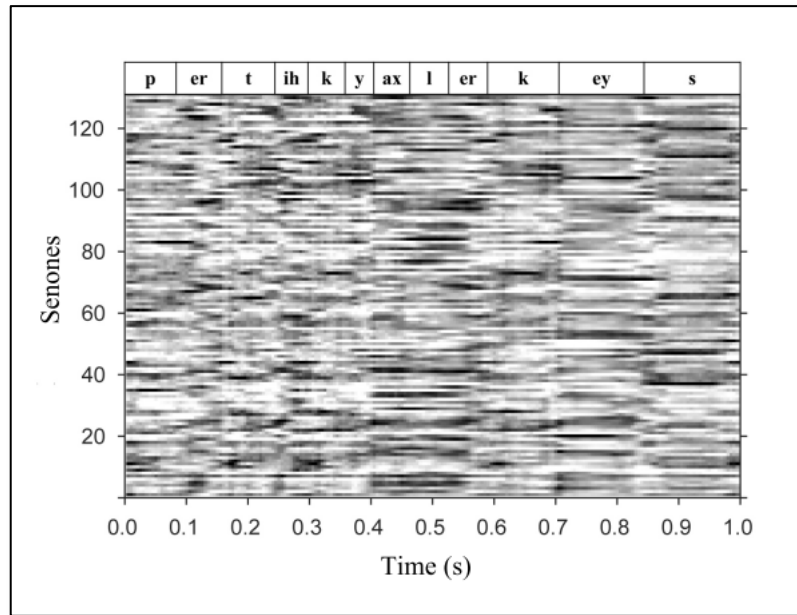
ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɜ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

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

EX1029-Jurafsky, 2.

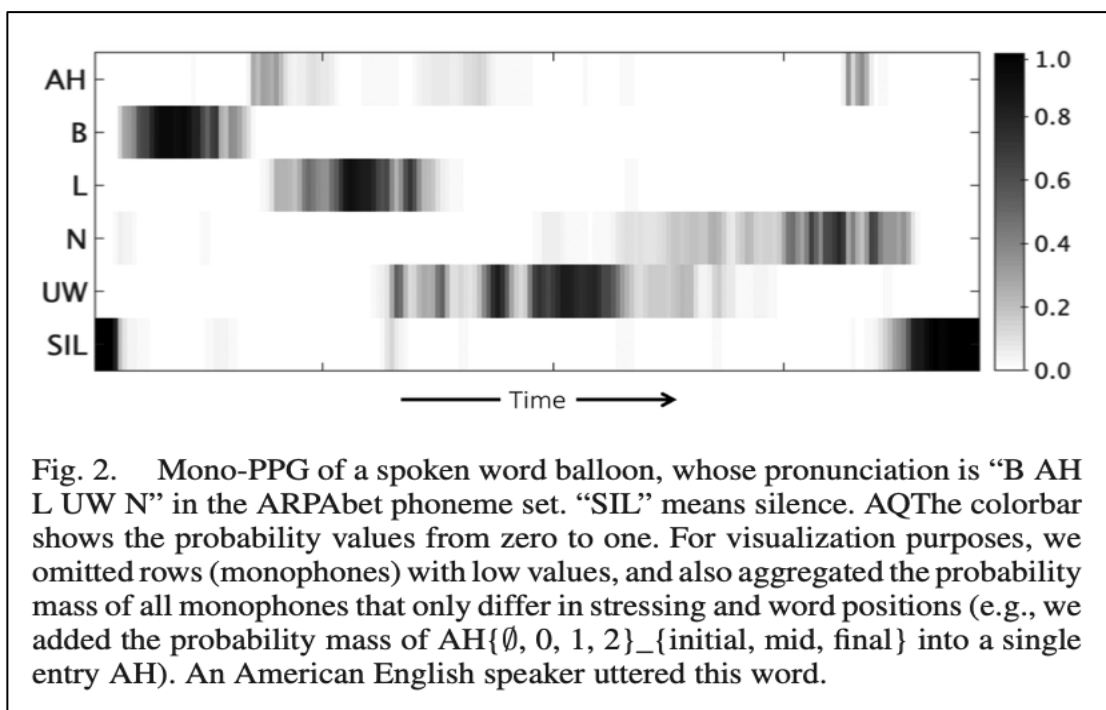
a. Phonetic Posteriorgrams (PPGs)

42. One common way to represent phonemes in accent conversion systems is in the form of a phonetic posteriorgram. *Phonetic Posteriorgrams For Many-To-One Voice Conversion Without Parallel Data Training* to Sun et al. (“Sun”) teaches a Phonetic Posteriorgram (PPG) “is a time-versus-class matrix representing the posterior probabilities of each phonetic class” per utterance. EX1030-Sun, 2–3. “A phonetic class may refer to a word, a phone or a senone.” EX1030-Sun, 3. An example of a PPG (using senones) is as follows:



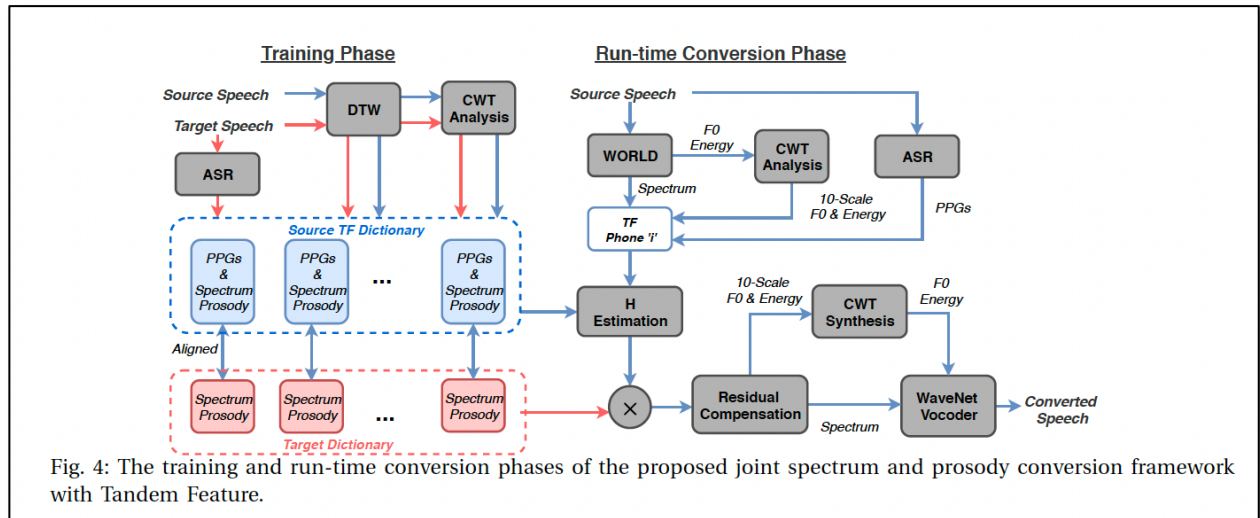
EX1030-Sun, Fig 4. In this example, the “horizontal axis represents time in seconds and the vertical one contain indices of phonetic classes.” EX1030-Sun, 3. A “[d]arker shade implies a higher posterior probability.” EX1030-Sun, 3.

43. Another example of a phonetic posteriorgrams used to represent phonemes is taught in EX1031, *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) using the ARPAbet phonetic alphabet and depicted below:



EX1031-Zhao, Fig. 2. *Zhao* teaches this “phonetic posteriorgram obtained by collapsing the senones into monophone symbols (346 monophones with word positions, e.g., word-initials, word-finals). For each monophone symbol, we aggregate the probability mass of all the senones that share the same root monophone.” EX1031-Zhao, 2370 rhc.

44. As another example, EX1023, *Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion* to *Sisman et al.* (“Sisman”) teaches “to convert all speaker’s voice into PPGs, and construct dictionaries that consist of source spectrum and prosody features as well as PPGs.” EX1023-Sisman, 6. *Sisman* teaches “a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary[.]” EX1023-Sisman, 6.



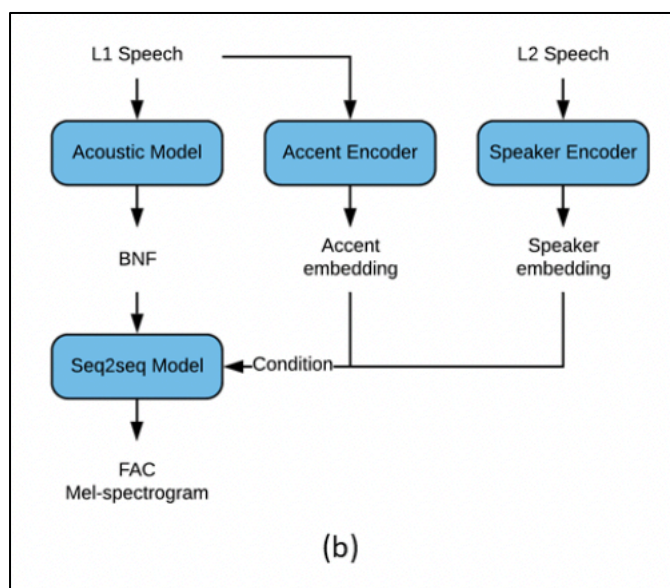
EX1023-Sisman, Fig. 4. *Sisman*’s PPG dictionary has “multiple coupled dictionaries $[A_i; B_i]$, one for each phone i , where $i = 1, \dots, N$, A_i is the source phonetic dictionary, and B_i the target phonetic dictionary. The source phonetic dictionary here consists of both PPGs, source spectrum features as well as CWT-based prosody features, called *source TF dictionary* while the target phonetic dictionary B_i only includes spectrum features and prosody features, called *target dictionary*.” EX1023-Sisman, 7.

D. Prior Art Accent Conversion Systems

45. After extracting the linguistic and acoustic features, along with any other needed information, it was common practice to perform accent conversion. For example, U.S. Patent Appl. No. 2022/0122579 to Biadsy et al. (“*Biadsy*”) teaches the speech-to-speech conversion server within the computing device “convert[s] audio data of an utterance spoken by a user who may speak with an accent into audio

data of a synthesized utterance that may have a different accent.” EX1032-Biadsy, [0023]; *see also* [0021].

46. As another example, *Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning* to Zhao et al. (“Accentron”) teaches an accent conversion system that “can generate accent conversions to arbitrary L2 speakers given a few seconds of audio (i.e., enough speech to compute a speaker embedding).” EX1010-Accentron, 3. *Accentron* teaches “Accentron, an approach that can generate accent-converted speech for arbitrary L2 speakers unseen during training. In the proposed approach, we first train a speaker-independent acoustic model on L1 corpora to extract bottleneck features that represent the linguistic content of utterances. Then, we develop a speaker encoder and an accent encoder to generate embedding vectors for the desired voice identity (L2 speaker’s) and accent (L1 accent), respectively.” EX1010-Accentron, 1.



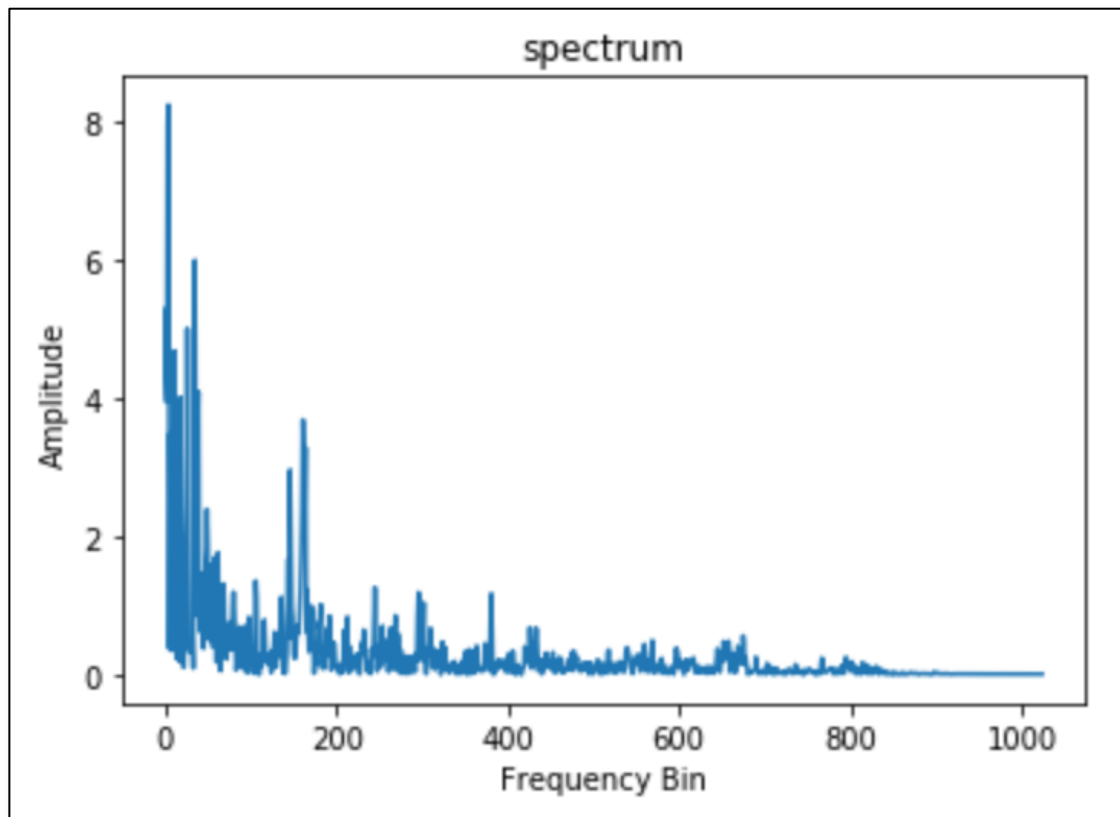
EX1010-Accentron, Fig. 1 (excerpt). *Accentron* “use[s] a sequence-to-sequence model to transform bottleneck-features to Mel-spectrograms, conditioned on the L2 speaker embedding and the L1 accent embedding. We conducted experiments on the L2-ARCTIC corpus under two testing conditions: the standard FAC setting where test L2 speakers were seen during training, and a zero-shot FAC setting where test L2 speakers were unseen during training.” EX1010-Accentron, 1.

47. Further, U.S. Patent Appl. No. 2018/0174595 to Dirac et al. (“*Dirac*”) teaches an accent conversion system that uses “collected audio samples to generate one or more accent translation models for translating speech from one accent to another.” EX1033- *Dirac*, [0015]. That is, audio samples “associated with at least a first accent” are compared to audio samples of “at least one other accent” to determine “various audio characteristics associated with...the accent.” EX1033- *Dirac*, [0048]. Resultingly, *Dirac* teaches outputting an “audio portion substantially associated with the second accent in the...same spoken language as the input audio” via the accent translation model, including adjusting the audio characteristics of the input audio ...to more closely resemble the audio characteristics of the second accent. EX1033-*Dirac*, [0053] and [0049].

a. Mel Spectrograms

48. An audio signal is “comprised of several single-frequency sound waves...The Fourier transform is a mathematical formula that allows us to

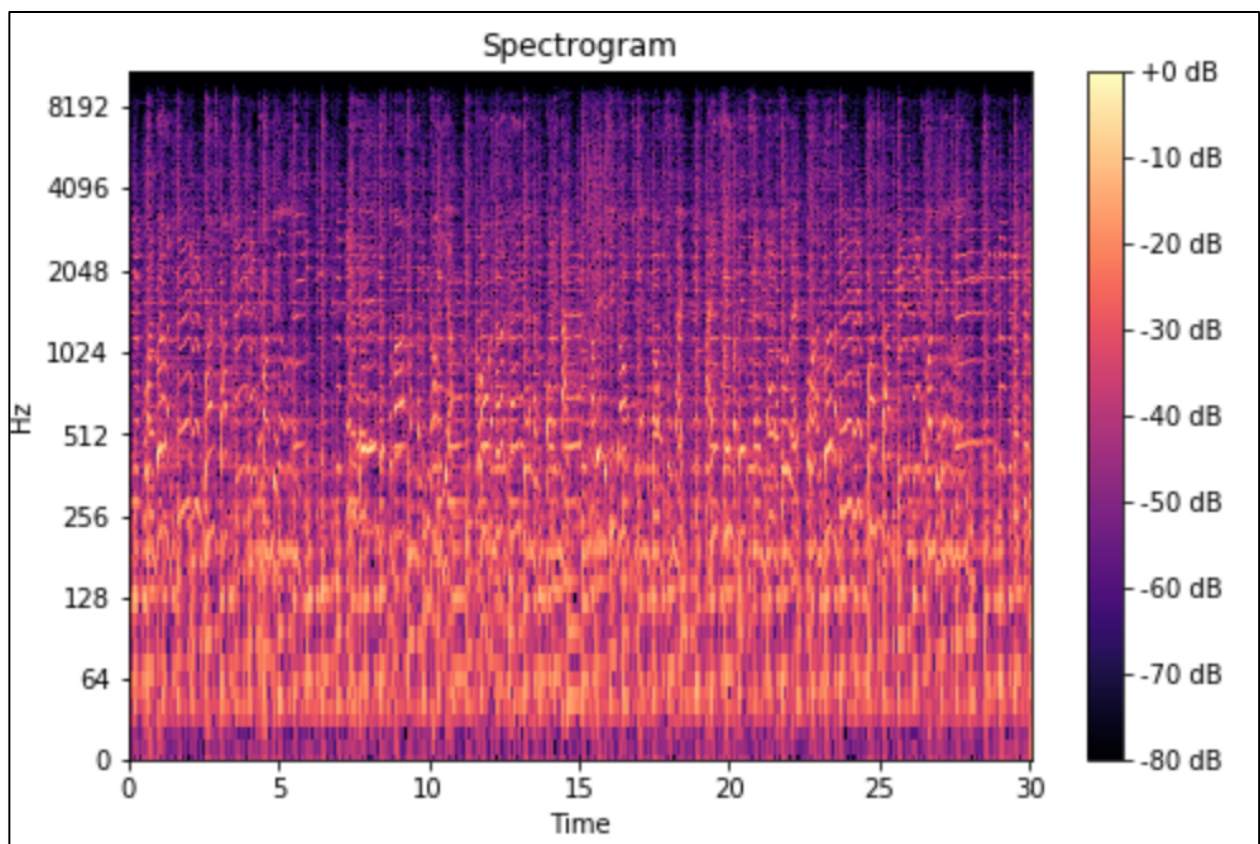
decompose a signal into its individual frequencies and the frequency's amplitude. In other words, it converts the signal from the time domain into the frequency domain. The result is called a spectrum.” EX1034-Roberts, 4-5.



EX1034-Roberts, 4.

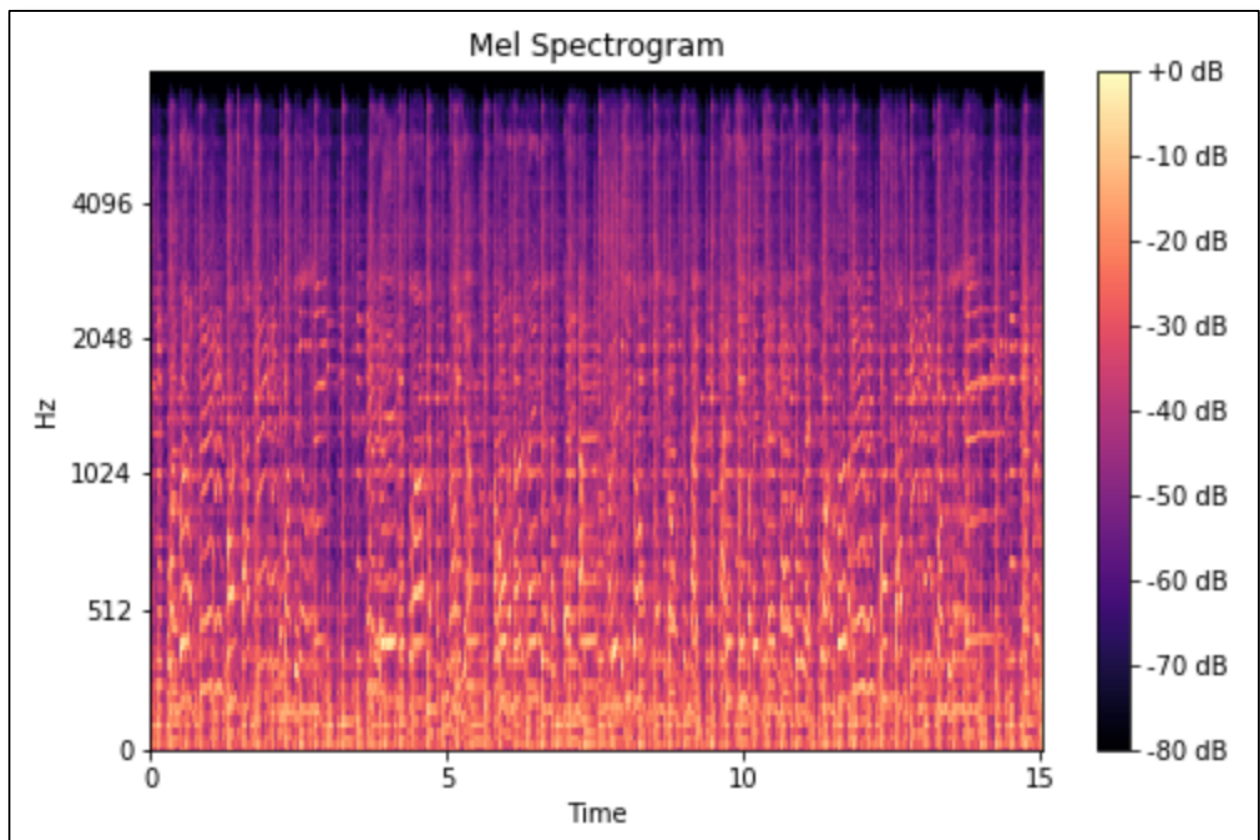
49. We “need a way to represent the spectrum of these signals as they vary over time.” EX1034-Roberts, 7. Therefore, a “short-time Fourier transform” or a “fast Fourier transform” (FFT) is performed on the frequency domain signal in order to “analyze the frequency content of a signal”, and if the “FFT is computed on overlapping windowed segments of the signal, and we get what is called the spectrogram” in order to visually represent the signals “loudness, or amplitude,

as it varies over time at different frequencies.” EX1034-Roberts, 7-8. *Voice Communication with Computers* to Schmandt (“Schmandt”) teaches a “spectrogram is a transformation of the waveform into the frequency domain.” EX1021-Schmandt, 13. In a spectrogram, the “horizontal axis represents time, the vertical axis represents frequency, and the intensity or blackness at a point indicates the acoustic energy at that frequency and time.” EX1021-Schmandt, 13. A spectrogram “requires much less storage than the original waveform and reveals acoustic features specific to speech.” EX1021-Schmandt, 13.



EX1034-Roberts, 6.

50. Studies “have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies.” EX1034-Roberts, 11. In order to combat this, in “1937, Stevens, Volkmann, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the mel scale.” EX1034-Roberts, 12. From the spectrogram in the prior paragraph, you “perform a mathematical operation on frequencies to convert them to the mel scale.” EX1034-Roberts, 12. A Mel-spectrogram is “is a spectrogram where the frequencies are converted to the mel scale.” EX1034-Roberts, 13.



EX1034-Roberts, 8.

b. Vocoder

51. *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches that to obtain hearable synthesized audio, mel spectrograms are transformed into waveforms using a real-time vocoder. EX1031-Zhao, 2371rhc (“The predicted Mel-spectrograms are converted back to audio waveforms using a WaveGlow neural vocoder trained on the L2 utterances[.]”); EX1031-Zhao, 2368rhc. (“we use a high-quality WaveGlow [15] real-time neural vocoder to convert Mel-spectrograms to speech waveform.”).

52. Another type of vocoder is a “WaveNet vocoder [21] to generate the speech waveforms.” EX1023-Sisman, 8 lhc.

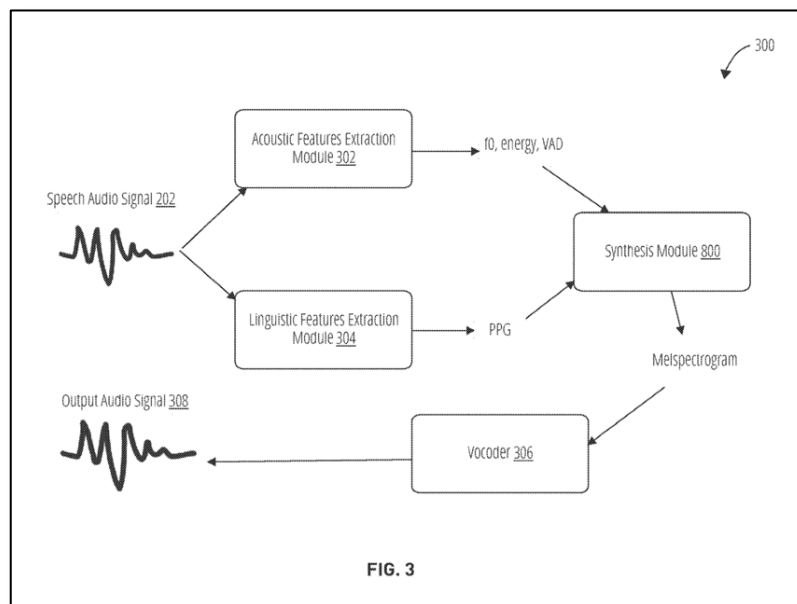
VI. OVERVIEW OF THE ’561 PATENT

53. I have reviewed the ’561 Patent (EX1001). Descriptions of the ’561 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the ’561 Patent are intended only as a summary of the ’561 Patent and/or intended as being pertinent to the Challenged Claims of the ’561 Patent. My opinions provided herein are not intended as a discussion of my entire understanding of the ’561 Patent or any other claims in any related patent to the ’561 Patent.

A. Summary of the ’561 Patent

54. The ’561 Patent describes “real-time correction of an accent in an input audio signal are provided.” EX1001, Abstract. “FIG. 3 is a block diagram showing a pipeline 300 for real-time correction of an accent in input audio signals, according

to an example embodiment. The pipeline 300 (also referred to as a computational graph) may include an acoustic features extraction module 302, a linguistic features extraction module 304, a synthesis module 800, and a vocoder 306.” EX1001, 6:61-67.



EX1001, Fig. 3.

B. Field of Endeavor of the '561 Patent

I have been informed that the field of endeavor of the claimed invention can be determined by reference to explanations of the invention's subject matter in the patent application, including the embodiments, function, and structure of the claimed invention.

55. The '561 Patent defines the “Technical Field” as relating to “audio processing” and, more specifically, to “systems and methods for real-time correction of accent in input audio signals.” EX1001, 1:19-22. After reviewing the '561 Patent,

it is my opinion that a PHOSITA would understand that the field of endeavor of the '561 Patent includes speech audio processing. *See* EX1001, Abstract, 1:19-22, Fig. 3.

C. Problem Purportedly Solved by the Inventors of the '561 Patent

56. I have been informed that a prior art reference is “reasonably pertinent” if a PHOSITA would have consulted it and applied its teachings when faced with the problem that the inventor was trying to solve. As such, I have been asked to analyze the '561 Patent and determine the problem that the inventors were trying to solve.

57. The '561 Patent notes that a concern exists regarding the understandability between speakers possessing different accents during live video or audio conversations and therefore “real-time correction of accent in input audio signals.” EX1001, 1:19-22, *See also* Title, Abstract, 1:19-22, 4:28-30, Fig. 3. That is, the '561 Patent details that a method for the “improved extraction” of acoustic and linguistic features for the generation of a modified, accented audio is desirable. EX1001, 15:11-21; *see also* 4:21-23, 15:11-21, Fig. 3. Thus, a PHOSITA would understand that the '561 Patent is directed to solving the problems of real-time audio modification, accent conversion and feature extraction from audio data.

D. File History of '561 Patent

58. I have reviewed the prosecution history for the '561 Patent. The '561 Patent claims priority to Provisional Application No. 63/297,901 filed January 10, 2022. The '561 PGR faced one office action with a double patenting rejection for claims 1, 3-4, 7-8, 13-14, 17-18, and 20-27 to U.S. Patent No. 11,715,457. EX1002, 149-157.

VII. OVERVIEW OF THE PRIOR ART

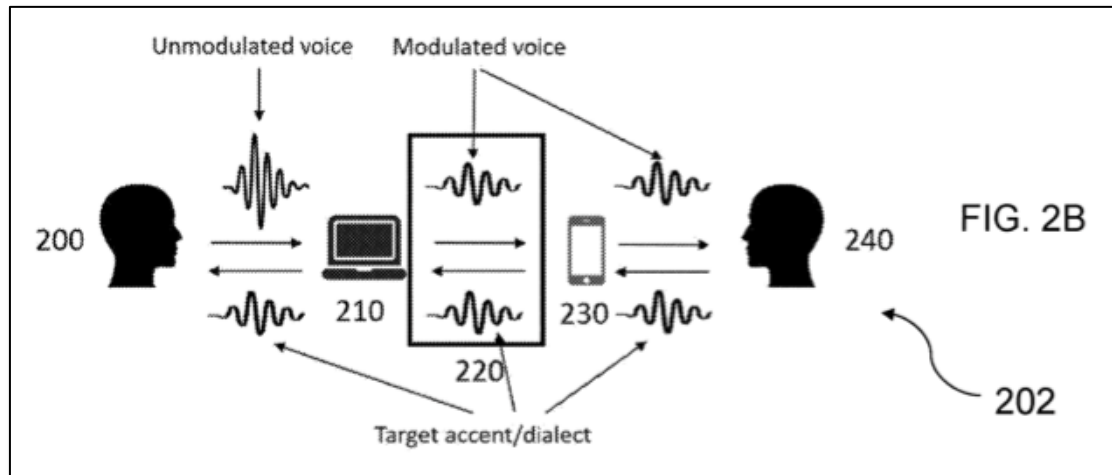
59. I have been informed that for a prior art reference to be proper for use in an obviousness analysis, the reference must be “analogous art” to the claimed invention. I have been informed that a prior art reference is analogous to the claimed invention if the reference is from the same field of endeavor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

60. The prior art is analogous to the claimed invention of the '561 Patent, as discussed with greater detail for each prior art reference included. In general, the prior art is from the same field of endeavor of the '561 Patent, namely speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3.

A. Overview of *Feinauer*

61. I have been informed U.S. Patent Appl. No. 2020/0193971 to Feinauer et al. (“*Feinauer*”) was filed on December 13, 2018, and published on June 18, 2020.

Feinauer teaches “methods for accent and dialect modification[.]” EX1004-Feinauer, Abstract.



EX1004-Feinauer, Fig. 2B.

62. A PHOSITA would have classified *Feinauer* within the same field of endeavor as the '561 Patent because both *Feinauer* and the '561 Patent relate to speech audio processing. EX1004-Feinauer, Title, Abstract, [0001], [0091] with EX1001, Title, Abstract, 1:19-22.

63. Additionally, a PHOSITA would have found *Feinauer* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Feinauer* and the '561 Patent describe the difficulties posed by accent conversion. EX1001, 1:19-22, 4:28-30, Fig. 3. For example, Feinauer teaches “[c]ommunication between people with different dialects”, including “miscommunication, misunderstanding, and negative opinions based on suboptimal pairings of language, dialect and/or accent between conversation partners” as potential outcomes

particularly without the context of visual cues. EX1004-Feinauer, [0002]-[0004]. *Feinauer* then identifies a system and method for addressing such problems via accent conversion, particularly the identification of an “accent and dialect modifier” utilized to modify “a spoken word to within a distance vector of the target threshold” of the desired accent. EX1004-Feinauer, [0086].

64. Further, the pertinence of *Feinauer* to the problem faced by the inventors of the '561 Patent would have been identifiable by a PHOSITA as real-time audio modification. EX1001, Title, Abstract, 1:19-22, 4:28-30, Fig. 3; EX1004-Feinauer, Abstract, [0002]- [0004], [0135].

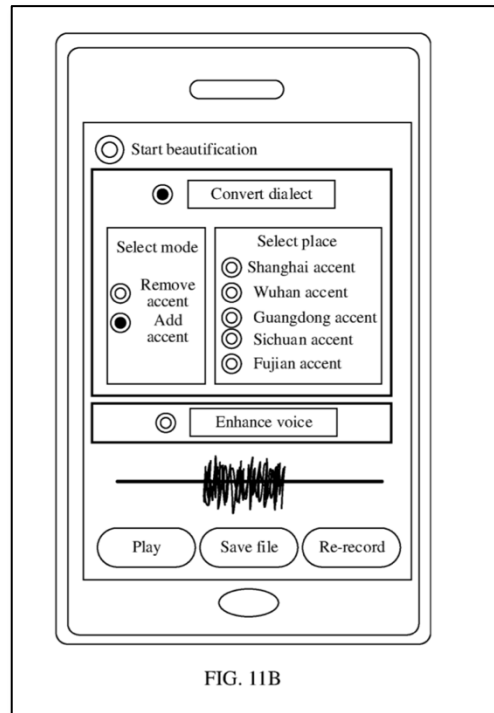
65. Further, the pertinence of *Feinauer* to the problem faced by the inventors of the '561 Patent would have been identifiable by a PHOSITA as feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3. *Feinauer's* real-time audio modification extracts “prosodic vocal features” from a detected dialect for the generation of the target audio data such that effective communication “across multiple dialects and/or accents” is achievable. EX1004-Feinauer, [0002]-[0004], [0086] [0095], [0099].

B. Overview of *Fan*

66. I have been informed U.S. Patent Appl. No. 2023/0223006 to Fan et al. (“*Fan*”)¹ was filed on March 20, 2023, published July 13, 2023, and is a continuation of PCT/CN2021/117945 (EX1015), which was filed on September 13, 2021, and claims priority to Chinese application 202010996501.5 (EX1013) filed September 21, 2020. *See* EX1014-*FanHistory*. The US application incorporates both the PCT application (EX1015) and the Chinese application (EX1013) by reference. The PCT application (EX1015) incorporates the Chinese application (EX1013) by reference.

67. *Fan* teaches “a dialect conversion mode, for adding an accent to or removing an accent from the first voice[.]” EX1006-*Fan*, Abstract.

¹ I have been informed that EX1006 is the U.S. Patent Appl. to *Fan*. I have been informed that also included is the Chinese priority application (EX1013), the English translation of the Chinese priority application (EX1013, 1), and a certification that the translation is true and correct (EX1013, 63). I have been informed that also included is the PCT priority application (EX1015), English translation of the PCT priority application (EX1015), and a certification that the translation is true and correct (EX1015, 64). I have been informed that *Fan* made a foreign priority claim within 12 months of the foreign priority application (EX1014, 3-4) and the Office acknowledged same (EX1014, 10).



EX1006-Fan, Fig. 11B.

68. A PHOSITA would have identified *Fan* as being within the same field of endeavor as the '561 Patent because both *Fan* and the '561 Patent relate, namely, to speech audio processing. EX1006-Fan, Abstract, [0134], [0201] with '561 Patent, Title, Abstract, 1:19-22, Fig. 3. *Fan* discloses a system having the capacity for “speaking style conversion on a to-be-converted first voice...[and] a dialect conversion mode[.]” EX1006-Fan, Abstract.

69. A PHOSITA would also have found *Fan* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Fan* and the '561 Patent describe accent conversion. EX1001, Title, Abstract, 1:19-22, 4:28-30, Fig. 3; EX1006-Fan, Abstract, [0120], [0135], [0201].

70. Further, a PHOSITA would have found both *Fan* and the ‘561 Patent directed to feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3. Comparatively, *Fan*’s dialect conversion mode utilizes a voice feature extraction model to extract a phoneme posteriorgram feature from a first voice (the input voice) having a first accent and a style separation model to extract a style feature from the input voice. EX1006-Fan, Abstract, [0113], [0120], [0127], [0135], [0201].

71. The subject matter relied upon from *Fan* is supported in *Fan*’s priority application and PCT, as shown in the table below. I rely on disclosures in the patent and priority application, as set forth below:

<i>Fan</i> Publication (EX1006) (figure or paragraph)	<i>Fan</i> PCT (EX1015) (page, figure or paragraph)	<i>Fan</i> Priority (EX1013) (page, paragraph)
Abstract	47, Abstract and cover page at (57)	39, Abstract
Fig. 4	50, Fig. 4	42, Fig. 4
Fig. 5	51, Fig. 5	43, Fig. 5
Fig. 7	54, Fig. 7	46, Fig. 7
Fig. 9	55, Fig. 9	47, Fig. 9
Fig. 11B	58, Fig. 11B	50, Fig. 11B
Fig. 12	59, Fig. 12	51, Fig. 12
Fig. 13	59, Fig. 13	51, Fig. 13
[0007]	2, ¶2	2, ¶2
[0010]	3, ¶2	3, ¶1
[0023]	6, ¶2	6, ¶2
[0024]	6, ¶3	6, ¶3
[0025]	6, ¶4	6, ¶4
[0026]	6, ¶5	6, ¶5
[0045]	8, ¶6	8, ¶4
[0046]	9, ¶1	9, ¶1
[0047]	9, ¶2	9, ¶2

[0048]	9, ¶3	9, ¶3
[0049]	9, ¶4	9, ¶4
[0050]	10, ¶1	9, ¶5
[0053]	10, ¶4	10, ¶3
[0075]	13, ¶5	13, ¶4
[0081]	14, ¶1	14, ¶1
[0082]	14, ¶2	14, ¶2
[0085]	14, ¶5	14, ¶5
[0089]	15, ¶4	15, ¶4
[0090]	15, ¶5	15, ¶5
[0091]	16, ¶1	15, ¶6
[0092]	16, ¶2	16, ¶1
[0093]	16, ¶3	16, ¶2
[0094]	16, ¶4	16, ¶3
[0095]	17, ¶3	16, ¶4
[0097]	17, ¶3	17, ¶2
[0099]	18, ¶2	18, ¶2
[0106]	19, ¶6	19, ¶5
[0113]	21, ¶1	20, ¶6
[0116]	21, ¶4	21, ¶3
[0120]	22, ¶4	22, ¶3
[0124]	23, ¶2	22, ¶7
[0125]	23, ¶3	23, ¶1
[0126]	23, ¶4	23, ¶2
[0127]	23, ¶5	23, ¶3
[0128]	24, ¶1	23, ¶4
[0129]	24, ¶2	24, ¶1
[0134]	25, ¶2	25, ¶1
[0135]	25, ¶3	25, ¶2
[0152]	28, ¶6	28, ¶4
[0153]	28, ¶7	28, ¶5
[0154]	28, ¶8	28, ¶6
[0155]	28, ¶9	28, ¶7
[0156]	28, ¶10	28, ¶8
[0158]	29, ¶2	28, ¶10
[0159]	29, ¶3	29, ¶1
[0160]	29, ¶4	29, ¶2
[0161]	29, ¶5	29, ¶3

[0162]	29, ¶6	29, ¶4
[0166]	30, ¶2	29, ¶8
[0167]	30, ¶3	30, ¶1
[0170]	30, ¶6	30, ¶4
[0171]	30, ¶7	30, ¶5
[0201]	35, ¶3	35, ¶3

C. Overview of *Prabhavalkar*

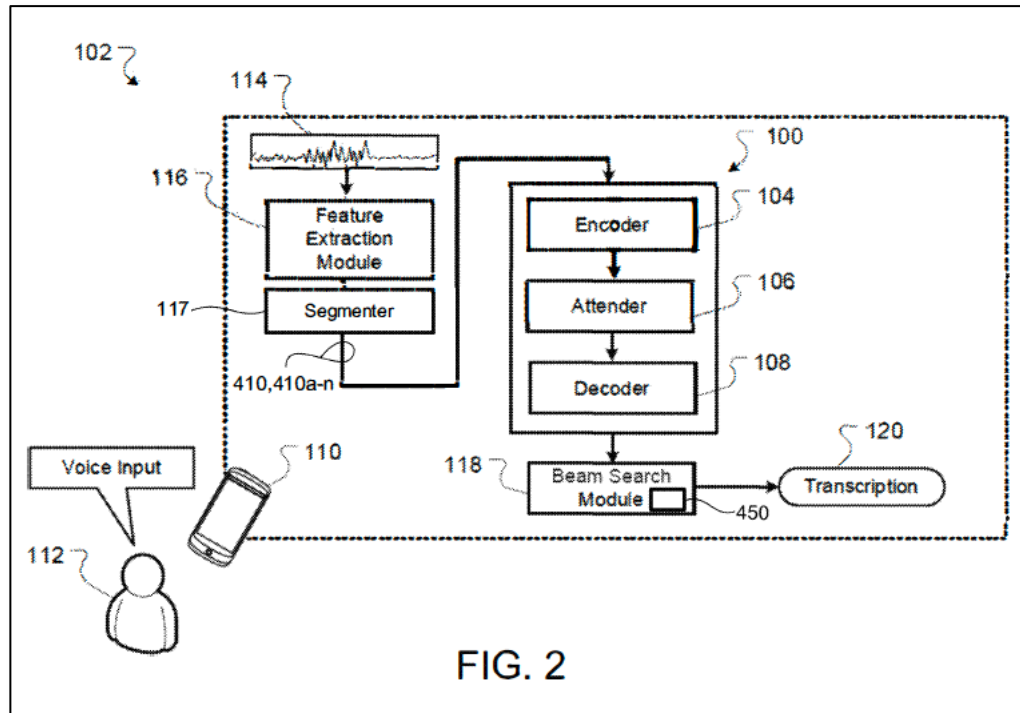
72. I have been informed U.S. Patent No. 12,106,749 to Prabhavalkar et al. (“*Prabhavalkar*”) was filed on September 20, 2021, and published on October 1, 2024. *Prabhavalkar* determines, “from the audio data for the utterance, speech frames,” and “successively processing chunks that each include different sets of speech frames with the speech recognition model.” EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. *Prabhavalkar* envisions “each chunk comprising a same predetermined number of speech frames, each chunk having a first predetermined number of speech frame representing speech.” EX1005-*Prabhavalkar*, 6:65-7:4.

73. A PHOSITA would have identified *Prabhavalkar* as being within the same field of endeavor as the ’561 Patent because both *Prabhavalkar* and the ’561 Patent are both directed to speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3 For example, *Prabhavalkar* teaches a method for “receiving audio data for an utterance, providing features indicative of acoustic characteristics of the utterance as input to an encoder neural network, and processing an output of the encoder neural network using an attender neural network to generate a context vector” as well as

“generating speech recognition scores using the context vector and a decoder neural network.” EX1005-Prabhavalkar, 2:41-3:56. *See also* EX1005-Prabhavalkar, 1:15-25, 6:56-7:4, 15:16-22.

74. Further, a PHOSITA would also have found *Prabhavalkar* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Prabhavalkar* and the '561 Patent describe real-time audio modification. EX1001, Title, Abstract, 1:19-22, 4:28-30, Fig. 3; EX1005-Prabhavalkar, 17:33-51, 18:33-37, 20:16-21; 22:16-27.

75. Further, a PHOSITA would also have found *Prabhavalkar* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Prabhavalkar* and the '561 Patent describe feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3. Specifically, *Prabhavalkar* discloses that “[a] feature extraction module 116 processes the audio data 114 to extract (e.g., generate) a set of feature values that are indicative of acoustic characteristics of the utterance. For example, the feature values may be mel-frequency cepstral coefficients.” EX1005-Prabhavalkar, 11:36-41.



EX1005-Prabhavalkar, Fig. 2. *See also* EX1005-Prabhavalkar, 2:44-50.

D. Overview of *Sargsyan*

76. I have been informed U.S. Patent Appl. No. 2020/0066296 to Sargsyan et al. (“*Sargsyan*”) was filed on August 21, 2019. *Sargsyan* claims priority to provisional application No. 62/720,702 filed on August 21, 2018.

77. *Sargsyan* teaches “speech enhancement and noise suppression” EX1007-Sargsyan, Abstract. *Sargsyan* teaches to “expand low bitrate audio (e.g., 8 kHz) in conference calls on the server side and make it so that people calling in from low bitrate cellular lines sound much richer for other participants in the call (from 16 kHz to 44.1 kHz).” EX1007-Sargsyan, [0027].

78. A PHOSITA would have identified *Sargsyan* as being within the same field of endeavor as the '561 Patent because both *Sargsyan* and the '561 Patent are both directed to speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3; EX1007-Sargsyan, Abstract, [0003]-[0004], [0027], Fig. 14.

79. Further, a PHOSITA would also have found *Sargsyan* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Sargsyan* and the '561 Patent describe feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3; EX1007-Sargsyan, [0045], [0070], [0072].

E. Overview of *Kalinli*

80. I have been informed U.S. Patent Application No. 2012/0116756 to Kalinli ("*Kalinli*") was filed on November 10, 2010, and published May 10, 2012. Kalinli teaches "spoken language processing method for tone/intonation recognition," using parallel processing. EX1009-Kalinli, Abstract, [0048].

81. A PHOSITA would have identified *Kalinli* as being within the same field of endeavor as the '561 Patent because both *Kalinli* and the '561 Patent are both directed to speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3; EX1009-Kalinli, Abstract, [0048].

82. Further, a PHOSITA would also have found *Kalinli* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both

Kalinli and the '561 Patent describe feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3; EX1009-Kalinli, Abstract, [0048], [0050].

F. Overview of *Watts*

83. I have been informed U.S. Patent No. 7,319,959 to Watts ("*Watts*") was filed May 14, 2003, and published January 15, 2008.

84. *Watts* teaches "processing an audio signal" to determine "phoneme-level classification." EX1008-Watts, Abstract.

85. A PHOSITA would have identified *Watts* as being within the same field of endeavor as the '561 Patent because both *Watts* and the '561 Patent are both directed to speech audio processing. EX1001, Abstract, 1:19-22, Fig. 3; EX1008-Watts, Abstract, 1:16-17.

86. Further, a PHOSITA would also have found *Watts* reasonably pertinent to the problem faced by the inventors of the '561 Patent because both *Watts* and the '561 Patent describe feature extraction from audio data. EX1001, 4:21-23, 15:11-21, Fig. 3; EX1008-Watts, 4:37-42, 5:48-51, 6:21-23, Fig. 2.

VIII. SUMMARY OF UNPATENTABILITY

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

Ground 1: Claims 1-2, 6-7, 11-12, and 16-18 are obvious over <i>Feinauer</i> , <i>Prabhavalkar</i> and <i>Fan</i> .	1004, 1005, 1006
Ground 2: Claims 3, 8 and 13 are obvious over <i>Feinauer</i> , <i>Prabhavalkar</i> , <i>Fan</i> and <i>Sargsyan</i> .	1004, 1005, 1006, 1007

Ground 3: Claims 5, 10 and 15 are obvious over <i>Feinauer, Prabhavalkar, Fan</i> and <i>Watts</i> .	1004, 1005, 1006, 1008
Ground 4: Claims 4, 9, and 14 are obvious over <i>Feinauer, Prabhavalkar, Fan</i> , and <i>Kalinli</i> .	1004, 1005, 1006, 1009

IX. OPINIONS REGARDING GROUND 1: CLAIMS 1-2, 6-7, 11-12, AND 16-18 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, AND FAN*.

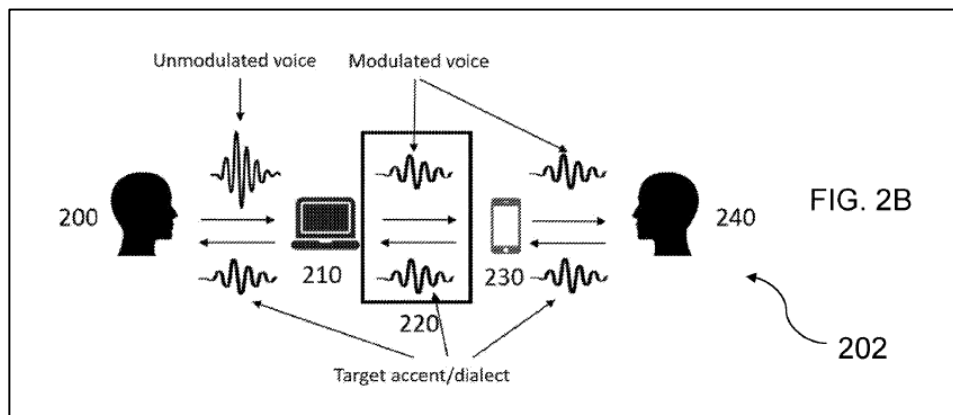
A. Claim 1

1. *Claim 1(Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”*

88. To the extent the preamble is limiting, *Feinauer* teaches a *real-time accent correction, the method implemented by a computing system* in an input audio signal² that receives an unmodulated voice signal, processes it, and selects “a target dialect and accent to use to modify [the] voice communication based on a context.” EX1004-*Feinauer*, [0001]-[0003], [0046], [0049]. For example, *Feinauer’s* accent conversion method and system may modify spoken words to within a spoken target accent strength threshold, such as from a Japanese accent to a German accent during

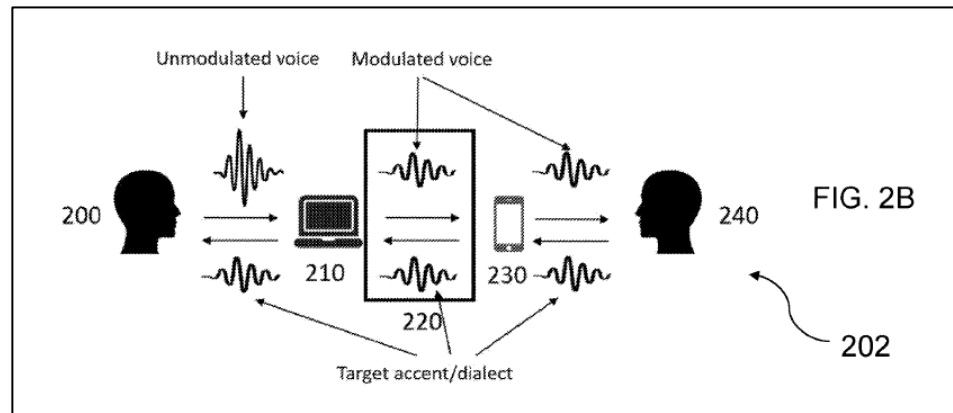
²Claim language is *italicized*. All drawing annotations added. Citations to a claim limitation and/or section number are to the limitation’s mapping in this Petition and are incorporated by reference, including the Motivations to Combine (MTC) and showing of reasonable expectation of success (REOS) provided in the cross-referenced section. All REOS include the benefits described for the respective MTC. All emphases added unless otherwise noted.

an audio/video call. EX1004-Feinauer, [0036], [0041], [0086]. As illustrated below, *Feinauer*'s method specifically “*modifies an accent* and/or dialect *in real-time* to sound more similar to an accent and/or dialect which would achieve a desired outcome.”



EX1004-Feinauer, [0035]-[0036], [0041], Fig. 2B. *Feinauer* explains its process may be performed within a “lag duration...that sounds like *real-time* to a user (e.g. 100ms).” EX1004-Feinauer, [0086]. It was well known that accent conversion may be performed in real time as demonstrated by *Feinauer*. As another example, *Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning* to Zhao et. al (“*Accentron*”) teaches, “accent conversion” where the “the inference can be accomplished in real time.” EX1010-Accentron, 33. A PHOSITA thus would have recognized *Feinauer* teaches a method for real-time correction of accent in a speech audio signal. A PHOSITA thus would have recognized *Feinauer* teaches a *real-time accent correction, the method implemented by a computing system*.

89. *Feinauer's* speech conversion process may be implemented as “software or programmable code embedded in a computer readable medium that is executed by a processor,” and is thus *implemented by a computing system*. For example, *Feinauer* teaches that “communication device 230 can be a computing device 210.” EX1004-Feinauer, [0041].



EX1004-Feinauer, Fig. 2B. *Feinauer* teaches that communication device 230 may be “any embodiments of device 101, described above in connection with FIG. 1.” EX1004-Feinauer, [0043].

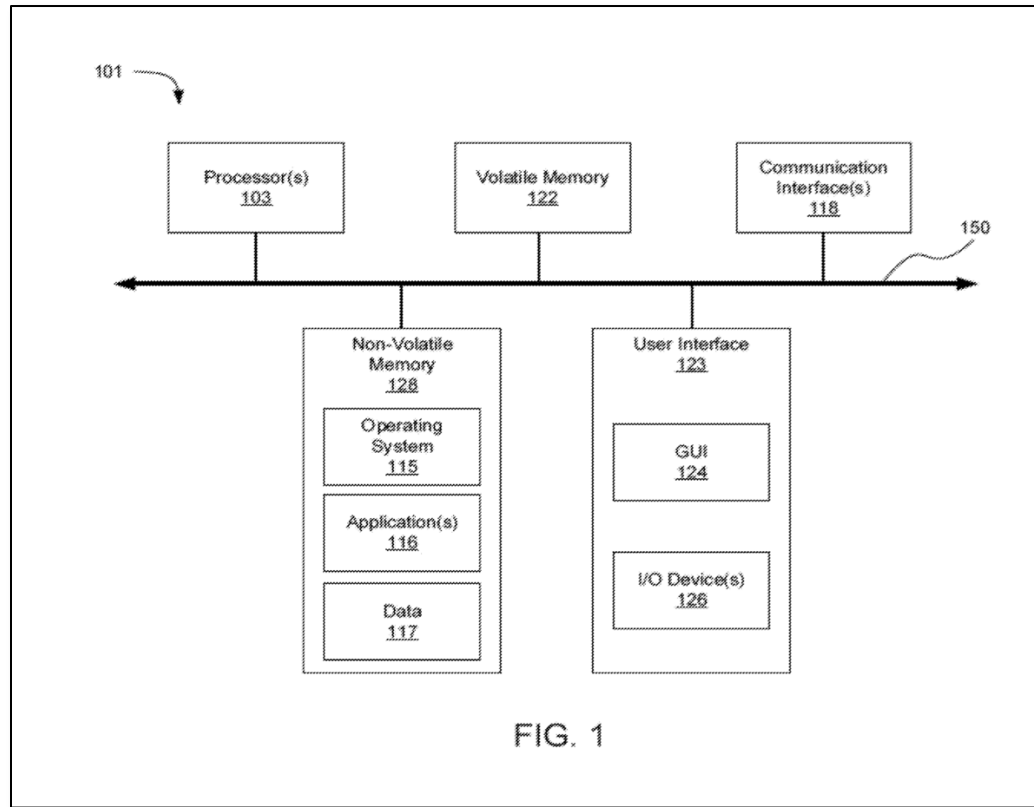


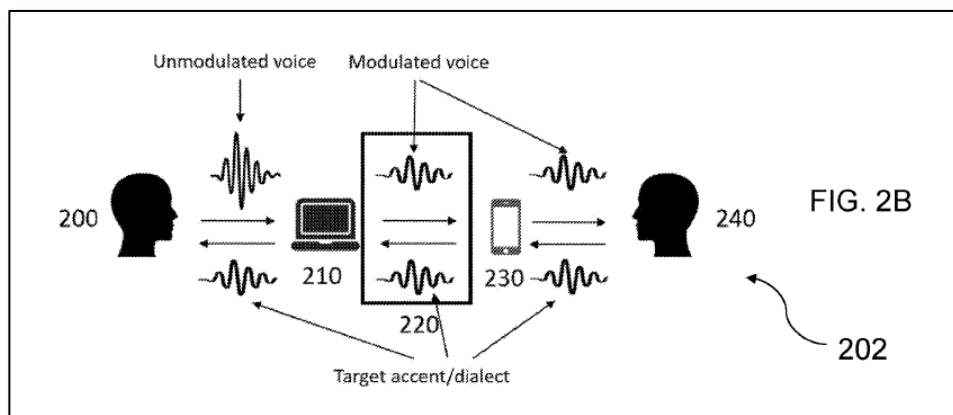
FIG. 1

EX1004-Feinauer, Fig. 1. *Feinauer's* computing device includes “one or more processors 103, volatile memory 122 (e.g., random access memory (RAM)), non-volatile memory 128” and processor “may be implemented by one or more programmable processors to execute one or more executable instructions, such as a computer program, to perform the functions of the system.” *See generally*, EX1004-Feinauer, [0031]-[0032]. *Feinauer's* function was to modify “an accent and/or dialect in real-time to sound more similar to an accent and/or dialect which would achieve a desired outcome.” EX1004-Feinauer, [0035]. Therefore, a PHOSITA would have recognized that *Feinauer* teaches *real-time accent correction, the method implemented by a computing system*.

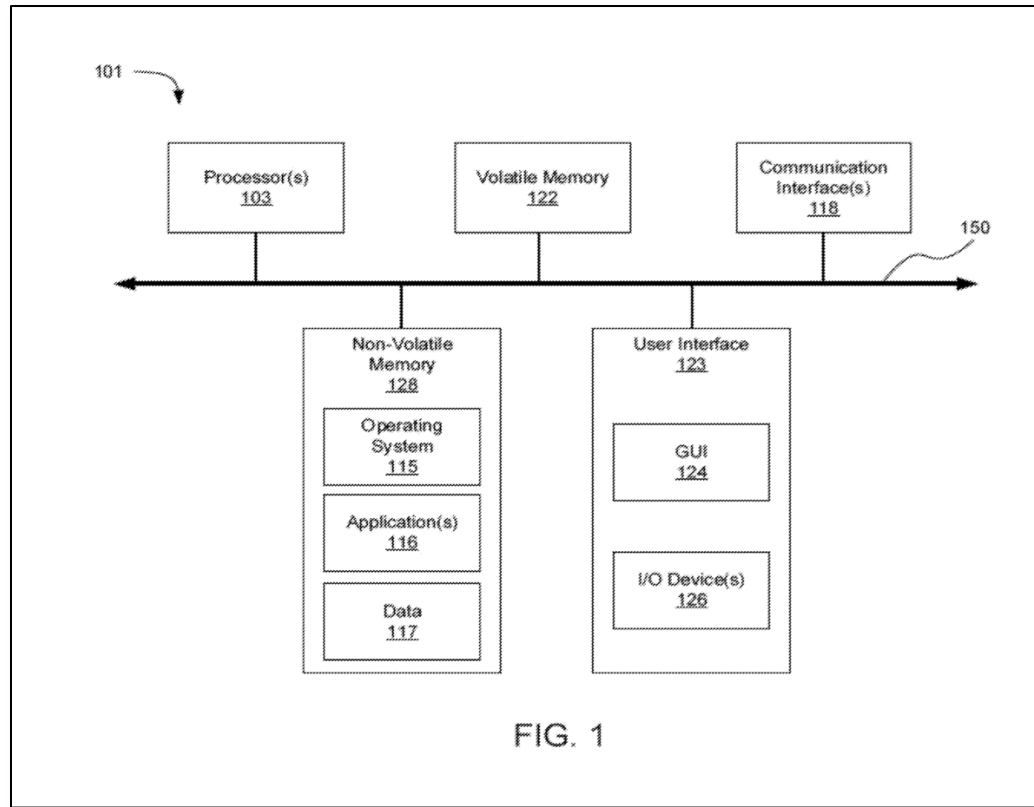
2. ***Claim 1(a)(1): “extracting, from a chunk of a stream of chunks of an input audio signal”***

a. *Feinauer’s Teachings*

90. *Feinauer* teaches receiving an unmodulated voice signal via a microphone. *Feinauer* teaches that communication device 230 “may include any type and form of hardware and software for receiving, detecting or **being provide[d] voice or audio as input, such as via a microphone**” and which includes “any type and form of hardware and software for playing, outputting or providing voice or audio as output, such as via a speaker.” EX1004-Feinauer, [0043]. *See also* EX1004-Feinauer, [0031]-[0032].



EX1004-Feinauer, Fig. 2B. Device 230 may be “any embodiments of device 101, described above in connection with FIG. 1” which includes “one or more input/output (I/O) devices 126 (e.g., ...a microphone...” EX1004-Feinauer, [0031], [0043].



EX1004-Feinauer, Fig. 1. It was well known to correct an accent spoken into a microphone. For example, *Accentron* teaches foreign accent conversion “has focused on acoustic methods, since they only require recording speech with a microphone.” EX1010-Accentron, 5.

91. *Feinauer* further teaches data may “represent...or contain...audio of conversation(s), such as audio fragments.” EX1004-Feinauer, [0053]. This audio fragment is represented by an “unmodulated voice or signal waveform. The unmodulated voice or signal waveform can be received by communication device 210 and transmitted via the communication network 220 without modification.” EX1004-Feinauer, [0044]. *Feinauer*’s device receives in real-time data in the form

of “data packets” that include unmodulated voice signal audio fragments from “I/O device(s)” which as discussed above includes a microphone. EX1004-Feinauer, [0031], [0044], [0053]. Therefore, a PHOSITA would have understood that *Feinauer* teaches receiving an unmodulated voice signal (*speech audio signal*) via a microphone. *See also* EX1004-Feinauer, [0042]-[0046], [0056].

92. *Feinauer* is clear that thereafter the unmodulated voice signal is processed “to provide a modulated voice with a target accent.” EX1004-Feinauer, [0045], Figs. 2B. Though *Feinauer* expressly states its method “***may modify the accent*** and/or dialect of a conversation, or portions thereof, ***on any level of granularity***, including but not limited on a per sound basis, per word basis, set or string of words basis, on a whole or portion of a sentence structure basis,” *Feinauer* provides no further detail regarding how its input unmodulated voice signal is divided into these “portions” for processing. *Feinauer* discloses “the accent and dialect modifier 318 can determine that certain data portions of the audio content of data packet including a conversation, a conversation fragment, a sentence or a single word should be modified.” EX1004-Feinauer, [0054], [0057]. It was well known in the prior art that in order to perform accent conversion in real time, a conversion should be processed in “portions.” As an example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”) teaches “a digital representation of a portion of an interaction may be analyzed in real-time” to maintain “low real

time delay” EX1018-Wasserblat, Abstract, [0025]. In sum, a PHOSITA would have looked to *Prabhavalkar* for detailed description of implementing known methods for dividing a voice signal into input chunks.

b. *Prabhavalkar*’s Teachings

93. *Prabhavalkar* describes methods of speech processing using neural networks. EX1005-Prabhavalkar, 1:15-25, 2:41-3:56. *Prabhavalkar*’s techniques are applicable to a variety of speech processing applications, including using neural networks to form a sequence-to-sequence model that “performs acoustic modelling, pronunciation modelling, and language modelling.” EX1005-Prabhavalkar, 3:22-24. Notably, *Prabhavalkar* specifically includes embodiments comprising a “sequence-to-sequence model without finite state transducers, a lexicon, or a text normalization module.” EX1005-Prabhavalkar, 3:25-29. A PHOSITA would have appreciated that *Prabhavalkar* teaches a quick, non-textual, real-time under 300ms latency, speech processing system. EX1005-*Prabhavalkar*, 22:24-26. A PHOSITA would have appreciated that it was common in the prior art for real-time speech processing systems to not translate received speech audio to text as demonstrated by *Prabhavalkar*, allowing the system to provide a quick, low-latency, real-time speech modification experience to the user. *See, e.g.*, EX1010-Accentron, 14; 33.

94. *Prabhavalkar*’s system includes a computer system and microphone that receives an utterance from a user and generates audio data representing the

utterance. EX1005-*Prabhavalkar*, 11:19-45, Fig. 2. *Prabhavalkar* then determines, “from the audio data for the utterance, *speech frames*,” and “successively processing *chunks* that each include different sets of speech frames with the speech recognition model.” EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. It is well known in speech processing to divide input audio into speech frames that include speech content from a speech utterance. Often these frames of speech are on the order of 10-40ms in length each, such as in *Prabhavalkar* where the speech frames are envisioned as being 30ms long. EX1005-*Prabhavalkar*, 15:18-20. *Prabhavalkar* explicitly envisions “*each chunk comprising a same predetermined number of speech frames*, each chunk having a first predetermined number of speech frames representing speech.” EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Here, *Prabhavalkar* is describing a process by which incoming speech is divided into a string of 30ms-long speech frames. Every five (5) of these speech frames are apportioned into a 150ms chunk (much like *Feinauer’s* “portions”) for subsequent processing. EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Because every chunk has 5 frames, *Prabhavalkar’s* chunks include a *predetermined number* of frames of the speech audio signal. The chunks are processed “successively,” that is, one-at-a-time, though successive processing such does not preclude content from one chunk being used and/or considered relative to another chunk, for example to provide context between chunks. Thus, *Prabhavalkar* teaches a *speech audio signal is divided into*

a stream of input chunks, each including 5 predetermined (pre-defined) frames of the speech audio signal. EX1005-*Prabhavalkar*, 20:14-20, 22:1-38.

95. *Prabhavalkar* characterizes its approach as “streaming,” wherein speech is processed “while audio data of an utterance is still being received and without waiting for the utterance to be completed.” EX1005-*Prabhavalkar*, 6:4-23, 18:28-41, 2:14-40. *Prabhavalkar* teaches that “streaming can be considered to mean that the system has a maximum allowable delay of 300 ms.” EX1005-*Prabhavalkar*, 18:32-35. A PHOSITA would have appreciated that one example of “streaming” occurs with an “audio and/or video conversation” as taught by *Feinauer*. EX1004-*Feinauer*, [0041]. Therefore, as also discussed in paragraph 146, a PHOSITA would have thus recognized *Prabhavalkar*’s system as performing a *real-time* speech processing method, like *Feinauer*.

c. Motivation to Combine *Feinauer* and *Prabhavalkar*

96. A PHOSITA would have been motivated to employ *Prabhavalkar*’s method of dividing audio into input chunks as an operable method of performing sampling (dividing speech into “portions”) in *Feinauer*’s real-time accent conversion method. *Feinauer* already teaches that “[t]he accent and dialect modifier 318 may modify the accent and/or dialect of a conversation, or portions thereof, on any level of granularity, including but not limited on a per sound basis, per word basis, set or string of words basis, on a whole or portion of a sentence structure basis.” EX1004-

Feinauer, [0057]. A PHOSITA would have understood that all this combination would entail is making sure there are frames within Feinauer's portions, which was well within a PHOSITA's skillset. A PHOSITA would have made any modifications necessary to make this modification operable since this combination was well within a PHOSITA's skillset.

97. Prabhavalkar further gives details that a PHOSITA would have been motivated to look towards for this idea of "portions" that Feinauer teaches such that "speech frames each representing audio characteristics over different periods having a same duration; successively processing chunks that each include different sets of speech frames with the speech recognition model, wherein the speech recognition model is configured to predict a variable number of word elements for each chunk processed, wherein each chunk corresponds to a time step of processing of the speech recognition model, each chunk comprising a same predetermined number of speech frames..." EX1005-Prabhavalkar, 6:56-7:4. See also EX1005-Prabhavalkar, 15:16-22, 17:33-51, 18:18-41, 22:1-38.

98. A PHOSITA would have been motivated by Prabhavalkar's express statement that a chunk-based approach may "limit the amount of latency incurred," increasing speed and performance compared to the unmodified Feinauer method. EX1005-Prabhavalkar, 17:47. Specifically, Prabhavalkar teaches that in a traditional LAS system, there is the "limitation that the entire utterance must be seen by the

encoder, before any labels can be decoded...” EX1005-Prabhavalkar, 17:33-36. One way to “address this limitation” is to “operate on segments of an utterance less than the full utterance.” EX1005-Prabhavalkar, 17:37-43. This is referred to in Prabhavalkar as “a “chunk”-based approach” to “process attention on small segments of the utterance, such as a certain number of frames or a certain duration of the utterance at a time.” EX1005-Prabhavalkar, 17:43-46. To do this, Prabhavalkar’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance while a second chunk or segment of the utterance is still being spoken.” EX1005-Prabhavalkar, 17:46-51.

99. A PHOSITA would have recognized that applying Prabhavalkar’s chunking techniques to Feinauer’s system would have only required a reprogramming according to known techniques to produce the reasonable result of reducing the latency of Feinauer’s accent conversion method by processing frames in chunks having a predetermined number of frames, such as 5 or 10. It was well known to process audio signals utterances in smaller portions, especially to process audio in real-time as taught by Feinauer. EX1004-Feinauer, [0057], [0086]. For example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“Wasserblat”) teaches “a digital representation of a portion of an interaction may be analyzed in

real-time” to maintain “low real time delay” EX1018-Wasserblat, Abstract, [0025].

A PHOSITA would have had a REOS at least because Prabhavalkar expressly describes the processor architecture necessary for implementing its method and extols its method’s use in speech applications, requiring only ordinary skill of a PHOSITA to combine. EX1005-Prabhavalkar, 17:15-16. (“an attention-based model can be designed for sequence-to-sequence speech recognition”) Prabhavalkar provides extensive support for processor architecture capable of implementing its method. For example, Prabhavalkar teaches the “processor 552 can process instructions for execution within the computing device 550, including instructions stored in the memory 564. The processor may also include separate analog and digital processors. The processor may provide, for example, for coordination of the other components of the device 550, such as control of user interfaces, 60 applications run by device 550, and wireless communication by device 550. Processor 552 may communicate with a user through control interface 558 and display interface 556 coupled to a display 554.” EX1005-Prabhavalkar, 25:55-65. *See also* EX1005-Prabhavalkar, 6:56-7:4, 15:16-22.

100. To the extent necessary to implement Prabhavalkar’s chunking approach, a PHOSITA would have incorporated the enabling hardware from Prabhavalkar into Feinauer’s system to result in an operable real-time accent

modification system performing chunking using ordinary skill without undue experimentation.

3. Claim 1(a)(2): “extracting, ... acoustic features”

101. The '561 Patent provides examples of “acoustic features” including “pitch 206 (or main frequency (F0)), energy 208 (signal amplitude), and voice activity detection.” EX1001, 6:35-37, Fig. 2. As I described in the background section, it was common to measure all of these acoustic features in speech audio signals. See Section IV.B.

a. Feinauer’s Teachings

102. *Feinauer* teaches extracting “prosodic vocal features and characteristics” from input speech audio. EX1004-Feinauer, [0099], Fig. 6 (step 620); As discussed above regarding Claim 1(a)(1), a PHOSITA would have modified *Feinauer*’s speech “portions” to be *input chunks* per *Prabhavalkar*. See Section V.

103. *Feinauer* as modified by *Prabhavalkar* teaches extracting “prosodic vocal features and characteristics” from chunks of input speech audio. EX1004-Feinauer, [0099], Fig. 6 (step 620). *Feinauer* further envisions prosodic features including “any voice characteristic such as *pitch, tone, volume, etc.*” EX1004-Feinauer, [0057]. A PHOSITA would have understood pitch, tone, and volume to each satisfy “acoustic features” of the input chunk of user’s speech. A PHOSITA

would have understood that an example of acoustic features are prosodic features. *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches that prosody “refers to the spoken style of discourse independent of lexical content[.]” EX1021-Schmandt, 301-302. *Schmandt* further teaches stress, pitch, duration, and intonation are all examples of prosody speech style. EX1021-Schmandt, 301-302. As another example, to calculate loudness, the energy of a signal is measured over time. For example, *Loudness, Its Definition, Measurement and Calculation* to Fletcher et al. (“*Fletcher*”) teaches the “sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point.” EX1025-Fletcher, 83. This sound intensity (energy) is referred to as the “loudness level” EX1025-Fletcher, 83. In order to maintain or modify the voice characteristics identified in [0057], *Feinauer* must extract them, but does not provide further detail.

b. Fan’s Teachings

104. *Fan* teaches a voice conversion method for “adding an accent to or removing an accent from” a user’s voice. EX1006-Fan, Abstract. Like *Feinauer*, *Fan* teaches modifying acoustic features such as the user’s tone, pitch, rhythm, and speaking speed which are examples of prosody speech style. EX1006-Fan, [0050], [0106], EX1021-Schmandt, 301-302. As previously discussed at paragraph 30-31, *Schmandt* teaches that prosody (including features alternatively referred to as

acoustic features) “refers to the spoken style of discourse independent of lexical content” such as stress, pitch, duration, rhythm, meter, and intonation. EX1021-Schmandt, 17, 301-302. Accordingly, a PHOSITA would have understood that the rhythm and speaking speed, defined broadly as prosody, are acoustic features.

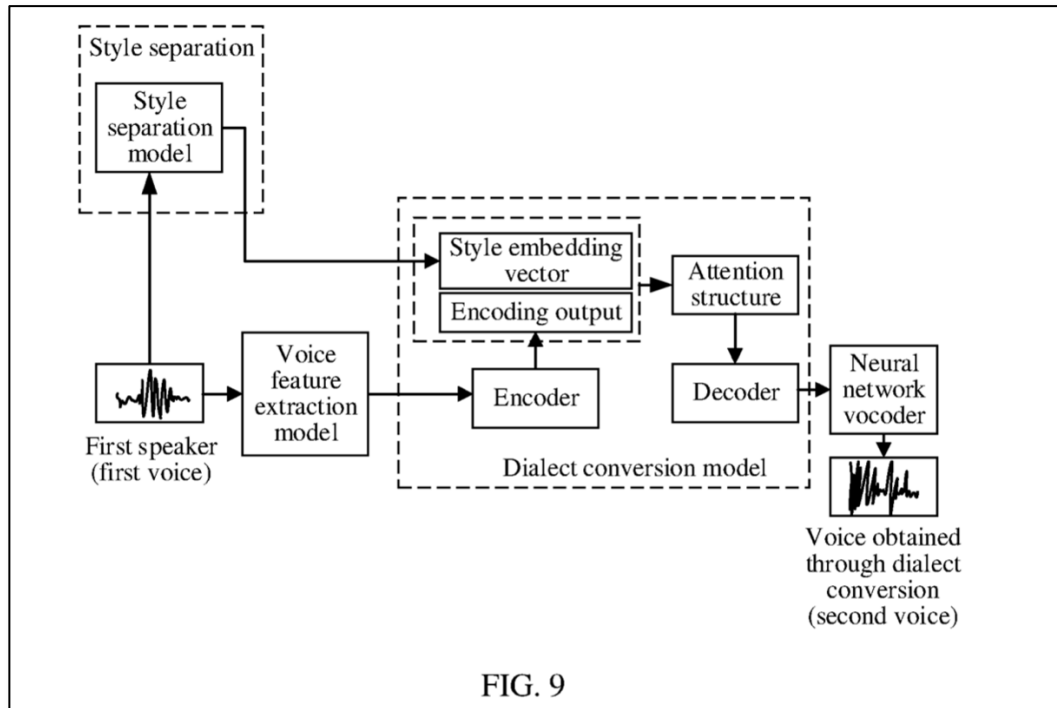
105. *Fan* expressly achieves its accent modification using “hardware or the software includes one or more **modules**” including a plurality of sub-feature extraction models, such as a style separation model. EX1006-Fan, [0023], [0089]-[0090], Fig. 5. I note the description of the style enhancement model is discussed in *Fan* with regard to an embodiment employing a “third voice,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable all of *Fan*’s envisioned embodiments.

c. Motivation to Combine *Feinauer-Prabhavalkar* and *Fan*

106. A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module (such as hardware employing *Fan*’s style separation model) as an operable method of extracting acoustic features including pitch, tone, volume, rhythm, and/or speaking speed in *Feinauer*’s real-time accent conversion

method. A PHOSITA would have appreciated the lengthy detail provided by Fan regarding hardware, software, and neural network configurations capable of extracting acoustic features, leading to a REOS. EX1006-Fan, [0045] (“the voice processing neural network...includes...a dialect conversion network”); [0047] (“Sequence to sequence (sequence to sequence, seq2seq) neural network: seq2seq belongs to an encoding (encoder)-decoding (decoder) structure.”); [0050] (“Style feature: The style feature includes but is not limited to a timbre. The timbre refers to quality of a voice, which is referred to as tone quality. The timbre can reflect unique quality of a sound made by a speaker. The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.”); [0090] (“Optionally, the style conversion network further includes a plurality of sub-feature extraction models and a multihead attention structure. The terminal inputs the third voice into the sub-feature extraction model, and extracts a vector of the sub-feature by using the sub-feature extraction model. For example, the plurality of sub-feature extraction models include a prosody extraction model, an accent extraction model, and a speaking speed extraction model. For example, the prosody extraction model is for extracting a prosody feature of the third voice, to obtain a prosody vector. The accent extraction model is for extracting an accent feature of the third voice, to obtain an accent vector. The speaking speed extraction model is for extracting a speaking speed feature of the third voice, to obtain a speaking speed vector.”); [0094] (“For example, a

structure output by the multihead attention mechanism may be a 256-dimensional style embedding vector (a style feature).”); [0127] (“Optionally, the terminal inputs a third voice into a style separation model, and separates a style feature of the first voice by using the style separation model, to obtain a timbre feature vector of the first voice. Then, the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation. Content of the second voice is the same as content of the input voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”) [0152]-[0155] (“The obtaining module 1330 is configured to obtain a third voice for extracting a style feature. The processing module 1310 is further configured to input the third voice into the style separation model, and separate the style feature of the third voice by using the style separation model. The processing module 1310 is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion, to obtain the second voice. In an optional implementation, the style feature includes a first feature, and the first feature includes a plurality of sub-features.”); *see also* [0166]-[0167], [0170]-[0171], [0090]-[0093], [0046], [0048]-[0049], Fig. 5, Fig. 7.



EX1006-Fan, Fig. 9. *Feinauer* already teaches machine learning and neural network capabilities. EX1004-Feinauer, [0036]-[0037] (“a neural network machine learning model can detect a correlation between accent features and their success in achieving the optimal desired outcome in the specified context”), [0039] (“a neural network machine learning model can detect a correlation between dialect features and their success in achieving the optimal desired outcome in the specified context”), [0098] (“the features of the vocal characteristics in each category can be analyzed by a supervised neural network and the identified vocal characteristics correlated with their respective accent and dialect strength category labels.”). A PHOSITA would have recognized the modification constitutes applying a known technique (*Feinauer* and *Fan*’s steps of acoustic feature extraction) to a known device (*Fan*’s hardware

and software modules) with predictable results, further motivating the combination. *Feinauer* already teaches the “machine learning model 338 can extract vocal features of pronounced utterances” and that vocal features are “prosodic” EX1004-Feinauer, [0067]; [0095] (“prosodic vocal features and characteristics of the dialects can be extracted from the audio samples”); [0099]; [0100].

4. Claim 1(a)(3): “extracting...linguistic features, wherein the linguistic features have a reduced accent;”

107. As discussed above, *Feinauer-Prabhavalkar-Fan* renders obvious extracting acoustic features from an input chunk. Further, *Feinauer* teaches identifying “individual words” in the user’s speech, and may determine “which words” to be accent-modified. EX1004-Feinauer, [0083] (“an algorithm will measure each individual word in the conversation to determine which words to modify”), [0094] (“an algorithm will measure each individual word in the conversation to determine which words to modify.”), [0099] (“At step **620**, the accent and dialect modifier can define feature values of words. For example, prosodic vocal features and characteristics of individual **word dialects can be extracted from the individual word audio samples** using preprocessing tools to convert the audio sample waveforms into a spectrogram of MFCC. In some embodiments, the accent and dialect modifier can query the database to obtain instructions, data or information to define feature values of words. In some embodiments, the accent and dialect modifier can define the feature values of all the

words in a conversation, conversation packet or voice data packet.”). A PHOSITA would have recognized that *Feinauer* thus extracts the linguistic features of words, but does not provide further detail regarding how linguistic features are extracted, or by which module.

108. The '561 Patent states “linguistic features 210 can be represented as Phonetic PosteriorGrams (PPGs).” EX1001, 7:6-26, Fig. 3.

a. Fan’s Teachings

109. *Fan* teaches a voice conversion method for “adding an accent to or removing an accent from” a user’s voice. EX1006-Fan, Abstract.

110. *Fan* teaches a voice feature extraction model that extracts phoneme posteriorgrams (PPGs) from the user’s speech representing the words spoken, as their phonemic constituents. EX1006-Fan, [0007] (“The terminal inputs the feature information of the first voice into a voice feature extraction model, and extracts a phoneme posteriorgram PPG feature of the first voice by using the voice feature extraction model, where the PPG feature is for retaining the content information of the first voice. The PPG feature describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and is equivalent to recognizing the phoneme. The PPG feature is for retaining the content information of the first voice. In this example, the voice does not need to be converted into a text to retain the content information of the voice, instead, voice input is directly performed, and the

content information of the first voice is retained by using the PPG feature, so that robustness can be improved.”), [0085]-[0086] (“The terminal inputs the first voice into the voice feature extraction model, and extracts the PPG feature of the first voice by using the voice feature extraction model.”), 100, Fig. 5, Fig. 9, [0126]-[0127], [0150] (“In an optional implementation, the processing module **1310** is further configured to input the feature information of the first voice into a voice feature extraction model, and extract a phoneme posteriorgram PPG feature of the first voice by using the voice feature extraction model, where the PPG feature is for retaining content information of the first voice.”). Fan’s PPGs are linguistic features because they represent phoneme probabilities and phonemes are a linguistic feature of any language. EX1006-Fan, [0085] (“The PPG feature describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and is equivalent to recognizing the phoneme.”). *See* EX1031-Zhao, 2368lhc.

111. As discussed, *Fan* expressly performs its methods using “hardware or the software includes one or more *modules*” including a plurality of sub-feature extraction models, such as a voice feature). EX1006-Fan, [0023], [0085], Fig. 5. I note the description of a voice feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style

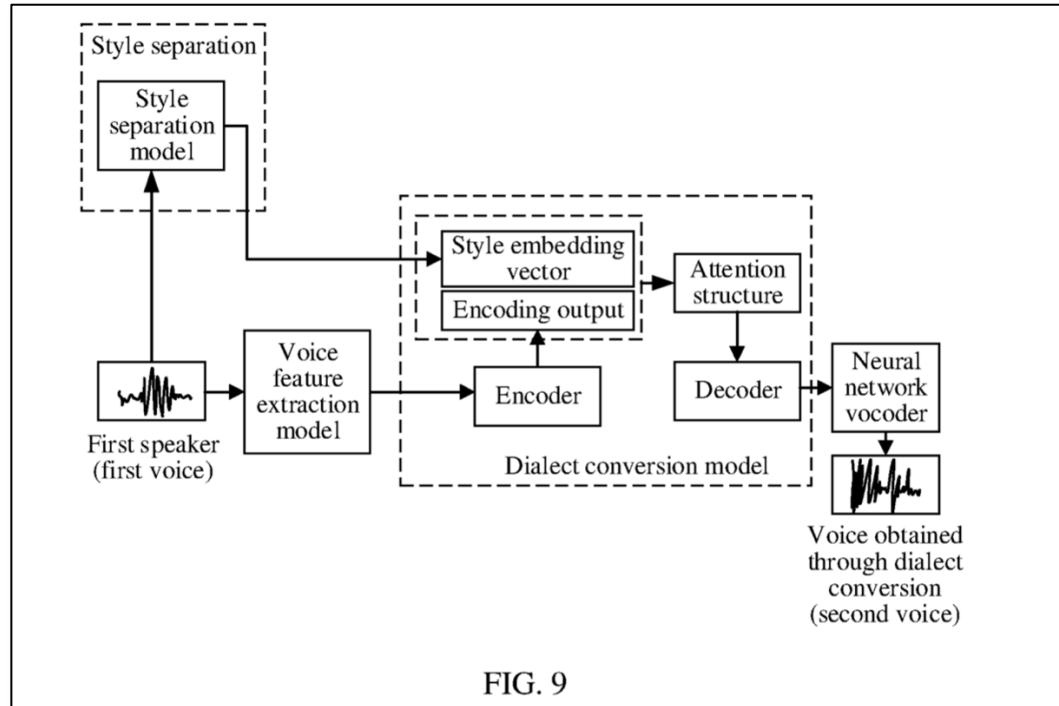
enhancement model in regard to a voice feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable to all of *Fan*’s envisioned embodiments. A PPG “describes a probability of each phoneme in a phoneme set corresponding to a voice frame, and is equivalent to recognizing the phoneme. The PPG feature is for retaining the content information of the first voice.” EX1006-Fan, [0007], [0085]. As supported by the prior art and the ’561 specification, a PHOSITA would have recognized that PPGs extracted from a user’s speech carry a reduced accent. *See* EX1001, 7:19-21, 15:22-32; EX1012-Ding, Abstract. Specifically, as supported by *Ding*, PPGs are mostly speaker-independent, but do carry some accent that may leak through to subsequently-generated speech. *Id.* For example, *Ding* teaches most methods “implicitly assume that PPGs are speaker-independent and contain only linguistic information in an utterance. In practice, however, PPGs carry speaker individuality cues, such as accent, intonation, and speaking rate. As a result, these cues can leak into the voice conversion, making it sound similar to the source speaker.” EX1012-Ding, Abstract. Therefore, it was well known that PPGs will still retain some accent within the PPG, though at a reduced amount. Therefore, PPGs satisfy *linguistic features with a reduced accent*. Thus, *Prabhavalkar-Fan*

renders obvious *extracting linguistic features with a reduced accent from an input chunk*.

b. Motivation to Combine *Feinauer-Prabhavalkar* and *Fan*

112. A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module (such as hardware employing *Fan's* voice feature extraction model) as an operable method of extracting linguistic features including PPGs in *Feinauer's* real-time accent conversion method. A PHOSITA would have appreciated the lengthy detail provided by *Fan* regarding hardware, software, and neural network configurations capable of extracting linguistic features, leading to a REOS. EX1006-Fan, [0045] (“the voice processing neural network...includes...a dialect conversion network”); [0047] (“Sequence to sequence (sequence to sequence, seq2seq) neural network: seq2seq belongs to an encoding (encoder)-decoding (decoder) structure.”); [0085] (“the voice feature extraction model may be a deep neural network.”); [0127] (“Optionally, the terminal inputs a third voice into a style separation model, and separates a style feature of the first voice by using the style separation model, to obtain a timbre feature vector of the first voice. Then, the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation. Content of the second voice is the same as content of the input

voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”); [0152]-[0155] (“The obtaining module 1330 is configured to obtain a third voice for extracting a style feature. The processing module 1310 is further configured to input the third voice into the style separation model, and separate the style feature of the third voice by using the style separation model. The processing module 1310 is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion, to obtain the second voice. In an optional implementation, the style feature includes a first feature, and the first feature includes a plurality of sub-features.”); see also [0166]-[0167]; [0170]-[0171]; [0090]-[0093]; [0046], [0048]-[0049]; Figs. 5, 7.



EX1006-Fan, Fig. 9. *Feinauer* already teaches machine learning and neural network capabilities. EX1004-Feinauer, [0036]-[0037] (“a neural network machine learning model can detect a correlation between accent features and their success in achieving the optimal desired outcome in the specified context”), [0039] (“a neural network machine learning model can detect a correlation between dialect features and their success in achieving the optimal desired outcome in the specified context”), [0098] (“the features of the vocal characteristics in each category can be analyzed by a supervised neural network and the identified vocal characteristics correlated with their respective accent and dialect strength category labels.”). A PHOSITA would have recognized the modification constitutes applying a known technique (*Feinauer* and *Fan*’s steps of linguistic feature extraction) to a known device (*Fan*’s hardware and software modules) with predictable results, further motivating the combination. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan*’s modules described functionality in *Fan*’s style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. A PHOSITA would have recognized PPGs were a well-known and widely-used tool for extracting linguistic features, and would have expected a speech modification system to be capable of extracting PPGs, further motivating the combination.

5. Claim 1(b): “synthesizing a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker; and”

113. As discussed in Claim 1(a)(2)-1(a)(3), *Feinauer-Prabhavalkar-Fan* teaches extracting acoustic features and linguistic features. *Feinauer-Prabhavalkar-Fan* teaches a speaker embedding for a human user and renders obvious synthesizing a spectrum representation based on the above.

114. The '561 Patent describes a speaker embedding as a “multidimensional vector” that is “pretrained based on audio data including a recorded speech of a target speaker having a further accent.” EX1001, 15:33-38. The '561 Patent further describes the speaker embedding may include other “information on target speakers’ voice style not related to the accent of the target speaker.” EX1001, 11:39-41. A speaker embedding may be trained fixed, pre-trained fixed, or extracted via a pre-trained model from a speech audio signal in real-time. EX1001, 11:42-53. The speaker embedding may be utilized in such a way that voice acoustic features corresponding to the speaker embedding match voice acoustic features of a target speaker. EX1001, 11:42-53.

115. *Feinauer* teaches adjusting the accent of the user to match a target accent. EX1004-Feinauer, [0045]-[0046], [0096], [0103]-[0110], Figs. 2B, 7. *Feinauer’s* system loads a relevant “translation model” corresponding to the particular target accent. EX1004-Feinauer, [0081], (“step 504, the accent and dialect

modifier preloads relevant translation models and translation dictionaries. For example, the relevant translation models and translation dictionaries can be installed on a local machine, such as computing device 210, 230”), [0105] (“step 704, the accent and dialect modifier loads translation models and dictionaries”), [0110], Fig. 7. *Feinauer’s* translation model determines “attributes of a user before the target word was spoken so that their voice profile can be replicated in the real time modification and the replaced word flows with the vocal biometric profile of the user.” EX1004-Feinauer, [0038]. *Feinauer’s* translation model includes “English received pronunciation with the American way of emphasizing certain words.” EX1004-Feinauer, [0065]. *Feinauer* teaches that “English received pronunciation may be recognized as being more persuasive or trustworthy, but American clients may understand familiar American words, such as brand names, more easily in their native American accent compared to with an English accent.” EX1004-Feinauer, [0065].

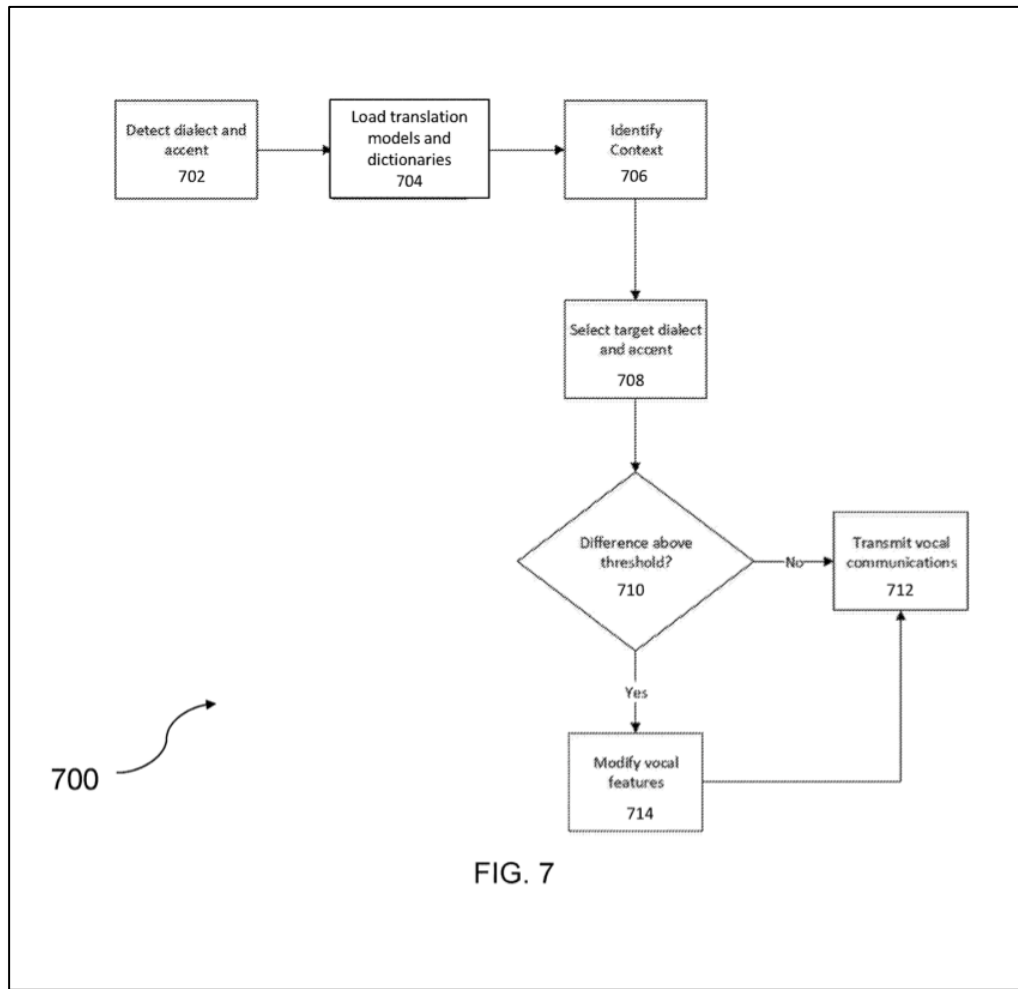


FIG. 7

EX1004-Feinauer, Fig. 7.

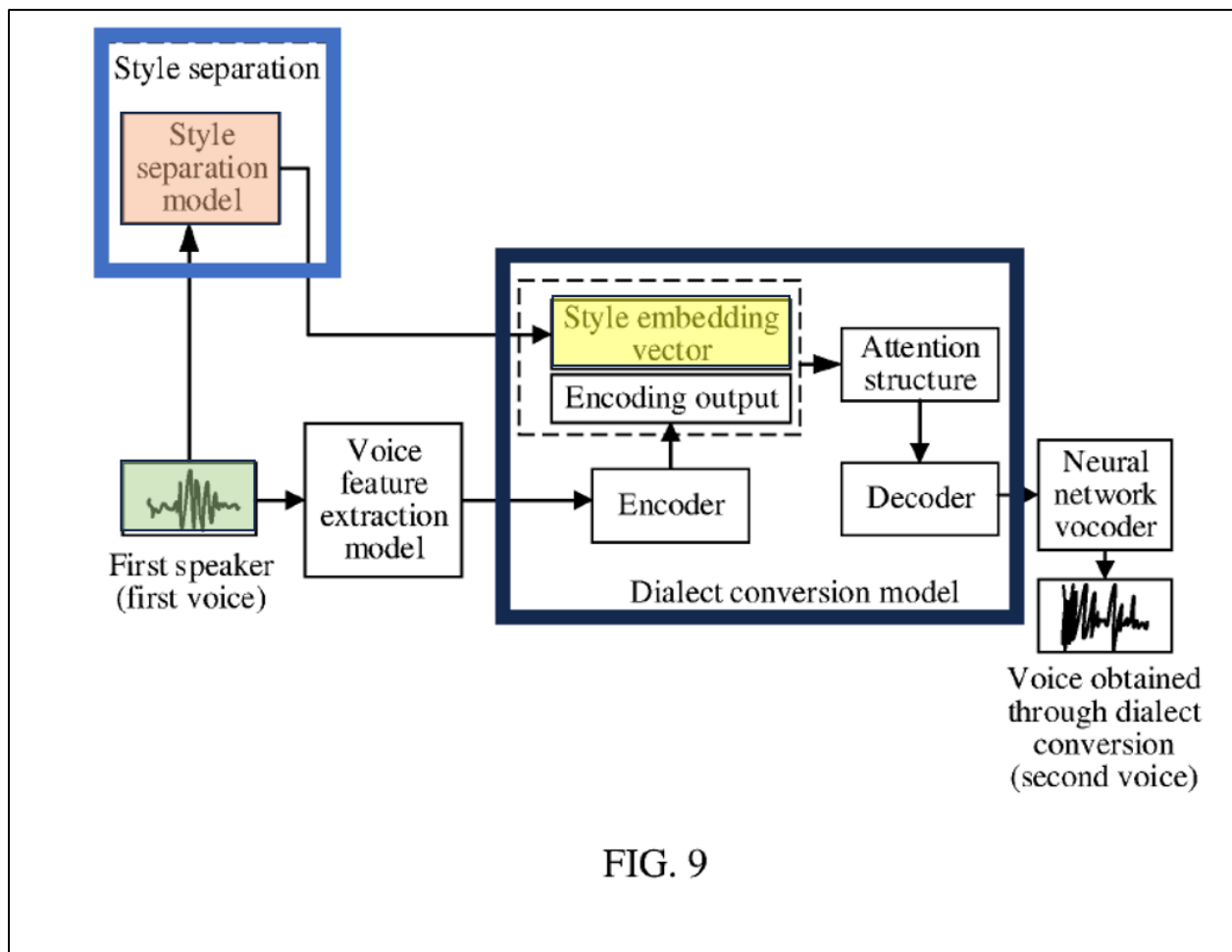
a. *Fan's Teachings*

116. Similarly, *Fan* teaches that a user may modify the particular accent of input source speech, such as to a “Sichuan accent” or a “Guangdong accent.”

EX1006-Fan, [0120], [0124], [0128]-[0130]. *Fan* envisions a “style embedding vector” that a PHOSITA would have recognized includes a representation of a target accent. EX1006-Fan, [0050], [0094], [0127], Figs. 7, 9, 12. It was well known in the prior art that a “speaker embedding vector” at least includes “timbre” as taught by

Fan. EX1010-Accentron, 21 (“speaker embedding mainly encodes information such as voice quality/timbre and pitch”).

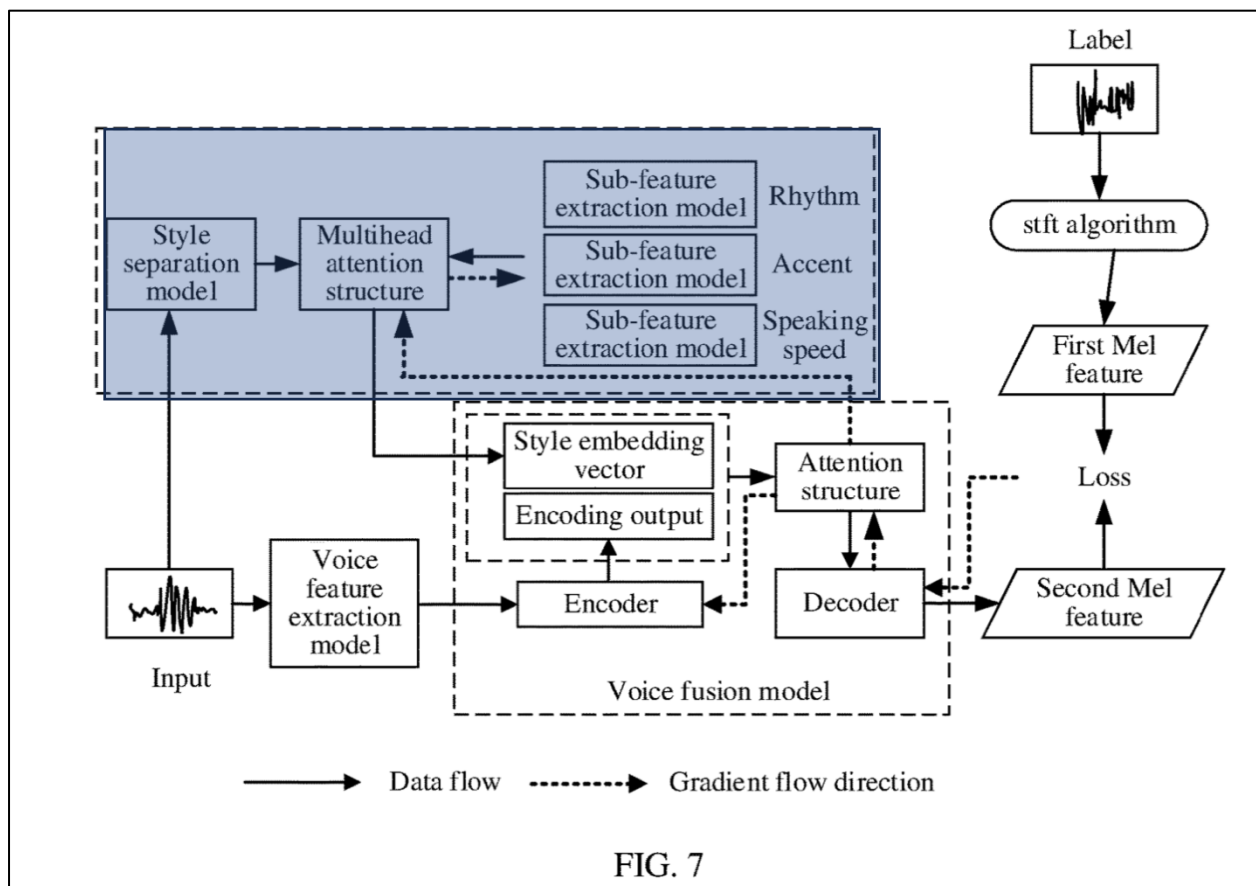
117. Specifically, *Fan* explains a target voice (annotated green below) may be input to a style separation model (annotated red). EX1006-Fan, [0093]-[0094], [0127].



EX1006-Fan, Fig. 9 (annotated) *see also* Figs. 7, 12. The style separation model extracts features from the target voice to output a style embedding vector (annotated

yellow) that represents the timbre of the input (target) voice. EX1006-Fan, [0094], Figs. 7, 9, 12.

114. *Fan* teaches that the features from the target voice are extracted in the style separation model and put into a “multihead attention structure” to enable “a model to extract a feature expression from different subspaces.” EX1006-Fan, [0094], Figs. 7, 9, 12. (annotated in blue) An output from the “the multihead attention mechanism includes some timbres of the third voice” in the form of a “256-dimensional style embedding vector (a style feature).” EX1006-Fan, [0094].



EX1006-Fan, Fig. 7. *Fan*'s timbre reflects the “unique quality of a sound made by a speaker.” EX1006-Fan, [0050].

115. *Fan* explicitly states that “timbre” comprises the accent of the voice. EX1006-Fan, [0050] (“The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.”). Thus, *Fan*'s style embedding vector (used in its dialect conversion model) represents the accent of a target speaker. EX1006-Fan, [0050], [0094], [0127], Figs. 7, 9, 12. *Fan* teaches this style feature includes “but is not limited to a timbre. The timbre refers to quality of a voice, which is referred to as tone quality. The timbre can reflect unique quality of a sound made by a speaker. The timbre includes but is not limited to a rhythm, an accent, and a speaking speed.” EX1006-Fan, [0050]. It was well known in the prior art that a “speaker embedding vector” at least includes “timbre” as taught by Fan. EX1010-Accentron, 21 (“speaker embedding mainly encodes information such as voice quality/timbre and pitch”). For these reasons, *Fan*'s style embedding vector teaches a *speaker embedding*.

116. In the exemplary embodiment corresponding to Figure 9, *Fan* describes linguistic and acoustic features combined with its speaker embedding (style embedding vector, yellow above) to produce a modified output. EX1006-Fan, [0126]-[0129]. Specifically, *Fan* explains the PPG feature (*e.g.* linguistic content) of the first (input) voice and style features (*e.g.* acoustic content) of the first (input) voice are input into the dialect conversion model (annotated with a black box)

corresponding to the selection operation, and the second (modified) voice is output by using the dialect (accent) conversion model corresponding to a style embedding vector (speaker embedding). EX1006-Fan, [0127], Fig. 9.

b. Motivation to Combine *Fan*'s speaker embedding

117. In the proposed modified method, a PHOSITA would have been motivated and found it obvious to employ the speaker embedding taught by *Fan* as part of the “translation model” within *Feinauer*'s voice profile corresponding to a particular target accent in *Feinauer* as an operable method of modifying the source voice's timbre to match the timbre of the target, modifying the resulting accent. EX1004-Feinauer, [0081] (“At step 504, the accent and dialect modifier pre-loads relevant translation models and translation dictionaries...the accent and dialect modifier identifies relevant translation models and dictionaries based on the content and/or language, accent and/or dialect the accent and dialect modifier is to use or operate[,]...[and] the accent and dialect modifier loads relevant translation models and dictionaries prior to, at the beginning of a conversation or during the conversation”), [0105] (“...[T]he accent and dialect modifier loads translation models and dictionaries. Loading translation models and dictionaries 704 can include preparing relative translation models and dictionaries that are to be accessed and used for the appropriate target dialect and/or accent, such as the dialect and/or accent to be selected for the context.”), 81, 105, Fig, 7; EX1006-Fan, [0010] (“a

similarity between styles of the to-be-converted voice and the third voice...is adjusted by using the weight that corresponds to each sub-feature and that is input by the user”); [0081]-[0082] (“the terminal device selects, based on the sub-template selection operation, a target voice corresponding to a sub-template...”); [0094] (“[t]he terminal multiplies the weight that is input by the user and that corresponds to each sub-feature and the vector of each sub-feature...[a] result obtained through the multiplication and the vector (for example, a timbre vector) of the first feature are simultaneously input into a multihead attention structure for attention alignment, so that the multihead attention structure outputs a style vector, where the style vector is the style feature of the third voice”); [0126]-[0129](“ Refer to FIG. 9 . When the mode selection operation is for selecting the dialect switching mode, the terminal inputs a PPG feature into a dialect switching model (a seq2seq neural network) based on the mode selection operation, and outputs a melt feature form of the second voice by using the dialect switching model. Then, a melt feature is converted into a playable voice form by using a neural network vocoder. The first voice is a voice of a first dialect, and the second voice is a voice of a second dialect...Then, the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation. Content of the second voice is the same as content of the input

voice (the first voice), and a speaking style of the input voice (the first voice) is retained.”);[0160]-[0162] (“In an optional implementation, the obtaining module 1330 is further configured to: receive a template selection operation input by the user[,...] obtain a voice corresponding to the target template, and use the voice corresponding to the target template as the third voice. [T]he obtaining module 1330 is further configured to receive the third voice input by a second speaker, where the first voice is a voice of a first speaker, and the second speaker is any person different from the first speaker. [T]he target conversion mode is the dialect conversion mode, and the target voice conversion network is a dialect conversion network. The processing module 1310 is further configured to input the feature information of the first voice into the dialect conversion network, and output the second voice over the dialect conversion network, where the first voice is a voice of a first dialect, and the second voice is a voice of a second dialect.”), Figs. 9, 12-13.

118. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan*’s modules described functionality in *Fan*’s style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. I note the description of a style feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male

announcers). However, the description of the style enhancement model in regard to a style feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable all of *Fan*’s envisioned embodiments.

119. As discussed, *Feinauer* itself teaches extracting linguistic and acoustic features from a user’s input speech. *See*, Sections IX.3-IX.4. *Feinauer* further teaches modifying an input, unmodulated audio signal to an output, modulated audio signal by adjusting the accent/dialect of the voice of a user (such as modifying the voice from an American accent to an Australian accent). EX1004-Feinauer, [0046]. *Fan* further provides detail as to how a PHOSITA would have implemented the modification, teaching “the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation.” EX1006-Fan, [0127].

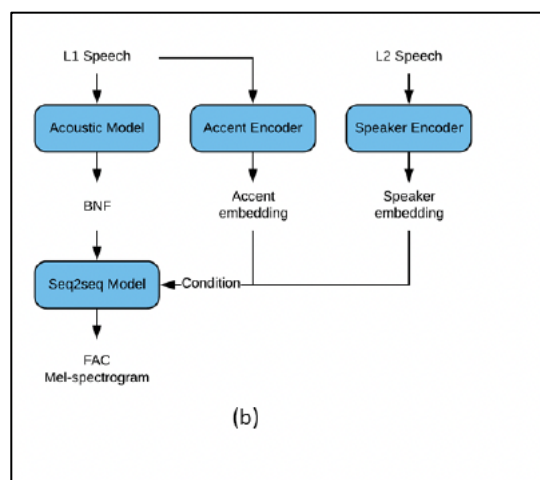
120. A PHOSITA would have combined the teachings such that the method of *Feinauer*, in which accent modification is performed as detailed by *Fan* regarding Fig. 9, in which linguistic content (*e.g.*, comprising PPGs) is combined with acoustic features (*e.g.*, pitch, tone, and/or volume per Claim 1(b)) and a speaker embedding (style embedding vector) to modify the accent (dialect) of the output audio signal. A

PHOSITA would have recognized the modification constitutes combining prior art elements (*Feinauer's* translation model and *Fan's* style embedding vector target template) according to known data storage and retrieval methods with predictable results, further motivating the combination. *Fan* provides several examples of how to store and retrieve the style embedding vector with *Fan's* accent conversion system. See EX1006-Fan, [0024]-[0026] (“According to a third aspect, an embodiment of this application provides a terminal, including a processor, where the processor is coupled to at least one memory, and the processor is configured to read a computer program stored in the at least one memory, so that the terminal performs the method in any implementation of the first aspect. According to a fourth aspect, an embodiment of this application provides a computer-readable medium. The computer-readable storage medium stores a computer program, and when the computer program is run on a computer, the computer is enabled to perform the method in the first aspect. According to a fifth aspect, this application provides a chip system. The chip system includes a processor, configured to support a terminal device in implementing functions in the foregoing aspects. In a possible design, the chip system further includes a memory. The memory is configured to store program instructions and data that are necessary for the terminal device. The chip system may include a chip, or may include a chip and another discrete component.”); [0094] (“For example, a structure output by the multihead attention mechanism may be a

256-dimensional style embedding vector (a style feature).”); [0156] (“The processing module **1310** is further configured to: input the third voice into the style separation model, and extract a vector of the first feature of the third voice by using the style separation model; and input the third voice into a sub-feature extraction model, and extract a vector of the sub-feature by using the sub-feature extraction model.”); [0158]-[0159] (“The processing module 1310 is further configured to determine the style feature of the third voice based on the vector of the first feature, the vector of the sub-feature received by the input module 1320, and the weight of the sub-feature received by the input module 1320. In an optional implementation, the processing module 1310 is further configured to input the vector of the first feature into a multihead attention structure, input the vector of the sub-feature and a product of the vector of the sub-feature and the weight corresponding to the sub-feature into the multihead attention structure, and output the style feature of the third voice by using the multihead attention structure.”). Because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan’s* software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification.

121. This would have been well within the skillset of PHOSITA, who would have been familiar with incorporating speaker embeddings to a sample of input speech, as seen in prior art such as *Accentron*. A PHOSITA would have done any

modifications necessary to make this modification operate since this combination was well within a PHOSITA's skillset. A PHOSITA would have understood receiving a speaker embedding from a chunk of audio in accent conversion was common practice in the prior art. For example, *Accentron* teaches extracting a "speaker embedding" from a segment of audio. EX1010-Accentron, 11 ("we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65])").



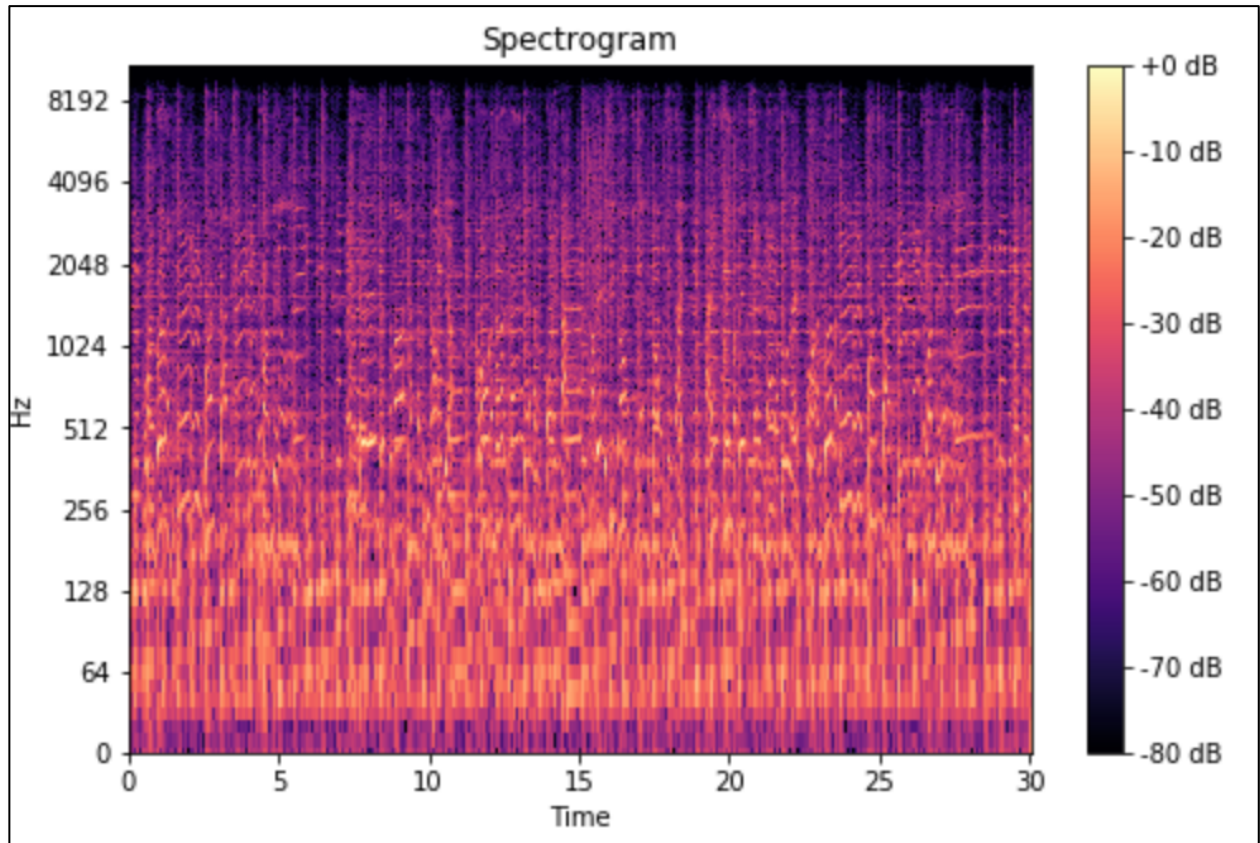
EX1010-Accentron, Fig. 1 (excerpt). *Accentron* teaches "foreign accent conversion." EX1010-Accentron, 1.

"synthesizing a spectrum representation"

122. *Feinauer* teaches "prosodic vocal features and characteristics of the dialects" extracted from the audio samples are converted into a "a spectrogram."

EX1004-Feinauer, 95, 99, 64. *Feinauer* already teaches generating a spectrogram

representing a signal's frequency spectrum. It was well known in the prior art that a spectrogram is a representation of a signal's frequency spectrum. For example, *Fan's* “mel spectrum feature is an intermediate representation. Although a mel spectrum contains information about an output voice, the mel spectrum cannot be directly played, and an inverse process of converting a voice signal into the mel spectrum needs to be performed. In this application, a playable audio file may be generated by using a vocoder by using an acoustic feature of the mel spectrum. The vocoder may be a neural network vocoder, and the network is responsible for converting the mel feature into a voice signal with a high natural degree.” EX1006-Fan, [0099]. As another example, *Voice Communication with Computers* to Schmandt (“*Schmandt*”) teaches a “spectrogram is a transformation of the waveform into the frequency domain.” EX1021-Schmandt, 13. In a spectrogram, the “horizontal axis represents time, the vertical axis represents frequency, and the intensity or blackness at a point indicates the acoustic energy at that frequency or time.” EX1021-Schmandt, 13. A spectrogram “requires much less storage than the original waveform and reveals acoustic features specific to speech.” EX1021-Schmandt, 13.



EX1034-Roberts, 1. *Fan* provides additional detail, teaching producing such a spectrogram specifically of the acoustic features, linguistic features, and style (speaker) embedding vector. EX1006-Fan, [0053], [0097]-[0099], [0126]-[0127].

123. A PHOSITA would have appreciated that in the *Feinauer-Prabhavalkar-Fan* modified system, the mapped acoustic features, linguistic features with a reduced accent, and speaker embedding for a human speaker are used to synthesize a mel spectrum representation, as in *Fan*. See ¶116. Specifically, as discussed above regarding Fig. 9 and its corresponding embodiment, a linguistic content (*e.g.*, PPG features) of the first (input) voice and the acoustic content (style features) of the first (input) voice are input into a dialect conversion model, and the

second (modified) voice is output by using the dialect (accent) conversion model corresponding to a style embedding vector (speaker embedding). EX1006-Fan, [0127], Fig. 9.

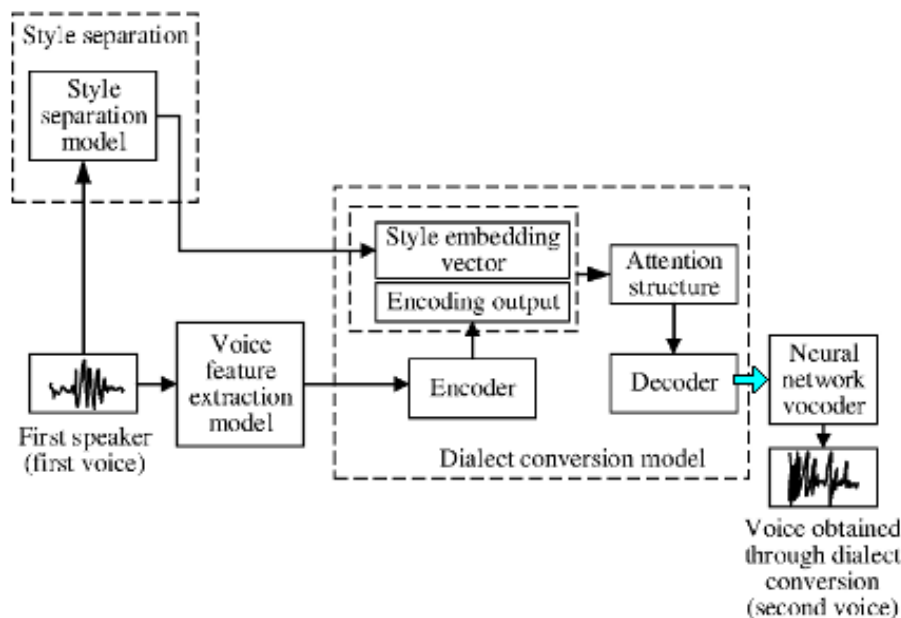


FIG. 9

EX1006-Fan, Fig. 9 (annotated) *see also* Figs. 7, 12. The output of the dialect conversion model, annotated with a blue arrow above, is contemplated by *Fan* as comprising a mel **spectrum** feature. EX1006-Fan, [0099], Fig. 9. *Fan's* “mel spectrum feature is an intermediate representation. Although a mel spectrum contains information about an output voice, the mel spectrum cannot be directly played, and an inverse process of converting a voice signal into the mel spectrum needs to be performed. In this application, a playable audio file may be generated by using a vocoder by using an acoustic feature of the mel spectrum. The vocoder may

be a neural network vocoder, and the network is responsible for converting the mel feature into a voice signal with a high natural degree.” EX1006-Fan, [0099]. Specifically, *Fan* explains that a mel spectrum feature is an intermediate representation that “contains information about an output voice” to be input to a vocoder, precisely as shown in Fig. 9. EX1006-Fan, [0099], Fig. 9. See ¶¶48-50.

124. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan*’s modules described functionality in *Fan*’s style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. I note the description of a style feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. See, e.g. EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model in regard to a style feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable all of *Fan*’s envisioned embodiments. Thus, because a mel spectrum teaches a *spectrum representation*, and because the *Feinauer-Prabhavalkar-Fan* system as combined would generate such mel spectrum representations as output, the modified system teaches *synthesizing a spectrum*

representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker.

6. Claim 1(c): “generating, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal.”

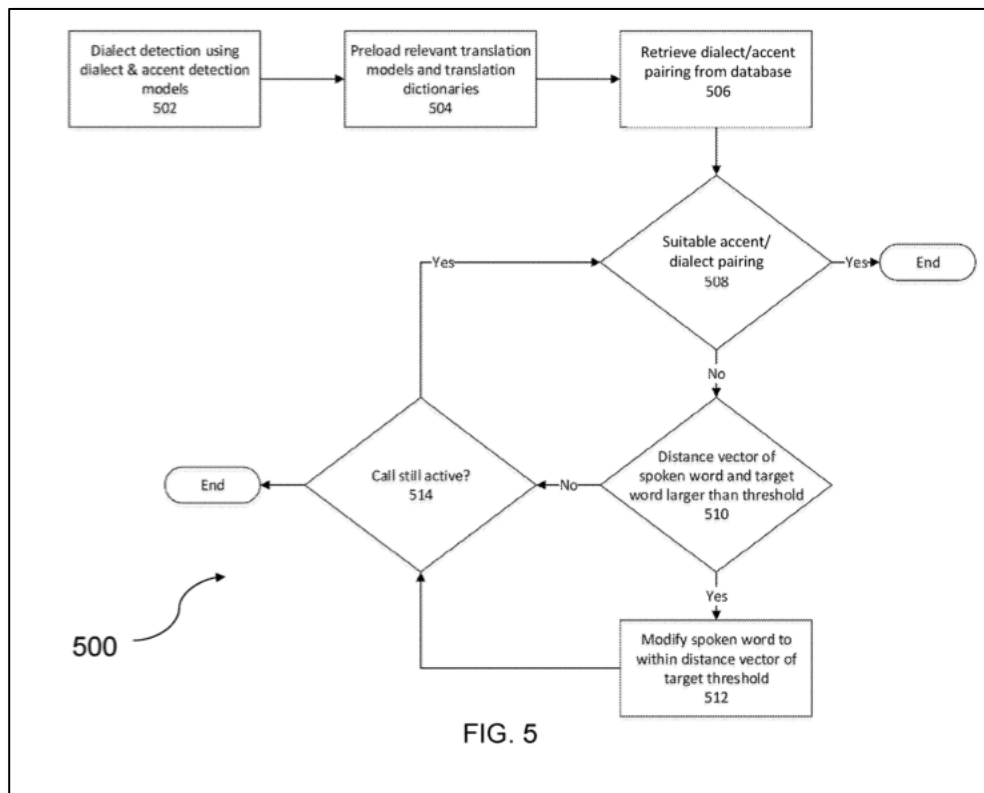
125. The ’561 Patent describes output chunks, like input chunks, comprising a plurality of frames. EX1001, 8:24-40, 9:10-14. Specifically, the ’561 Patent describes output chunks as corresponding in size to that of the input chunk from which they were generated. EX1001, 14:13-18.

126. *Feinauer* teaches combining a target accent *speaker embedding*, *acoustic features* such as pitch, and *linguistic features* including spoken words into an output audio signal having a modified accent. EX1004-Feinauer, [0057]. In the *Feinauer-Prabhavalkar-Fan* method, the output audio signal is generated in *output chunks* based on the spectrum representation by a *vocoder*. EX1005-Prabhavalkar, 18:1-18. In the *Feinauer-Prabhavalkar-Fan* method, the output audio signal is generated in *output chunks* comprising a plurality of frames corresponding to the input chunk from which it was generated. EX1005-Prabhavalkar, 18:1-18.

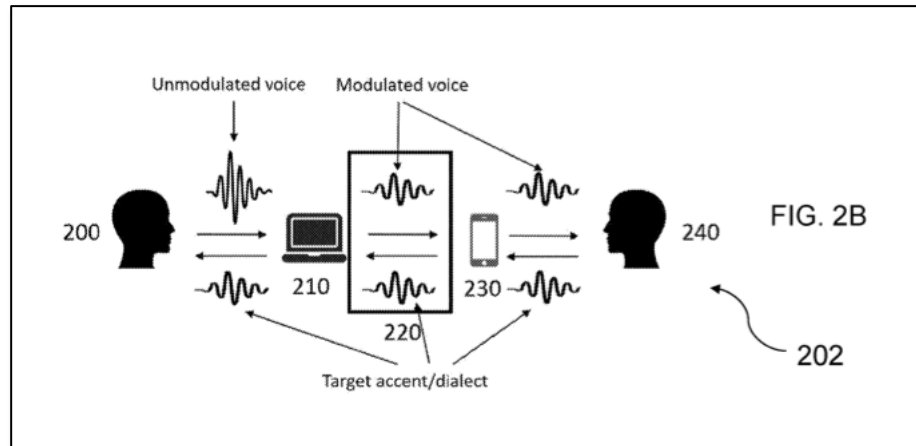
a. Feinauer’s Teachings

127. *Feinauer* employs an accent modifier 318 to modify the accent in “portions” of a user’s speech, modified per Claim 1(a)(1) to be *Prabhavalkar*’s input chunks. EX1004-Feinauer, 57, Fig. 3; *See* Section IX discussing “chunk-based”

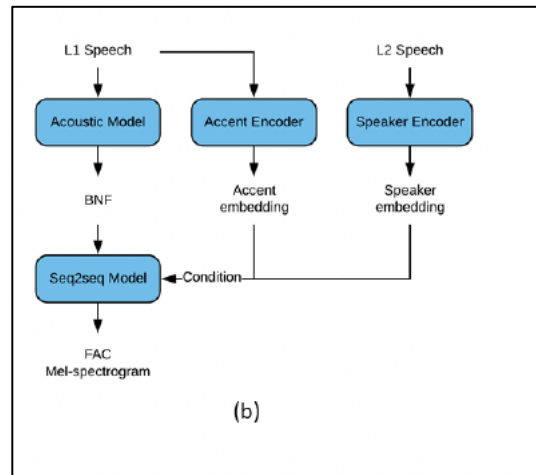
processing. Specifically, *Feinauer's* method generates a conversation output audio signal with a modified accent. *Feinauer* teaches “the conversation can be broadcast on a lag of which the duration must be long enough for a word to be expressed and analyzed for its accent and dialect strength and then modified to be expressed by the modification software. For example, the lag duration can be a duration that sounds like real-time to a user (e.g. 100ms).” EX1004-Feinauer, [0086].



EX1004-Feinauer, Fig. 5. A PHOSITA would have understood that *Feinauer's* broadcast” conversation teaches an output audio signal as is consistent with Fig. 2B.



EX1004-Feinauer, Fig. 2B. Each of steps 512 (Fig. 5), 714 (Fig. 7), and 812 (Fig. 8) provide steps in which *Feinauer's* method generates an output audio signal based on a *speaker embedding*, (*Feinauer's* template, modified to include *Fan's* style embedding vector per Claim 1(b)), *acoustic features* (*Feinauer's* pitch, tone, and/or volume per Claim 1(a)(2)), and *linguistic features* (*Feinauer's* individual words, modified to be represented by *Fan's* PPGs per Claim 1(a)(3)). EX1004-Feinauer, 86, 110, 117, Figs. 5-7. In the combined system, the spectrum representation as discussed in limitation 1(b) is input to a vocoder, as seen in *Fan*. A PHOSITA would have understood this was common practice in the prior art. For example, *Accentron* teaches “foreign accent conversion” EX1010-Accentron, 1.



EX1010-Accentron, Fig. 1 (excerpt). *Accentron's* seq2seq model consumes bottleneck features, accent embedding, and speaker embedding “to synthesize Mel-spectrograms for an arbitrary L2 speaker.” EX1010-Accentron, 12; 10; Table 1. The decoder “produces an 80-dimensional Mel-spectrogram as the prediction of the L2 speech” with a “25ms window and 10ms shift.” EX1010-Accentron, 14, 18. After this, “the output Mel-spectrogram is converted back into a waveform through...a separately trained vocoder (e.g., WaveNet [37], WaveRNN [38]).” EX1010-Accentron, 3. *Accentron's* “Mel-spectrogram is converted back to a waveform” using a WaveRNN vocoder. EX1010-Accentron, 9, 18.

128. A PHOSITA would have appreciated that *Feinauer's* method generates an output audio signal also based on a spectrum representation, for example *Feinauer's* method uses the “spectrogram” to “determine a correlation between the prosodic vocal features of the audio samples and the input dialects.” EX1004-Feinauer, [0095]. This spectrogram correlation is used in the process of generating

an output audio signal when determining “how much to modify the individually spoken word.” EX1004-Feinauer, [0096], Fig. 7, Fig. 8 (808-812). *Feinauer* teaches providing a modified output, envisioning its system may “modif[y] the voice of user 200 into an accent/dialect more familiar to call participant 240 (such as modifying the voice from an American accent to an Australian accent).” EX1004-Feinauer, [0046]. *Feinauer* teaches to “modify the unmodulated voice of one or more of the participants to provide a modulated voice with a target accent and/or dialect.” EX1004-Feinauer, [0045]. *Feinauer*’s modified voice signal with “a different accent and/or dialect” is “transmitted via the communication network 220” and “received by communication device 230 which can then be communicated to, provided to or played to the participant 240.” EX1004-Feinauer, [0044]-[0045]. *Fan* further provides detail, stating, “then, the PPG feature of the first voice and the timbre feature vector of the first voice are input into the dialect conversion model corresponding to the selection operation, and the second voice is output by using the dialect conversion model corresponding to the selection operation.” EX1006-Fan, [0127].

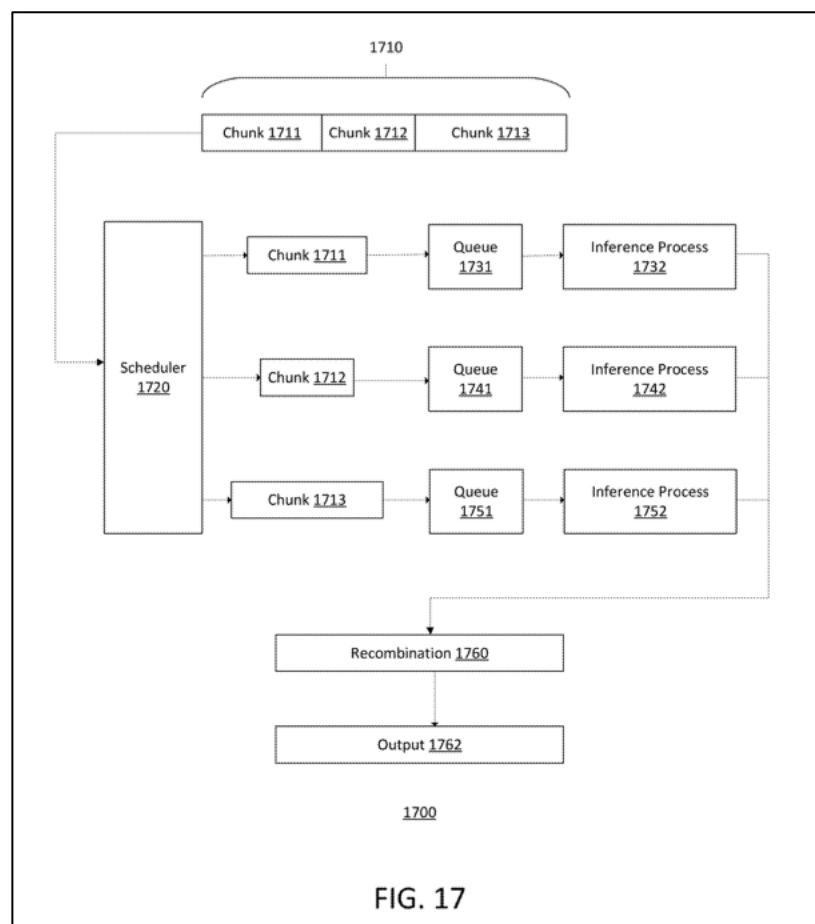
“generating an output chunk by a vocoder”

129. *Prabhavalkar* teaches a method of real-time speech processing in which input is processed as chunks, and an output corresponding to each chunk is produced sequentially before processing the next input chunk. EX1005-

Prabhavalkar, 6:24-28, 18:1-41. *Prabhavalkar* teaches its “chunk-based” processing technique consumes all frames in a first input chunk, processing them and outputting results “before it consumes the next chunk.” EX1005-*Prabhavalkar*, 18:20-23, 17:33-51. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-*Prabhavalkar*, 17:46-51. (emphasis added) Thus, *Prabhavalkar*’s generates output from a given chunk (an “output chunk”) before moving on to processing succeeding chunks. Based on this disclosure, a PHOSITA would have recognized or at least found obvious that *Prabhavalkar*’s “chunk-based approach,” which consumes input as chunks, correspondingly generates output as chunks. *Prabhavalkar*’s “chunk-based approach,” which consumes input as chunks, would apply the same sequential processing when applied to *Feinauer*’s “portions” of user’s speech to produce output chunks of processed audio frames as an “output chunk.” This was well known to do.

130. As an example, U.S. Patent Appl. No. 2020/0294492 to Sypniewski et al. (“*Sypniewski*”) teaches a front end module 201 splits raw audio data into a series of frames. EX1020-Sypniewski, [0059], Fig. 2. *Sypniewski* provides an example in which a 7-10 second long portion of input audio “may be segmented on relative

silence such that the input audio comprises an entire utterance.” EX1020-Sypniewski, [0057], [0110]. *Sypniewski* characterizes these 7-10 second portions of audio as “chunks.” EX1020-Sypniewski, [0162]. Fig. 17 below illustrates an example of audio 1710 split into chunks 1711-1713 and then that each chunk is later recombined after the inference process. EX1020-Sypniewski, [0019] (“After inference, the resulting outputs may be recombined.”)



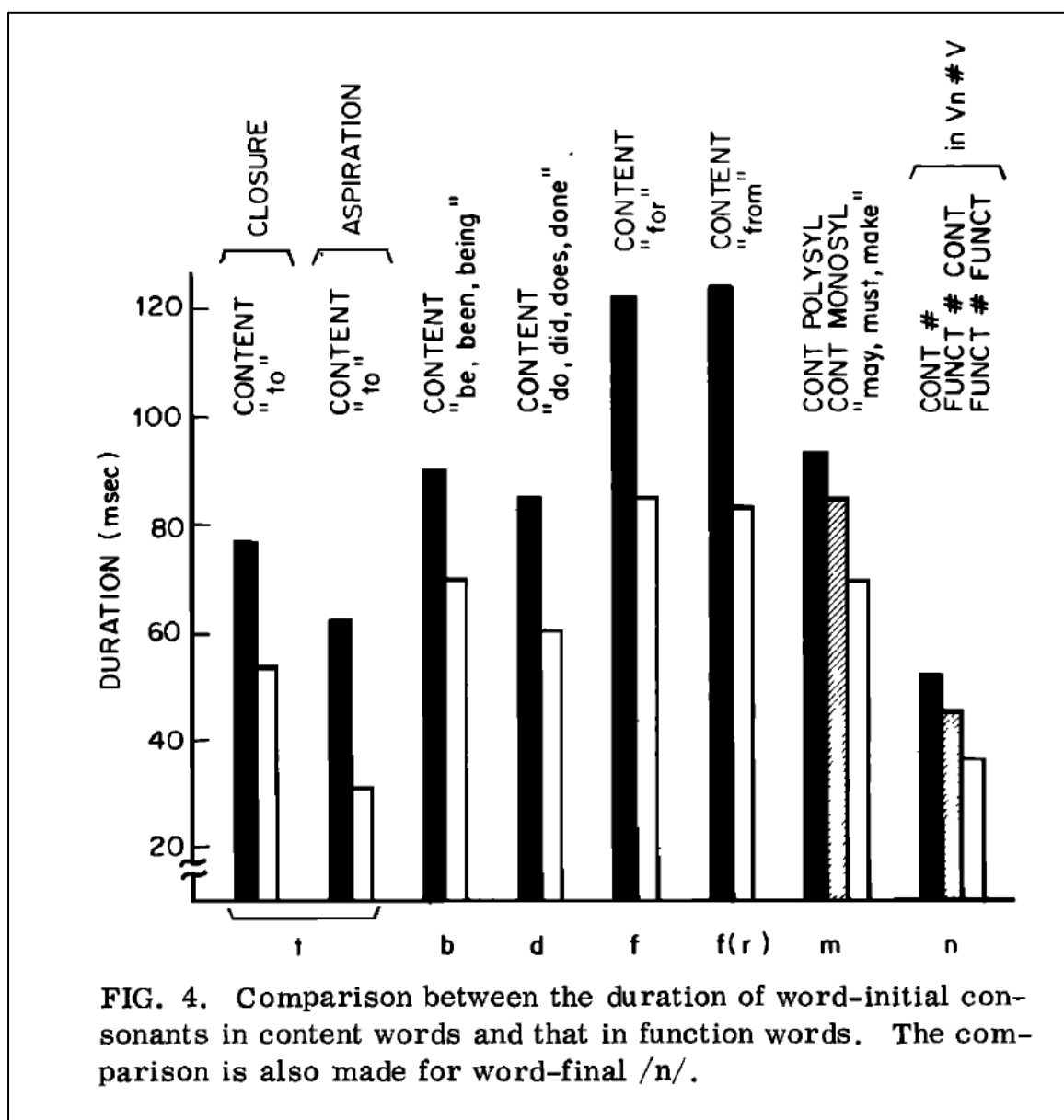
EX1020-Sypniewski, Fig. 17.

131. A PHOSITA would have been motivated to configure the *Feinauer-Prabhavalkar-Fan* system to generate its output voice modified with the target

accent in “output chunks” as taught by *Prabhavalkar* to enable the increase in speed and performance intended when dividing the speech signal into input chunks as discussed regarding Claim 1(a)(1). EX1005-*Prabhavalkar*, 17:33-51. *Prabhavalkar* characterizes its approach as “streaming,” wherein speech is processed “while audio data of an utterance is still being received and without waiting for the utterance to be completed.” EX1005-*Prabhavalkar*, 5:63-6:15, 2:10-17. *Prabhavalkar* teaches that “streaming can be considered to mean that the system has a maximum allowable delay of 300s.” EX1005-*Prabhavalkar*, 18:32-35. A PHOSITA would have appreciated that one example of “streaming” occurs with an “real-time “audio and/or video conversation” as taught by *Feinauer*. EX1004-*Feinauer*, [0041]; [0086]. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-*Prabhavalkar*, 17:46-51. (emphasis added) Therefore, a PHOSITA would have thus recognized *Prabhavalkar*’s system as performing a *real-time* speech processing method, like *Feinauer*. It was well known in the prior art that in order to perform accent conversion in real time, a conversion should be processed in “portions.” As an example, U.S. Patent Appl. No. 2013/0246064 to Wasserblat et al. (“*Wasserblat*”)

teaches “a digital representation of a portion of an interaction may be analyzed in real-time” to maintain “low real time delay” EX1018-Wasserblat, Abstract, [0025].

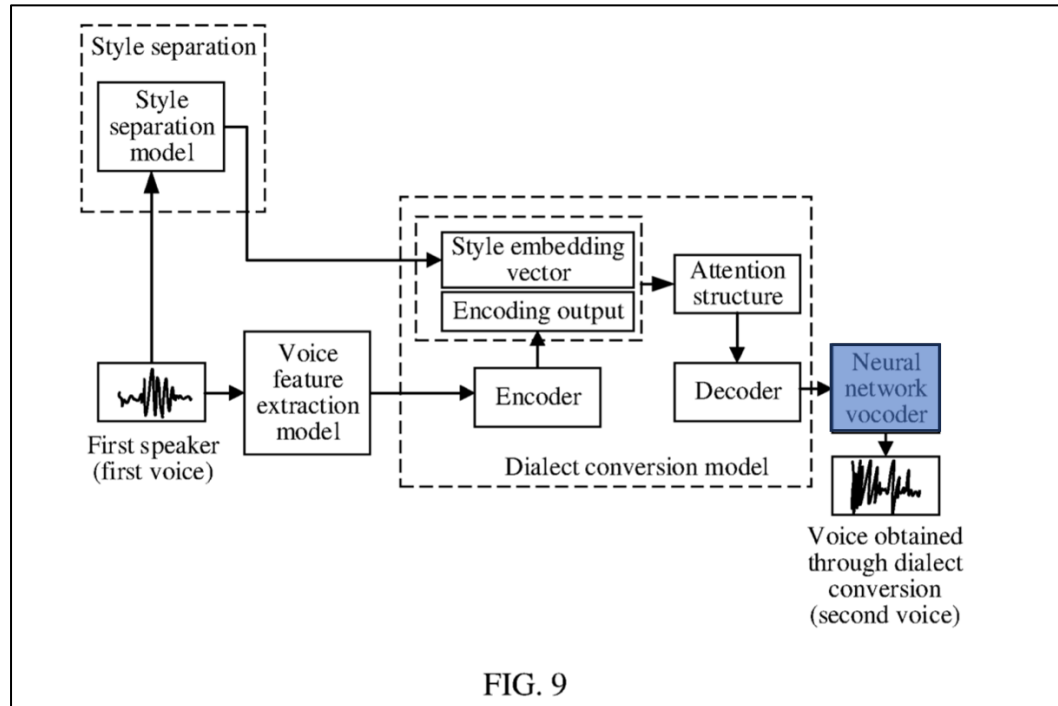
132. Further, it was well known that audio frames are short (~30ms) as taught by *Prabhavalkar* and are commonly used as input to signal processing algorithms (e.g., spectra extraction), while speech events such as phonemes or words are longer in duration than a single frame EX1005-Prabhavalkar, 15:16-22. For example, Umeda teaches many phonemes span multiple 30ms frames as shown in Figure 4.



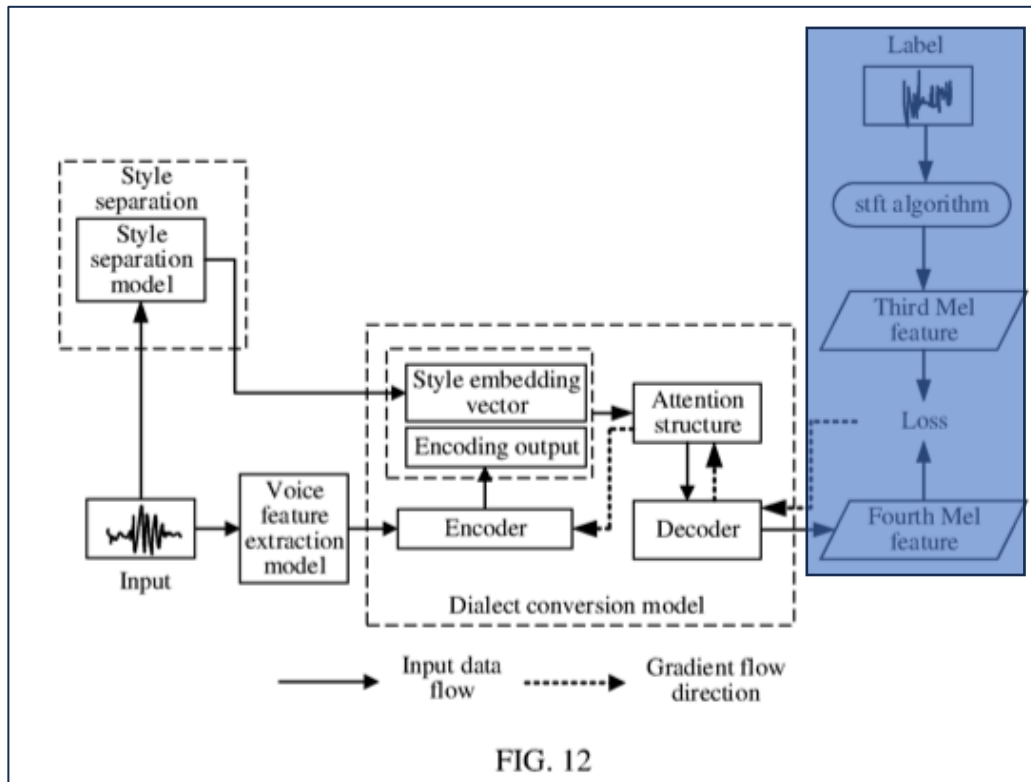
EX1035-Umeda, Fig. 4. Thus, to analyze speech looking for linguistic events, a PHOSITA would have found it advantageous to analyze a longer series of frames, together in sequence as an output chunk.

b. Fan's Teachings

133. *Fan* teaches using a “neural network vocoder” to “convert an output feature of a voice processing neural network into a highly natural voice.” EX1006-Fan, [0049].



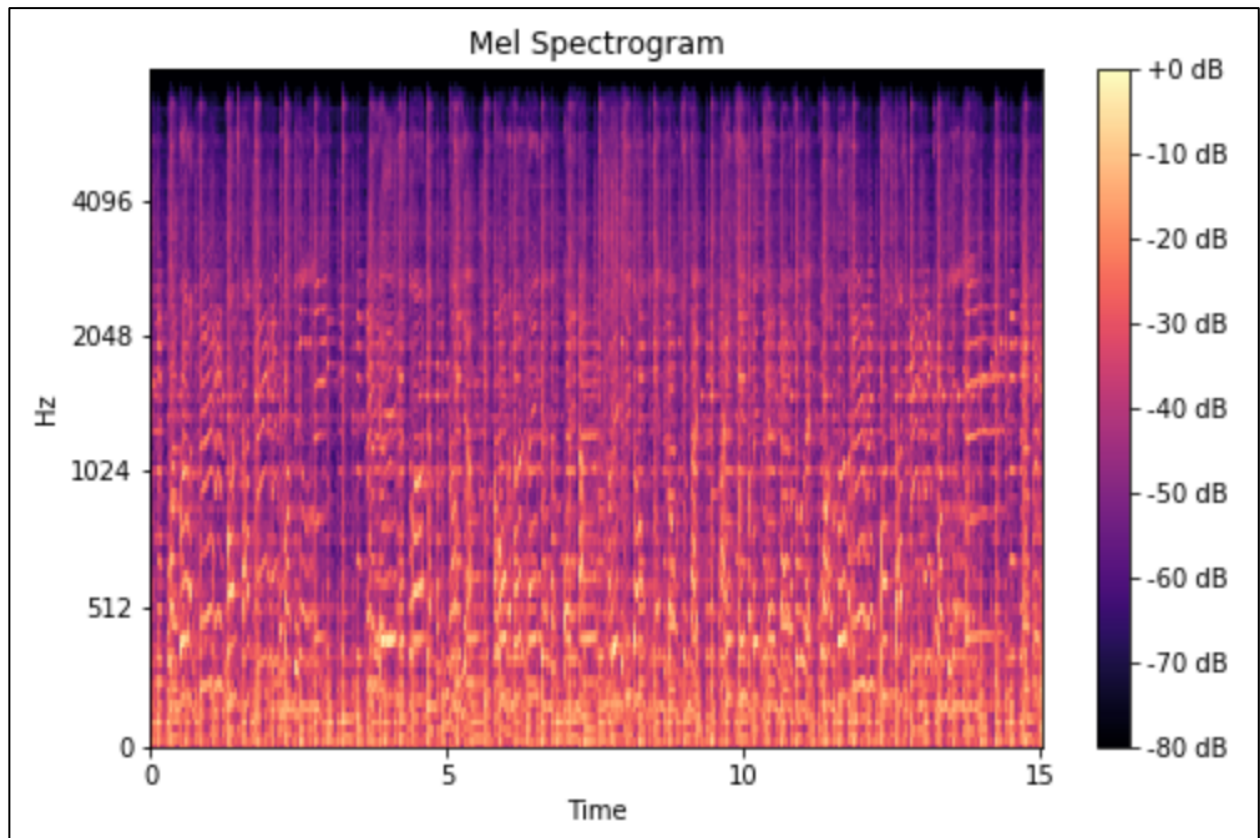
EX1006-Fan, Fig. 9 (annotated).



EX1006-Fan, Fig. 12 (annotated).

134. Fan teaches a mel “feature is converted into a playable voice form by using a neural network vocoder.” EX1006-Fan, [0126]. Fan discusses the third and fourth “melt” features. A PHOSITA would have recognized that “melt” is a typographic error throughout Fan’s specification. Fan describes mel features as “a voice signal being stored as a waveform diagram feature is represented as feature transformation in time domain of the voice signal.” EX1006-Fan, [0053]. In some paragraphs Fan’s specification properly uses the term mel feature. *See* EX1006-Fan, [0053], [0097], [0099]. In other paragraphs the specification uses the term “melt” feature. *See* EX1006-Fan, [0116], [0125]. As a melt feature is non-sensical in plain meaning and was not a term in the art, a PHOSITA would have understood “melt

feature” to be properly read as “mel feature” throughout the specification. A PHOSITA would have understood a mel feature a represents a frequency as measured in a scale which reflects the human auditory system. Studies “have shown that humans do not perceive frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies.” EX1034-Roberts, 11. In order to combat this, in “1937, Stevens, Volkman, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the mel scale.” EX1034-Roberts, 12. From the spectrogram in the prior paragraph, you “perform a mathematical operation on frequencies to convert them to the mel scale.” EX1034-Roberts, 12. A Mel-spectrogram is “is a spectrogram where the frequencies are converted to the mel scale.” EX1034-Roberts, 13.



EX1034-Roberts, 8.

c. Motivation to Combine *Fan's* vocoder

135. In the proposed modified method, A PHOSITA would have been motivated and found it obvious to utilize a hardware or software module such as a vocoder to generate an output chunk based on *Feinauer's* spectrogram-spectrum representation as taught by *Fan*. A PHOSITA would have appreciated the lengthy detail provided by *Fan* regarding hardware, software, and neural network configurations capable of generating an output signal based on a spectrum representation using a vocoder, leading to a REOS. EX1006-Fan, [0049] ("The

neural network vocoder is essentially a neural network, and is configured to convert an output feature of a voice processing neural network into a highly natural voice.”), [0099] (“The vocoder may be a neural network vocoder, and the network is responsible for converting the mel feature into a voice signal with a high natural degree.”), [0126] (“Then, a melt feature is converted into a playable voice form by using a neural network vocoder.”), Figs. 9 and 12. A vocoder was common to use in accent conversion systems to convert a computed representation of speech, such as a spectrogram or LPC coefficients, into an audible speech waveform. *Converting Foreign Accent Speech Without a Reference* to Zhao et al. (“Zhao”) teaches that to obtain hearable synthesized audio, mel spectrograms are transformed into waveforms using a real-time vocoder. EX1031-Zhao, 6rhc (“The predicted Mel-spectrograms are converted back to audio waveforms using a WaveGlow neural vocoder trained on the L2 utterances[.]”); EX1031-Zhao, 2rhc. (“we use a high-quality WaveGlow [15] real-time neural vocoder to convert Mel-spectrograms to speech waveform.”). Another type of vocoder is a “WaveNet vocoder [21] to generate the target speech waveforms.” EX1023-Sisman, 9 lhc. A PHOSITA would have recognized the modification constitutes applying a known technique (*Feinauer* and *Fan*’s steps of producing an output signal using a vocoder) to a known device (*Fan*’s vocoder software and any necessary supporting hardware) with predictable results, further motivating the combination and because the combination required

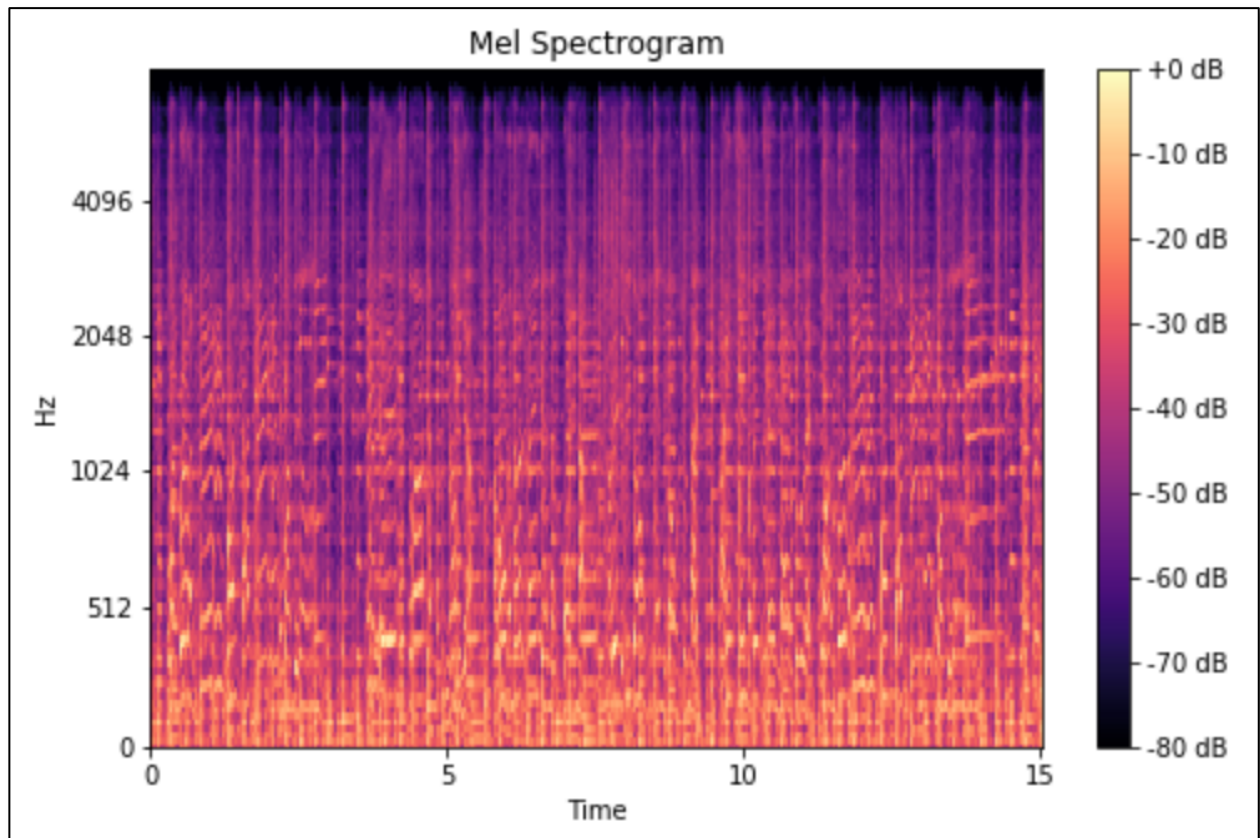
only reprogramming of *Feinauer-Prabhavalkar-Fan*'s software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification. It was common in the prior art to have spectrograms as input to a vocoder. For example, *Foreign Accent Speech Without a Reference* to Zhao et al. ("*Zhao*") teaches that to obtain hearable synthesized audio, mel spectrograms are transformed into waveforms using a real-time vocoder. EX1031-Zhao, 6rhc ("The predicted Mel-spectrograms are converted back to audio waveforms using a WaveGlow neural vocoder trained on the L2 utterances[.]"); EX1031-Zhao, 2rhc. ("we use a high-quality WaveGlow [15] real-time neural vocoder to convert Mel-spectrograms to speech waveform.").

B. Claim 2

- 1. Claim 2: "The method of claim 1, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal."**

136. As discussed in Claim 1(b), *Feinauer*'s spectrogram in *Feinauer-Prabhavalkar-Fan*, includes the linguistic features with a reduced accent. *Feinauer* teaches the acoustic features include a pitch of the input audio signal and Mel

Frequency Cepstral Coefficients (MFCC) of the input audio signal. EX1004-Feinauer, 57, 99. *Feinauer* teaches to “convert the audio sample waveforms into a spectrogram of MFCC [mel-frequency cepstral coefficients].” EX1004-Feinauer, [0099], [0095]. A PHOSITA would have recognized descriptions related to the spectrogram of MFCC in *Fan* are applicable all of *Fan*’s envisioned embodiments. A PHOSITA would have understood *Feinauer*’s spectrogram of MFCC (mel frequency cepstral coefficients) teaches a mel-spectrogram. A PHOSITA would have understood that *Feinauer*’s teachings of a spectrogram of mel frequency cepstral coefficients would be in the mel scale already. For example, a Mel-spectrogram is “is a spectrogram where the frequencies are converted to the mel scale.” EX1034-Roberts, 13.



EX1034-Roberts, 8.

C. Claim 6

1. Claim 6 (Pre): “A computing system, comprising an input device, a vocoder, memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”

137. As discussed in Claim 1(Pre), *Feinauer* teaches a computing device. As discussed in Claim 1(c), *Feinauer-Prabhavalkar-Fan* teach a vocoder. Further, *Feinauer*’s computing device includes an input device such as a microphone where speech is received and detected. EX1004-*Feinauer*, [0031], [0041]. *Feinauer* teaches a computing system device that includes a memory storing instructions that has “one

or more processors, coupled to memory” that are configured to execute the stored instructions. EX1004-Feinauer, [0010], [0012], [0031]-[0032]. *Feinauer’s* processors are configured to execute the stored instructions. EX1004-Feinauer, [0032]. Thus, *Feinauer* also teaches the claimed *input device*.

2. Claim 6 (a): “extract, from a chunk of a stream of chunks of an input audio signal obtained via the input device, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”

138. As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent. As discussed in Claim 6[Pre], this input audio is obtained via *Feinauer’s* input device.

3. Claim 6 (b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”

139. See Claim 1(b).

4. Claim 6 (c): “and generate, via the vocoder and based on the spectrum representation, an output chunk of an output audio signal.”

140. See Claim 1(c).

D. Claim 7

1. Claim 7: “The computing system for real-time correction of claim 6, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio

signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”

141. See Claim 2.

E. Claim 11

1. Claim 11(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”

142. As discussed in Claim 1(Pre), *Feinauer* teaches a computing device that includes a non-transitory computer-readable storage medium including instructions. EX1004-Feinauer, [0031]-[0032], [0043], [0081], [0118]-[0121]; See also Section IX.

2. Claim 11(a): “extract, from a chunk of a stream of chunks of an input audio signal, linguistic features and acoustic features, wherein the linguistic features have a reduced accent;”

143. As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent.

3. Claim 11(b): “synthesize a spectrum representation based on the acoustic features, the linguistic features, and a speaker embedding for a human speaker;”

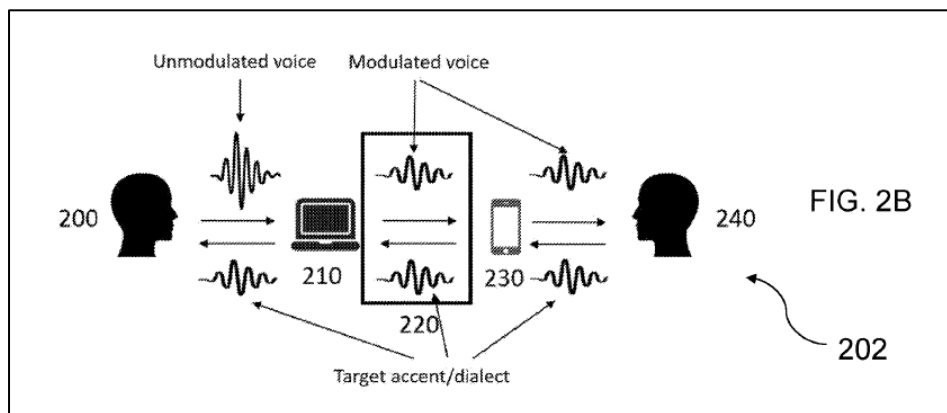
144. See Claim 1(b).

4. Claim 11(c): “generate, by a vocoder and based on the spectrum representation, an output chunk of an output audio signal;”

145. See Claim 1(c).

5. Claim 11(d): “and provide the output audio signal for output via an auditory channel of an output device.”

146. *Feinauer* teaches to provide the output audio signal for output via speakers (auditory channel) of an output device. EX1004-Feinauer, [0014], [0016], [0018], [0041]- [0042]. *Feinauer* teaches “user 200 will have his or her voice modified from a first accent and/or dialect to a second accent and/or dialect” and that participants will communicate “with each other over the communication channel using modulated audio signals of one or more of their voices in the accent and/or dialect selected and modified.” EX1004-Feinauer, [0042].



EX1004-Feinauer, Fig. 2B. Once the modified modulated signal is sent via *Feinauer*’s communication channel, a user may listen to the signal via “one or more input/output (I/O) devices 126 (e.g., a mouse, a keyboard, a microphone, one or

more speakers...” EX1004-Feinauer, [0031]. A PHOSITA would have recognized a “speaker” teaches an auditory channel, since it was well known that a speaker allows a user to listen to audio. For example, Fan teaches a “loudspeaker 1561 is configured to output a processed second voice.” EX1006-Fan, [0201].

F. Claim 12

- 1. Claim 12: “The non-transitory computer-readable storage medium of claim 11, wherein: the spectrum representation includes a melspectrogram with the reduced accent; and the acoustic features include one or more of a pitch of the input audio signal, a standard deviation of the pitch of the input audio signal, a mean of the pitch of the input audio signal, an energy of the input audio signal, a noise floor of the input audio signal, Linear Prediction Cepstral Coefficients (LPCC) of the input audio signal, Mel Frequency Cepstral Coefficients (MFCC) of the input audio signal, or Bark Frequency Cepstral coefficients (BFCC) of the input audio signal.”***

147. See Claim 2.

G. Claim 16

- 1. Claim 16 (Pre): “A method for real-time accent correction, the method implemented by a computing system and comprising:”***

148. See Claim 1[Pre].

- 2. Claim 16 (a): “dividing a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”***

149. Feinauer in view of Prabhavalkar teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks

including a 5 predetermined (*pre-defined*) number of frames of the speech audio signal. See Claim 1(a)(1).

3. Claim 16 (b): “extracting, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”

150. As discussed in Claim 1(a), *Feinauer-Prabhavalkar-Fan* teaches extracting, from the input chunk acoustic features and linguistic features, wherein the linguistic features have a reduced accent. *Feinauer-Prabhavalkar-Fan* further teach a context associated with the input chunk.

151. The '561 Patent explains “context may include a pre-determined number of the frames belonging to chunks preceding the input chunk in the stream of input chunks.” EX1001, 15:11-16.

152. As discussed regarding Claim 1(a)(1), in the *Feinauer-Prabhavalkar* method, speech audio is divided into *input chunks* comprising a *pre-defined number of frames*. See Section IC citing EX1005-*Prabhavalkar*, 6:56-7:4, 15:16-22. Further, in the *Feinauer-Prabhavalkar-Fan* method, the style separation model (part of the mapped *acoustic feature module*) uses “attention” to “focus more on different key parts of an input instead of the entire input.” EX1006-Fan, [0048], [0090], [0094]. This was a very common way to focus on different parts of audio input in the prior art. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to

focus more on different key parts of an input instead of the entire input.”). As another example, Zhao teaches forcing its decoder to compute the attention alignments from both the forward and backward directions during training, making the decoder incorporate useful contextual information from both the past and future when producing the alignment. EX1031-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1010-Accentron, 11. (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”) *Prabhavalkar* explains that “attention” may be computed using chunks immediately preceding each input chunk. For example, *Prabhavalkar* teaches “the attender of the speech recognition model is configured to compute attention based speech frames in the current chunk and one or more previous chunks of the utterance.” EX1005-*Prabhavalkar*, 6:56-11, 15:16-22. These preceding chunks (including a pre-determined number of frames) teach *a context associated with the input chunk* in the *Feinauer-Prabhavalkar-Fan* method. These preceding chunks (including a pre-determined number of frames) teach *a context associated with the input chunk* in the *Feinauer-Prabhavalkar-Fan* method.

a. Prabhavalkar’s Teachings

153. In speech recognition, “attention” refers to a mechanism that allows a neural network model to dynamically focus on the most relevant segments of an audio input. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”); [0097] (“Because a length of the input sequence may be long, it is difficult for a vector to express abundant information during decoding. Therefore, the attention mechanism is introduced. Then, a vector of the style feature and a PPG vector output by the encoder are spliced in width, to obtain an attention input matrix. Then, the attention input matrix is sent to the attention structure frame by frame, and the attention structure cooperates with the encoder to output a mel (or may also be referred to as a mel spectrum) feature of the second voice. The mel spectrum feature is a feature representation form of the second voice. After encoding ends, different weighting is performed on an output vector of an encoding module to obtain several vectors, and each vector corresponds to one output. This ensures that an output sequence is focused on an input sequence instead of depending on all input sequences.”). In another example, Zhao teaches forcing its decoder to compute the attention alignments from both the forward and backward directions during training, making the decoder incorporate useful contextual information from both the past and future when producing the alignment. EX1031-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1010-Accentron, 11. (“we

divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”)

154. *Prabhavalkar* teaches computing attention based speech frames in the current chunk and one or more previous chunks of the utterance. EX1005-*Prabhavalkar*, 6:56-7:11. In a specific example, *Prabhavalkar* computes attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:5-12, 18:28-41. Because each chunk, in an embodiment, includes 5 speech frames, the “10 previous chunks” to a particular input chunk would include a pre-determined number (50) of speech frames. EX1005-*Prabhavalkar*, 6:65-7:4 (“each chunk having a first predetermined number of speech frames representing speech occurring before speech content being predicted at the current time step and a second predetermined number of speech frames representing speech occurring after the speech content being predicted in the current time step.”), 22:1-38. *Prabhavalkar* teaches “the attender is configured to compute attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:9-11. Thus, these predetermined number of frames belonging to chunks preceding the input chunk teach a context associated with the input chunk. EX1005-*Prabhavalkar*, 7:1-3, 18:1-41 (“As discussed above with reference to FIG. 1, attention based models consist of

three modules. First, an encoder module, represented by a multi-layer recurrent neural net-work (RNN), models the acoustics. Second, a decoder module, which consists of multiple RNN layers, predicts the output sub-word unit sequence. Third, an attention layer module selects frames in the encoder representation that the decoder should attend to when predicting each sub-word unit. Attention-based models, such as Listen, Attend and Spell (LAS) can be explored in "full-sequence" mode, meaning attention is computed by seeing the entire input sequence. Thus, during inference, the model can produce the first output token only after all input speech frames have been consumed.”), 22:1-3638 EX1001, 1:64-66 (“The context may include a pre-determined number of the frames belonging to chunks preceding the input chunk in the 65 stream of input chunks.”), 6:55-63 (“An input to module 302 may include a chunk 408 of frames concatenated with a context 402. The context 402 may include a pre-determined number of frames of the speech audio signal 202 preceding the chunk 408. Context 402 can be stored in a cache of module 302 and continuously updated. Thus, at each state of real time, input of the modules can include the chunk 408 ended at the previous state of real time and the context 402 corresponding to the chunk.”), 17:1-4 (Claim 8). It was well known that attention is a type of context. EX1010-Accentron, (“the target speech using the *attention context*”); EX1010-Accentron, Table 1.

Table 1: Hyperparameters of the seq2seq foreign accent conversion model.		
Block	Component	Parameters
Inputs	BNF	256-dimensional
	Speaker d-vector	256-dimensional
	Accent d-vector	256-dimensional
Encoder	3 × Conv layers	512 5×1 kernel with 1×1 stride; ReLU; batch norm
	2 × p-Bi-LSTM	256 cells per direction;
Attention	Location-sensitive attention	128-dim attention context; 32 31×1 attention conv kernels
Decoder	PreNet	2 × Fully-connected layer ; 256 neurons; ReLU
	2 × LSTM	1024 cells
	Linear (Mel)	1 × Fully-connect layer; 80 units; no activation
	Linear (stop token)	1 × Fully-connect layer; 1 unit; no activation
	PostNet	5 × Conv layers; 512 5×1 kernel with 1×1 stride; tanh; batch norm
Outputs	Mel-spectrogram	80-dimensional
	Stop token	2-dimensional

b. Motivation to Combine *Prabhavalkar*’s Context

155. A PHOSITA would have been motivated to extract acoustic features using an attention-based approach as taught by *Feinauer-Prabhavalkar-Fan* by considering the context of the frames of the 10 previous chunks as taught by *Prabhavalkar*. EX1005-*Prabhavalkar*, 22:1-38. (“a chunk size of 5 (300 ms latency) is the requirement for allowed streaming delay”) *Prabhavalkar teaches* “the attender is configured to compute attention based on speech frames in the current chunk and at least 10 previous chunks of the utterance.” EX1005-*Prabhavalkar*, 7:9-11. *Prabhavalkar teaches* that “streaming can be considered to mean that the system has a maximum allowable delay of 300s.” EX1005-*Prabhavalkar*, 18:1-41. A PHOSITA would have appreciated that one example of “streaming” occurs with an “real-time

“audio and/or video conversation” as taught by *Feinauer*. EX1004-Feinauer, [0041]; [0086]. *Prabhavalkar*’s model shifts “attention from one chunk to the next to limit the amount of latency incurred. As a result, the attention mechanism can provide outputs to the decoder allowing a transcription of a first chunk or segment of an utterance **while a second chunk or segment of the utterance is still being spoken.**” EX1005-Prabhavalkar, 17:33-51.

156. A PHOSITA would have appreciated *Fan*’s discussion of “attention” providing “focus” on “key parts” of an input as an express motivation to apply *Prabhavalkar*’s “previous chunks” *context* approach to the system’s acoustic feature extraction module, allowing the combined method to focus more on the key *current input chunk*. EX1006-Fan, [0048] (“The purpose is to enable the output of the decoder to focus more on different key parts of an input instead of the entire input.”); [0097] (“Because a length of the input sequence may be long, it is difficult for a vector to express abundant information during decoding. Therefore, the attention mechanism is introduced. Then, a vector of the style feature and a PPG vector output by the encoder are spliced in width, to obtain an attention input matrix. Then, the attention input matrix is sent to the attention structure frame by frame, and the attention structure cooperates with the encoder to output a mel (or may also be referred to as a mel spectrum) feature of the second voice. The mel spectrum feature is a feature representation form of the second voice. After encoding ends, different

weighting is performed on an output vector of an encoding module to obtain several vectors, and each vector corresponds to one output. This ensures that an output sequence is focused on an input sequence instead of depending on all input sequences.”). As another example, Zhao teaches “by forcing the decoder to compute the attention alignments from both the forward and backward directions during training, we can make the decoder incorporate useful contextual information from both the past and future when producing the alignment.” EX1031-Zhao, 2368. As another example of context, *Accentron* teaches a 150-frame overlap. EX1010-Accentron, 11. (“we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).”)

157. A PHOSITA would have recognized the modification constitutes applying a known technique (*Prabhavalkar’s* context provided by frames of previous chunks) to a known device (hardware employing a style separation model, mapped as the *acoustic feature extraction module* of *Feinauer-Prabhavalkar-Fan*) with the predictable result of accounting for the context of previous chunks when extracting acoustic features from a particular input chunk. As one example, Zhao teaches “by forcing the decoder to compute the attention alignments from both the forward and backward directions during training, we can make the decoder

incorporate useful contextual information from both the past and future when producing the alignment.” EX1031-Zhao, 2368. Because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan*’s software using ordinary skill and hardware already present, a PHOSITA would have appreciated a REOS for this modification.

4.Claim 16 (c): “obtaining a speaker embedding for a human speaker; and”

158. *Feinauer-Prabhavalkar-Fan* teaches to obtain a speaker embedding for a human speaker. *See* Claim 1(b).

5.Claim 16 (f): “generating an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”

159. *Feinauer-Prabhavalkar-Fan* teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. *See* Claim 1(c).

H. Claim 17

- 1. Claim 17(Pre): “A computing system, comprising memory comprising instructions stored thereon, and one or more processors coupled to the memory and configured to execute the stored instructions to:”***

160. *See* Claim 6[Pre].

- 2. Claim 17(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks***

including a pre-defined number of frames of the speech audio signal;

161. *Feinauer* in view of *Prabhavalkar* teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a 5 predetermined (pre-defined) number of frames of the speech audio signal. See Claim 1(a)(1).

3. ***Claim 17(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”***

162. See Claim 16(b), 1(a).

4. ***Claim 17(c): “obtain a speaker embedding for a human speaker; and generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”***

163. *Feinauer-Prabhavalkar-Fan* teaches to obtain a speaker embedding for a human speaker. See Claim 1(b). *Feinauer-Prabhavalkar-Fan* teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. See Claim 1(c).

I. Claim 18

1. ***Claim 18(Pre): “A non-transitory computer-readable storage medium comprising instructions stored thereon that, when executed by one or more processors, cause the processors to:”***

164. See Claim 11[Pre].

2. ***Claim 18(a): “divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a pre-defined number of frames of the speech audio signal;”***

165. *Feinauer* in view of *Prabhavalkar* teach to divide a speech audio signal into a stream of input chunks, an input chunk from the stream of input chunks including a 5 predetermined (pre-defined) number of frames of the speech audio signal. See Claim 1(a)(1).

3. ***Claim 18(b): “extract, from the input chunk and a context associated with the input chunk, acoustic features and linguistic features, wherein the linguistic features have a reduced accent;”***

166. See Claim 16(b), 1(a).

4. ***Claim 18(c): “obtain a speaker embedding for a human speaker; and”***

167. *Feinauer-Prabhavalkar-Fan* teaches to obtain a speaker embedding for a human speaker. See Claim 1(b).

5. ***Claim 18(d): “generate an output chunk of an output audio signal based on the speaker embedding, the acoustic features, and the linguistic features.”***

168. *Feinauer-Prabhavalkar-Fan* teach generating an output chunk of an output audio signal based on the speaker embedding, acoustic features, and linguistic features. See Claim 1(c).

X. OPINIONS REGARDING GROUND 2: CLAIMS 3, 8, 13 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND SARGSYAN*

A. Claim 3

1. ***Claim 3: “The method of claim 1, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”***

169. *Feinauer’s* processor is able to process digital and analog signals. EX1004-*Feinauer*, [0032]. A PHOSITA would have expected that *Feinauer’s* system would digitize the input audio signal a sampling rate to make the signal usable in speech processing. *Prabhavalkar* teaches to “receive spoken information from a user and convert it to usable digital information.” EX1005-*Prabhavalkar*, 26:57-59. *Feinauer-Prabhavalkar-Fan* in view of *Sargsyan* teach the *input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate*.

170. It was well known to process audio in a digital format. For example, *Wasserblat* teaches “segments of a digital representation of a voice interaction may be analyzed to produce analysis results.” EX1018-*Wasserblat*, [0104] A PHOSITA would have recognized that applying *Prabhavalkar’s* digitization techniques to *Feinauer’s* system would have only required a reprogramming according to known techniques to produce the reasonable result of being able to process audio quicker in *Feinauer’s* accent conversion method by processing a digital signal, resulting in a

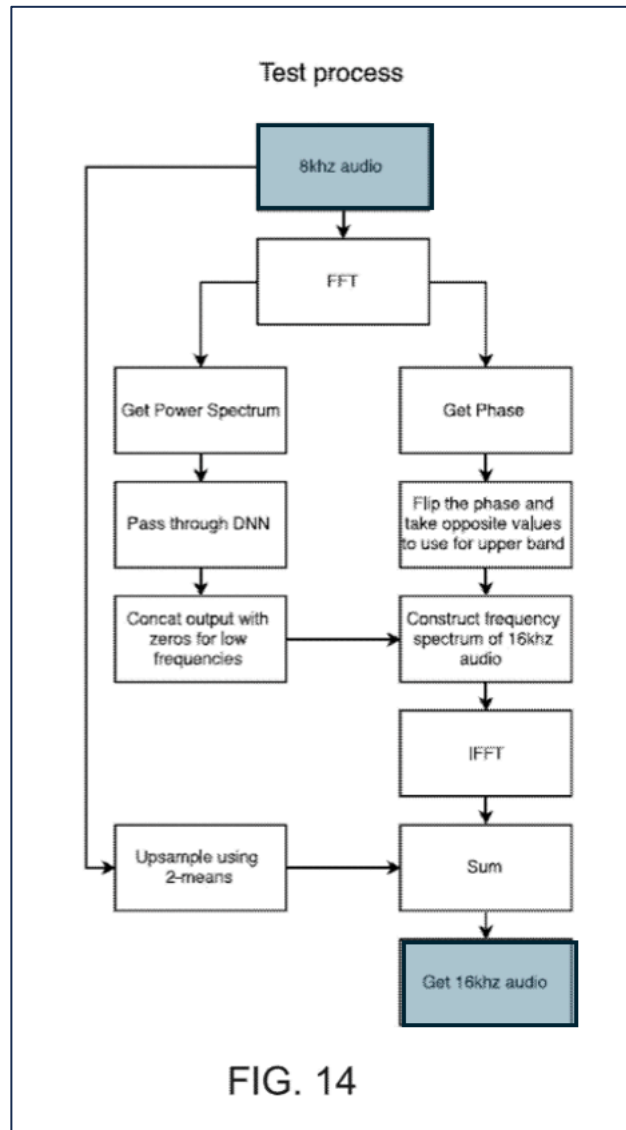
reasonable expectation of success. This was well known to do. For example, *Balasubramaniam* teaches an “audio capture device 16 is a device configured to capture audio of the conversation between the information gatherer and information provider and generate audio data comprising a digital representation of the conversation.” EX1019-Balasubramaniam, [0083]. It was well known that converting the audio signal into a digital signal will be easier to process. For example, *Prabhavalkar* teaches to “receive spoken information from a user and convert it to **usable** digital information.” EX1005-Prabhavalkar, 26:57-59. (emphasis added) As another example, *Xydeas* teaches “[d]igital computers and thus complex signal processing, not easily accomplished otherwise, can be achieved. Information in a digital format can be encrypted and hence, secrecy, especially important in military communications, is obtained.” EX1017-Xydeas, 18.

171. *Feinauer* does not specifically discuss its sample rate. However, *Feinauer-Prabhavalkar-Fan* in view of *Sargsyan* teach an *input audio signal digitized with a first sample rate and an output audio signal digitized with a second sample rate different from the first sample rate*.

a. *Sargsyan’s Teachings*

172. U.S. Patent Appl. No. 2020/0066296 to Sargsyan et al. (“*Sargsyan*”) teaches a speech processing system in which input audio sampled at a first sample rate is converted to a second, different sample rate. EX1007-Sargsyan, [0036]. In a

specific example, “8 kHz sampled audio” (*a first sample rate*) “is converted to 16 kHz” (*second sample rate different from the first sample rate*). EX1007-Sargsyan, [0036]. *Sargsyan* teaches sample rate conversion may improve the received output sound quality and/or save network traffic. (“people calling in from low bitrate cellular lines [to] sound much richer for other participants in the call (from 16 kHz to 44.1 kHz). The methods and systems can also be used to expand low bitrate audio (e.g., 8 kHz) on user device so that audio can be sent in low bandwidth mode and then expanded on the other end (from 16 kHz to 44.1 kHz), thereby saving network traffic.” EX1007-Sargsyan, [0027]. *Sargsyan*’s sample rate conversion is illustrated in Fig. 14 below.



EX1007-Sargsyan, Fig. 14 (annotated).

b. Motivation to Combine *Sargsyan's* sampling rate

173. A PHOSITA would have been motivated to employ *Sargsyan's* method of converting a “8 kHz sampled audio” (*a first sample rate*) “to 16 kHz” (*second sample rate different from the first sample rate*) in *Feinauer's* real-time accent conversion method. A PHOSITA would have appreciated such a modification would

have resulted in “much richer” sound. EX1007-Sargsyan, [0027]. A PHOSITA would have recognized that applying *Sargsyan’s* sampling techniques to *Feinauer’s* system would have only required a reprogramming according to known techniques to produce the reasonable result of saving network traffic in *Feinauer’s* accent conversion method by expanding low bitrate audio. Specifically, *Sargsyan* teaches “8 kHz sampled audio” (*a first sample rate*) “is converted to 16 kHz” (*second sample rate different from the first sample rate*). EX1007-Sargsyan, [0036]. *Sargsyan* teaches this allows “people calling in from low bitrate cellular lines [to] sound much richer for other participants in the call (from 16 kHz to 44.1 kHz). The methods and systems can also be used to expand low bitrate audio (e.g., 8 kHz) on user device so that audio can be sent in low bandwidth mode and then expanded on the other end (from 16 kHz to 44.1 kHz), thereby saving network traffic.” EX1007-Sargsyan, [0027]. A PHOSITA would have recognized that applying *Sargsyan’s* sampling techniques to *Feinauer’s* system would have only required a reprogramming according to known techniques to produce the reasonable result of being able to save network traffic in *Feinauer’s* accent conversion method when processing a digital signal, resulting in a REOS. EX1007-Sargsyan, [0027] emphasis added (“people calling in from low bitrate cellular lines [to] sound much richer for other participants in the call (from 16 kHz to 44.1 kHz). The methods and systems can also be used to expand low bitrate audio (e.g., 8 kHz) on user device so that audio can be sent in

low bandwidth mode and then expanded on the other end (from 16 kHz to 44.1 kHz), **thereby saving network traffic.**”) In one example, *Sargsyan*’s resampling rate may be advantageous to do when *Feinauer*’s incoming audio is coming from a phone with a 8kHz rate and would have been motivated to up sample the audio to sound “much richer” as *Sargsyan teaches*. EX1007-Sargsyan, [0027]; EX1004-Feinauer, [0008], [0013], and [0041]. See EX1021-Schmandt, 220. *Feinauer* already envisions sending its signal over a network. EX1004-Feinauer, [0041] (“The computing device **210** can send and receive communication signals (e.g. via voice over internet protocol (VoIP) or public switched telephone network (PSTN))...The communication network **220** can send and receive signals originating from computing device **210** and communication device **230**, such as audio and/or video signals, for example VoIP communications.”), [0043], [0047]-[0048] (“In some implementations, the computing device **210** can include an audio transmitter**312** and audio receiver **314** for sending and receiving audio-based communications. The audio transmitter **312** can comprise any combination of hardware and software for communicating, sending or transmitting signals comprising audio content in any type or form.”).

B. Claim 8

1. ***Claim 8: “The computing system for real-time correction of claim 6, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.”***

174. See Claim 3.

C. Claim 13

1. Claim 13: “*The non-transitory computer-readable storage medium of claim 11, wherein: the input audio signal is digitized with a first sample rate; and the output audio signal is digitized with a second sample rate different from the first sample rate.*”

175. See Claim 3.

XI. OPINIONS REGARDING GROUND 3: CLAIMS 4, 9, AND 14 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND KALINLI*

A. Claim 4

2. Claim 4: “*The method of claim 1, further comprising processing, in parallel via at least two processing units of the computing system, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk.*”

176. As discussed regarding limitations 1(a)(2)-1(a)(3), *Feinauer-Prabhavalkar-Fan* teach extracting acoustic features and linguistic features. See Claim 1(a)(2)-1(a)(3). Further, *Feinauer* teaches “multiple processors may provide functionality for *parallel*, simultaneous execution of instructions or for *parallel*, simultaneous execution of one instruction **on more than one piece of data.**” EX1004-Feinauer, [0032]. However, *Feinauer* does not describe performing such parallel processing *via at least two processing units of the computing system.*

a. Kalinli’s Teachings

177. *Kalinli* teaches “multiple processors could extract different features from an input sound window in parallel.” EX1009-Kalinli, [0048]. *Kalinli* teaches the “ability to process data in parallel saves valuable processing time, leading to a more efficient and streamlined system for tone/intonation recognition.” EX1009-Kalinli, [0048].

178. *Kalinli* teaches a speech processing system that extracts features using a plurality of processors. EX1009-Kalinli, [0017] (speech processing), [0021]-[0023] (feature extraction), [0048]-[0049] (multiple processors). Specifically, *Kalinli* envisions “multiple processors extract[ing] different features from an input sound window in parallel.” EX1009-Kalinli, [0048]. *Kalinli* provides an example in which a parallel speech processing system has 8 “synergistic” processing elements (SPE) operating under a single power processing element (PPE). EX1009-Kalinli, [0048]-[0051], Fig. 4. Different processors are dedicated to extraction of particular features such as intensity or frequency features. EX1009-Kalinli, [0048]-[0049] (“multiple processors could extract different features from an input sound window in parallel. The intensity extraction filter, frequency contrast extraction filter, temporal contrast extraction filter, and orientation extraction filter could all process the auditory spectrum simultaneously, generating corresponding multi-scale features much more efficiently if implemented on different processor elements in parallel.”). Thus, *Kalinli* provides express disclosure of performing parallel processing of

different speech features *via at least two processing units of the computing system*. As explained below, a PHOSITA would have been motivated to modify the *Feinauer* system according to *Kalinli's* disclosure to extract the different linguistic and acoustic features using a plurality of parallel processing units. As a specific example, a PHOSITA would have been motivated to extract linguistic features from frames of input speech on a first processor while, in parallel, extracting acoustic features from the same frames on a second processor. Thus, in combination, *Feinauer-Prabhavalkar-Fan* in view of *Kalinli* teaches *processing, in parallel via at least two processing units of the computing system, extracting the acoustic features, extracting the linguistic features*.

b. Motivation to Combine

179. A PHOSITA would have been motivated to modify *Feinauer-Prabhavalkar-Fan* parallel processing teachings to include *Kalini's* teachings of multiple processors to extract features. This was common in the prior art and advantageous to do. For example, *Prabhavalkar* teaches “multiple model replicas are trained in parallel and adjustments to the parameters of the model replicas are combined synchronously.” EX1005-Prabhavalkar, 5:39-43. *Prabhavalkar* teaches “multitasking and parallel processing may be advantageous.” EX1005-Prabhavalkar, 28:20-21. A PHOSITA would have appreciated that *Feinauer* already teaches “multiple processors may provide functionality for parallel, simultaneous

execution of instructions or for *parallel*, simultaneous execution of one instruction on more than one piece of data” but does not provide further detail and would have been motivated to look towards *Kalini*’s teachings for that detail. EX1004-Feinauer, [0032].

180. A PHOSITA would have appreciated that *Kalinli* teaches the “ability to process data in parallel saves valuable processing time, leading to a more efficient and streamlined system for tone/intonation recognition.” EX1009-Kalinli, [0048]. A PHOSITA would have been motivated to apply *Kalini*’s teachings so that one of *Feinauer*’s processor extracts the acoustic features as discussed in Claim 1(a)(2) and *Feinauer*’s second processor extracts the linguistic features as discussed in Claim 1(a)(3). *Kalinli* already teaches “generating corresponding multi-scale features much more efficiently if implemented on different processor elements in parallel.” EX1009-Kalinli, [0048]. A PHOSITA would have been motivated to perform the Claim 1(a)(2)-1(a)(3) process would be completed in parallel via at least two processing units of the computing system, dedicating particular processors to particular feature extraction as in *Kalinli*, improving the accuracy and speed of the system. EX1009-Kalinli, [0039]-[0042]. (“resulting in an accuracy of 79.0%.”) A PHOSITA would have recognized that applying *Kalinli*’s techniques to *Feinauer*’s system would have only required software reprogramming according to known techniques to produce the reasonable result of increased synchronization in

Feinauer-Prabhavalkar-Fan accent conversion method resulting in a REOS. The combination constitutes use of a known technique (*Kalinli*'s delegation of particular feature extraction responsibilities to particular processors) to improve similar devices (the modified *Feinauer* and *Kalinli* ML speech processing systems) in the same way.

2. ***Claim 9: "The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk."***

181. See Claim 4.

3. ***Claim 14: "The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to process, in parallel, two or more of the extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, or generating the output chunk."***

182. See Claim 4.

XII. OPINIONS REGARDING GROUND 4: CLAIMS 5, 10, 15 ARE OBVIOUS OVER *FEINAUER, PRABHAVALKAR, FAN, AND WATTS*

A. Claim 5

1. ***Claim 5: "The method of claim 1, further comprising: extracting the acoustic features, extracting the linguistic features, synthesizing the spectrum representation, and generating the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and providing, as***

an input to at least one of the modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.”

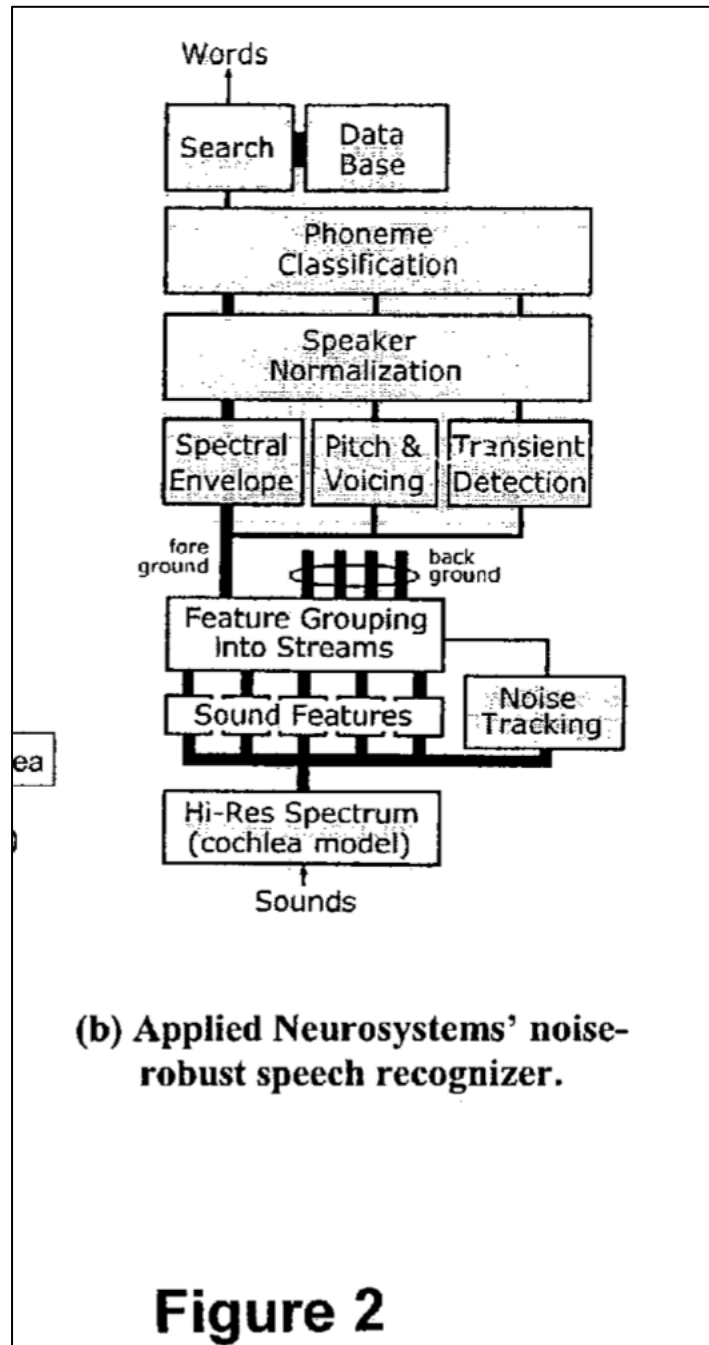
183. The '561 describes a “computational graph” as including a “pipeline.” EX1001-7:43-44, 16:31-34, 17:19-20. Though *Feinauer* envisions parallel processing, it does not teach the recited computational graph time-shifting details of Claim 5. Specifically, *Feinauer* teaches “multiple processors may provide functionality for parallel, simultaneous execution of instructions or for *parallel*, simultaneous execution of one instruction on more than one piece of data.” EX1004-*Feinauer*, [0032]. Further, the '561 describes time-shifts as being “selected to find a trade-off for quality of output feature vs provided delay. Higher shift values (for example, 100 ms) can be selected for linguistic features and lower shift values can be selected for acoustic features (for example, 20 ms) and input for vocoding (for example, 40 ms).” EX1001-17:50-56.

184. In the combination discussed above, Fan’s style separation model, voice feature extraction model, and vocoder are combined to *Feinauer* to extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk as recited in claim 5. *See* Claim 1(a)(2)-1(a)(3) and 1(c). Further, each of these modules in the system receives input

data for processing and subsequently outputs data, as claimed. The modified Feinauer system does not teach computational graphs (pipelines) using time-shifts.

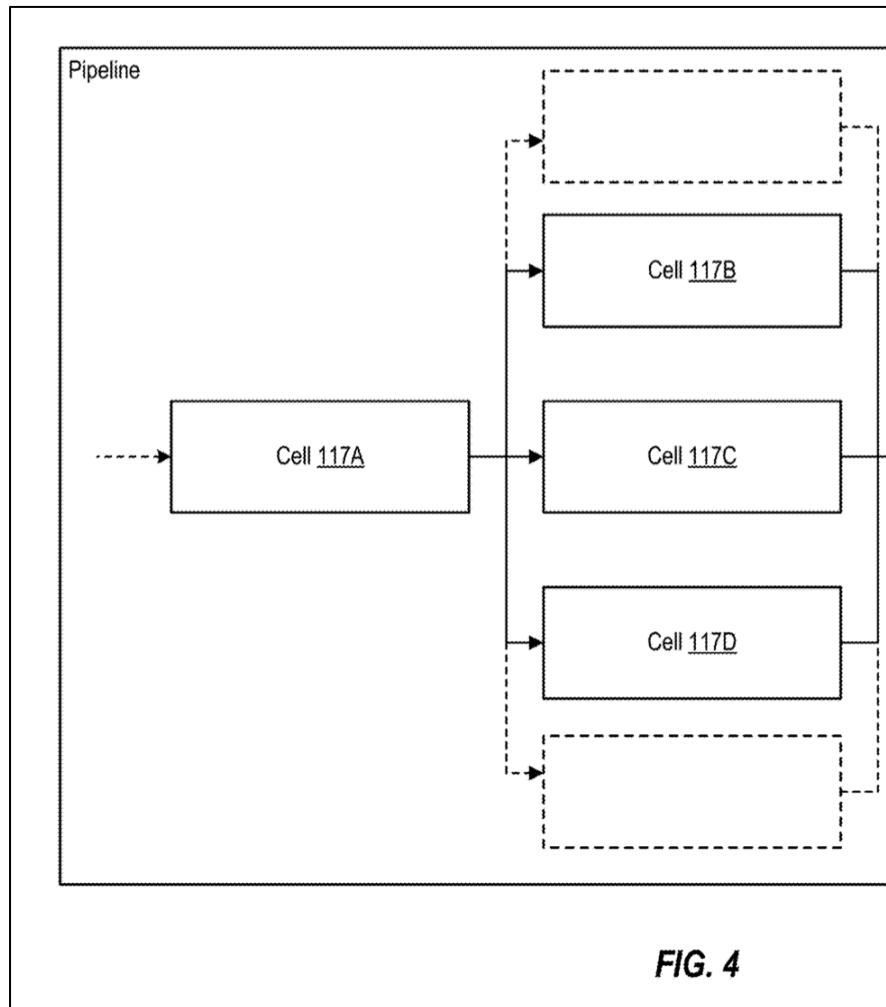
a. *Watt's Teachings*

185. However, *Watts* teaches a “computational pipeline” for phoneme and speech sound classification. EX1008-Watts, 5:7-47, Fig. 2B.



EX1008-Watts, Fig. 2B. Computational graphs (pipelines) were well-known and widely-used approaches used in speech processing for improving speed and efficiency by coordinating the timing of the processing. It was well known that pipeline's improve efficiency and speed of a speech processing system. EX1037-

Ma, 8:14-23. (“the disclosed pipeline may provide greatly improved flexibility and maintainability. Further, by utilizing specific functional cells to replace certain tasks that conventionally require human manipulation of data, the disclosed pipeline provides improved consistency, while at the same time decreasing the management overhead of a pipeline for language modeling.”) A PHOSITA would have appreciated that a “pipeline will operate exactly the same way using the selected default parameters each time, avoiding costly human errors leading to backtracking and repeating the data sampling process.” EX1037-Ma, 14:8-11. A pipeline “provides a highly efficient process, greatly saving labor, cost and time for language modeling.” EX1037-Ma, 8:1-2. It was well known that a pipeline receives audio data, selects audio data, and slices the audio data into “chunks of desired sizes” and after slicing the audio data into chunks, the audio is further processed in each part of the pipeline, thus coordinating the timing of the processing. EX1037-Ma, 3:30-55, Fig. 4.



EX1037-Ma, Fig. 4. *Watts* provides details of numbers suggesting that some speech events may be classified at different time scales in a speech processing system. EX1008-Watts, 4:43-47, 7:5-20. Thus, *Watts* teaches the claimed *computational graph*.

186. *Watts* teaches time parameters for acoustic and linguistic feature extraction. *Watts's* phoneme classifier extracts linguistic features on a “150 ms context” EX1008-Watts, 7:5-20. Further, *Watts* teaches extraction of acoustic features such as pitch “on a much finer time-scale (e.g., 1 ms) so as to be able to

track fast formant changes associated with consonant closures.” EX1008-Watts, 4:43-47. Thus, a PHOSITA would have recognized that *Watts* teaches the *time-shift of acoustic features* in its computational graph are *lower* than the *time-shift of linguistic features*. It was well known acoustic and linguistic features have different time-shift parameters. For example, Danjyo teaches “the unit of time for linguistic features differs from that for acoustic features.” EX1038-Danijyo, [0088]. A PHOSITA would have expected these time-shift parameters would have been provided to *Feinauer-Prabhavalkar-Fan*’s style separation model and voice feature extraction model respectively in the combined system, since those are the components receiving voice audio data for feature extraction. (*providing, as an input to at least one of the modules, a time-shift parameter*)

187. *Watt*’s 1ms acoustic time shift parameter and 150 ms linguistic time shift parameter in view of *Feinauer-Prabhavalkar-Fan*’s chunk teachings in Claim 1(a)(1)-1(c), *indicates a difference in timestamps of the chunk of the stream of chunks of the input audio signal and Fan*’s style separation model and voice feature extraction model (*one of the input data or the output data of the at least one of the respective modules*). A motivation to combine with *Feinauer-Prabhavalkar-Fan*’s chunks for the input and output data was provided above in Claim 1(a)(1)-1(c) and is incorporated by reference.

b. *Fan*’s Teachings

188. Each of *Fan's* respective modules (style separation model, voice feature extraction model, and vocoder) are associated with input data and output data. EX1006-Fan, Figs.7 and 9. A motivation to combine with *Fan* was provided above in Claim 1(a)(2)-1(a)(3)-1(c) and is incorporated by reference. For example, *Fan's* output of the style separation model is a style embedding vector and the output of the voice feature extraction model is a PPG. See Claim 1(a)(3)-1(b); EX1006-Fan, [0127], Fig. 9.

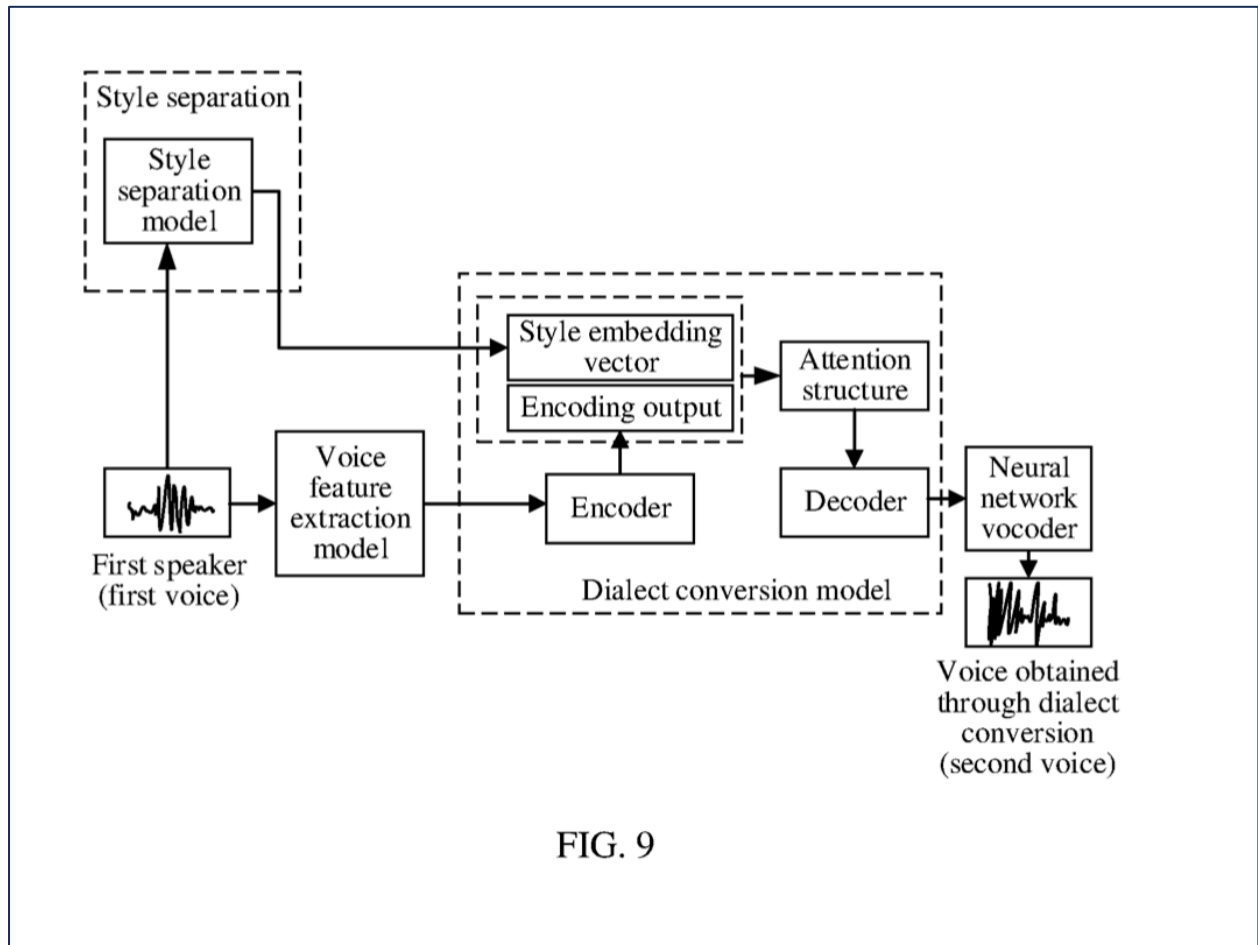
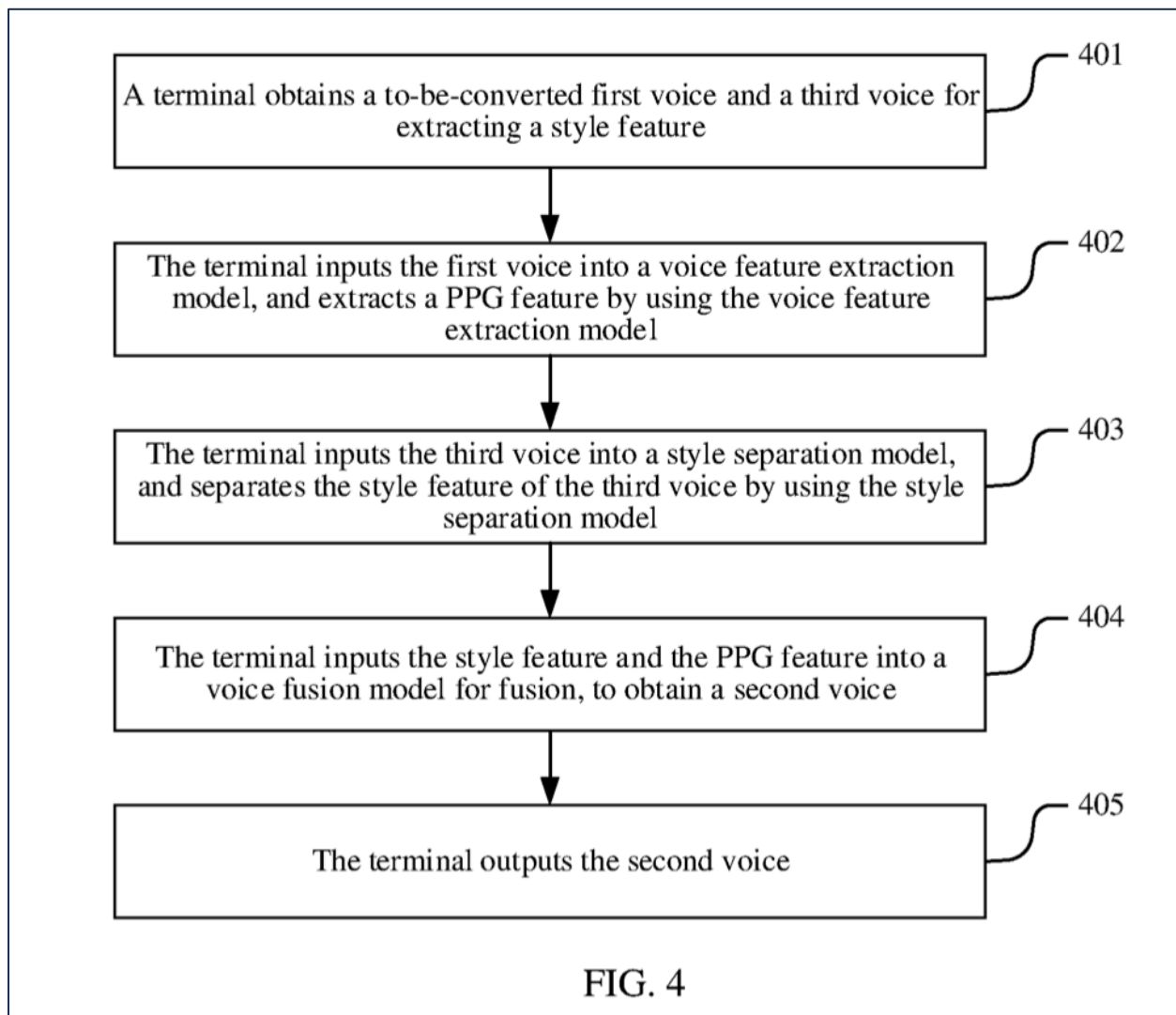


FIG. 9

EX1006-Fan, Fig. 9. As another example, the input of *Fan's* style separation model and voice feature extraction model is voice audio. EX1006-Fan, [0127], Fig. 4.



EX1006-Fan, Fig. 4. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan's* modules described functionality in *Fan's* style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. A PHOSITA would have recognized descriptions related to any hardware and/or software related *Fan's* modules described functionality in *Fan's* style enhancement embodiment would have been equivalently applicable to the second dialect conversion embodiment as well. I note

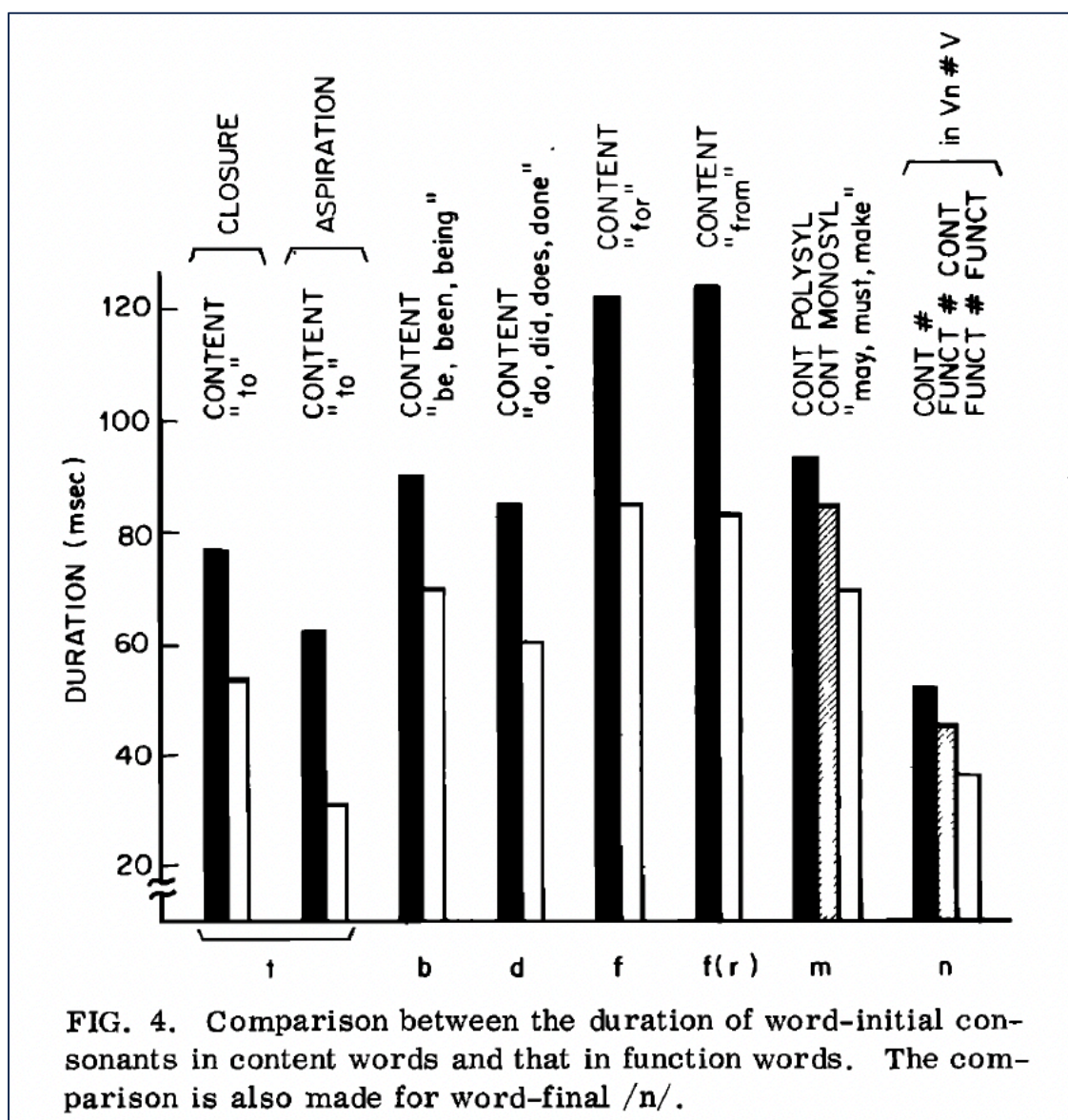
the description of a style feature discussed in *Fan* with regard to an embodiment employing a “style enhancement model,” an embodiment which is not directly relied-upon in this mapping. *See, e.g.* EX1006-Fan, [0075], [0095] (discussing female and male announcers). However, the description of the style enhancement model in regard to a style feature provided is applicable to the later “dialect conversion” embodiment that is being relied-upon in this mapping. Thus, a PHOSITA would have recognized descriptions related to any hardware and/or software related to *Fan*’s models/modules are applicable all of *Fan*’s envisioned embodiments.

c. Motivation to Combine *Watt*’s

189. A PHOSITA would have been motivated to modify *Feinauer-Prabhavalkar-Fan* in view of *Watts* to teach *generating the output chunk* using *Fan*’s respective modules (style separation model, voice feature extraction model, and vocoder) of a *computational graph*, as demonstrated by *Watts*. A PHOSITA would have appreciated that such a modification would have been the use of known technique (*Watts*’s computational graph architecture) to improve similar devices (*Feinauer-Prabhavalkar-Fan*’s accent conversion system) in the same way, thus resulting in increased processing speed and efficiency. It was well known that pipeline’s improve efficiency and speed of a speech processing system. EX1037-Ma, 8:14-23. (“the disclosed pipeline may provide greatly improved flexibility and

maintainability. Further, by utilizing specific functional cells to replace certain tasks that conventionally require human manipulation of data, the disclosed pipeline provides improved consistency, while at the same time decreasing the management overhead of a pipeline for language modeling.”) A PHOSITA would have appreciated that a “pipeline will operate exactly the same way using the selected default parameters each time, avoiding costly human errors leading to backtracking and repeating the data sampling process.” EX1037-Ma, 14:8-11. A pipeline “provides a highly efficient process, greatly saving labor, cost and time for language modeling.” EX1037-Ma, 8:1-2.

190. A PHOSITA would have been further motivated to specifically configure the combined system to utilize *a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features* as demonstrated by *Watts*. It was well known in the prior art that linguistic features (phonemes such as represented by PPGs) occur slower and across multiple frames than acoustic features, such as pitch. For example, Consonant duration in American English to Umeda teaches phonemes span 20+ms per phoneme as depicted below in Fig. 4.



EX1035-Umeda, Fig. 4. As another example, *Orr* teaches “a phoneme (phoneme size being ~30-280 ms)” EX1036-Orr, [0058]. *Orr* also teaches acoustic features such as intonation can occur as quickly as “1ms” EX1036-Orr, [0057], Claim 21. A PHOSITA would have appreciated that such a modification would have been the use of known technique (*Watts’s* time shift parameters) to improve similar devices (*Feinauer-Prabhavalkar-Fan’s* accent conversion system) in the same way, thus

resulting in increased processing speed and accuracy, allowing *Feinauer-Prabhavalkar-Fan*'s accent conversion system to extract even small nuances of pitch change as taught by Watts. *Watts* teaches extraction of acoustic features such as pitch “on a much finer time-scale (e.g., 1 ms) so as to be able to track fast formant changes associated with consonant closures.” EX1008-Watts, 4:43-47. A PHOSITA would have had a reasonable expectation of success in making such a modification because the combination required only reprogramming of *Feinauer-Prabhavalkar-Fan*'s software using ordinary skill of a PHOSITA and hardware already present.

B. Claim 10

1. *Claim 10: “The computing system for real-time correction of claim 6, wherein the processors are further configured to execute the stored instructions to: extract the acoustic features, extract the linguistic features, synthesize the spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.”*

191. See Claim 5.

C. Claim 15

1. *Claim 15: “The non-transitory computer-readable storage medium of claim 11, wherein the instructions, when executed by the processors, further cause the processors to: extract the acoustic features, extract the linguistic features, synthesize the*

spectrum representation, and generate the output chunk using respective modules of a computational graph, wherein each of the respective modules is associated with input data and output data; and provide, as an input to at least one of the respective modules, a time-shift parameter indicating a difference in timestamps of the chunk of the stream of chunks of the input audio signal and one of the input data or the output data of the at least one of the respective modules, wherein the acoustic features are assigned a lower time-shift parameter than the linguistic features.”

192. See Claim 5.

XIII. CONCLUSION

193. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: March 18, 2026

By: 

Christopher Schmandt

Christopher Schmandt

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

Education

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

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

Professional Experience -- MIT

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

Architecture Machine Group, Research Associate, 1980-1984

Architecture Machine Group, Research Assistant, 1979-1980

Architecture Machine Group, Graphics Programmer, 1977-1979

Departmental Undergraduate Research Opportunities Program Coordinator, 1984-2018

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

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

Sponsored Research Activities

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

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

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

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

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

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

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

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

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

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

Professional experience -- intellectual property

Testifying legal engagements within the last five years, representing party in italics

- *Microsoft v. SpeakWare*, 2019
- *Bumble Trading LLC v. Match Group LLC*, 2019
- *Tile Inc. v. Cellwitch Inc.*, 2019
- *Blackberry Ltd v. Facebook, Inc* 2020
- *EROAD Ltd v. PerDiem Co.*, 2020
- *Snap Inc. v. SRK Technology LLC*, 2020
- *Shopify Inc Ltd v. Express Mobile Inc.*, 2020
- *Express Mobile Inc, v Wix Ltd* 2020
- *Motorola Mobility LLC v. Ironworks patents, LLC.* 2021
- *Express Mobile Inc, v GoDaddy Inc* 2021
- *Bumble Trading LLC v. Kinectus, LLC.* 2021
- *Tile, Inc v. Linquet Technologies, Inc.* 2021
- *Facebook, Inc. v. Palo Alto Research Center, Inc.* 2021
- *Unified Patents, LLC v. Gesture Technology Partners, LLC.* 2021
- *Quantum Metric, Inc v. Content Square Isreal, Ltd.* 2021
- *Meta Platforms Inc. v. AlmondNet, Inc* 2022
- *Apple Inc. v. Zentian Limited* 2022
- *Google LLC v. Mira Advanced Technology Systems Inc* 2022
- *Google LLC v. Robocast* 2023
- *Google LLC v. Buffalo Patents LLC* 2023
- *Amazon v. AlmondNet* 2023
- *Google LLC v. Valtrus* 2023
- *Samsung Electronics Co., Ltd v. Stanton Techiya, LLC* 2024
- *Roku, Inc. v. AlmondNet, Inc.* 2024
- *LinkedIn v. AlmondNet, Inc* 2024
- *Google, LLC v. Dialect, LLC* 2024
- *SAP America, Inc. v. Cyandia, Inc.* 2024
- *Tableau Software LLC v. iCharts LLC* 2024
- *Google LLC and Samsung Electronics Company., Ltd. v. Cerence Operating Company* 2024
- *Samsung Electronics Company., Ltd. v. Cerence Operating Company* 2025
- *Amazon.com, Inc. v. KAIFI* 2025

Co-inventor on U.S. Patents

- 5,177,685 for "Automobile navigation system using real time spoken driving instructions"
- 6,728,348 for "System for storing voice recognizable identifiers using a limited input device such as a telephone key pad"
- 6,937,986 for "Automatic dynamic speech recognition vocabulary based on external sources of information"
- 7,098,776 for "Methods and apparatus for vibrotactile communication"
- 7,392,280 for "Method for summarization of threads in electronic mail"
- 7,738,637 "Interactive voice message retrieval"
- 7,443,283 for "Methods and apparatus for connecting an intimate group by exchanging awareness cues and text, voice instant messages, and two-way voice communications"
- 7,865,560 for "System for summarization of threads in electronic mail"
- 8,121,653 for "Methods and apparatus for autonomously managing communications using an intelligent intermediary"
- 8,135,128 for "Animatronic creatures that act as intermediaries between human users and a telephone system"

Co-inventor on U.S. Patent Applications

- 20020076009 "International dialing using spoken commands"
- 20020087328 "Automatic dynamic speech recognition vocabulary based on external sources of information"
- 20020064257 "System for storing voice recognizable identifiers using a limited input device such as a telephone keypad"
- 20030023688 "Voice-based message sorting and retrieval method"
- 20030081738 "Method and apparatus for improving access to numerical information in voice messages"
- 20030144846 "Method and system for modifying the behavior of an application based upon the application's grammar"
- 20030158903 "Calendar bar interface for electronic mail interaction"

Publications

Field Study of a Tactile Sound Awareness Device for Deaf and Hard of Hearing Users ISWC 2020. (with D. Jain, B. Chiu, S. Goodman, L. Findlater, J. Froehlich)

SkinMorph: Texture-Tunable On-Skin Interface Through Thin, Programmable Gel ISWC 2018. (with Cindy Kao, M. Banforth, D. Kim)

Technical Interventions to Detect, Communicate, and Deter Sexual Assault. ISWC 2017. (with Manisha Mohan)

Exploring Interactions and Perceptions of Kinetic Wearables. DIS 2017. (with Cindy Hsin-Liu Kao, D. Ajilo, O. Anilonyte, A. Dementyev, I. Choi and S. Follmer)

Leveraging User-made Predictions to Help Understand Personal Behavior. MobileHCI 2017. (with Miriam Greis, Tilman Dangler and Albrecht Schmidt)

Rovables: On-Body Robots as Mobile Wearables. UIST 2016. (with Cindy Hsin-Liu Kao, A. Dementyev, I. Choi, D. Ajilo, M. Xu, and S. Follmer)

DuoSkin: Rapidly Prototyping On-Skin User Interfaces Using Skin-Friendly Materials. ISWC 2016. (with Cindy Hsin-Liu Kao, Christian Holz, Asta Roseway, and Andres Calvo)

Immersive Terrestrial Scuba Diving Using Virtual Reality (with Dhruv Jain, Misha Sra, Jingru Go, Rodrigo Margues, Raymond Wu and Justin Chiu) Proceedings, UIST 2016

Expanding social mobile games beyond the device screen (with Misha Sra) Journal of Personal and Ubiquitous Computing, 2015

NailO: Fingernails as an input surface (with Cindy Hsin-Liu Kao, Artem Dementyev, Joseph Paradiso) CHI 2015

Mugshots: A mug display for front and back stage social interaction in the workplace (with Cindy Hsin-Liu Kao) TEI (Tangible and Embedded Interfaces) 2015

Mime: compact, low power 3D gesture sensing for interaction with head mounted displays (with Andrea Colaco, Ahmend Kirmani, Hye Soo Yang, Nan-Wei Gong, and Vivek Goyal) Proceedings of UIST 2013.

Spotz: A location-based approach to self-awareness (with Misha Sra) Proceedings of Persuasive 2013.

Setting the stage for interaction: A tablet application to augment group discussion in a seminar class (with Drew Harry and Eric Gordon) Proceedings of CSCW 2012.

Indoor Location Sensing using Geo-Magnetism (with Jaewoo Chung, Matt Donahoe, Ig-Jae Kim, Pedram Razavai and Micaela Wiseman) Proceedings of International Conference on Mobile Systems, Applications, and

Services (Mobisys) 2011.

My second bike: a TV-enabled social and interactive riding experience (with Jaewoo Chung, Kuang Xu, Andrea Colaco, and Victor Li) Proceedings of IEEE Communications and Networking Conference, Jan 2010.

Going my way?: User-aware route planner (with Jaewoo Chung) Proceedings CHI 2009.

Globetoddler: Designing for remote interaction between preschoolers and their traveling parents (with Paulina Modlitba) CHI 2008 Extended Abstracts

Are we there yet? - a temporally aware media player (with Matt Adcock and Jaewoo Chung), Australian User Interface Conference (AUIC) 2008

Physical embodiments for mobile communication agents (with Stefan Marti), UIST 2005

Giving the caller the finger: collaborative responsibility for cellphone interruptions (with Stefan Marti) Extended Abstracts, CHI 2005

Active Messenger: email filtering and delivery in a heterogeneous network (with Stefan Marti) Human-Computer Interaction Journal (HCI) Volume 20 (2005)

WatchMe: communication and awareness between members of a closely-knit group (with Natalia Marmasse) Proceedings of Ubicomp 2004

An audio-based personal memory aid (with S. Vemuri, W. Bender, S. Tellex and B. Lassey) Proceedings of Ubicomp 2004

Improving speech playback using time-compression and speech recognition (with Sunil Vemuri, Philip DeCamp, and Walter Bender) Proceedings of CHI 2004

Impromptu: managing networked audio applications for mobile users (with Kwan Lee, Jang Kim, and Mark Ackerman), Proceedings of MobiSys 2004

TalkBack: a conversational answering machine (with Vidya Lakshminpathy and Natalia Marmasse) Proceedings of UIST 2003

"ListenIn" to domestic environments from remote locations (with Gerardo Vallejo) Proceedings of the 2003 International Conference on Auditory Display (ICAD)

Safe & Sound: a wireless leash (with Natalia Marmasse) Extended Abstracts, Proceedings of CHI 2003

Mediated voice communication via mobile IP (with Jang Kim, Kwan Lee, Gerardo Vallejo, and Mark Ackerman), Proceedings of UIST 2002.

The Audio Notebook: Paper and pen interaction with structured speech (with Lisa Stifelman and Barry Arons), Proceedings of CHI 2001

Synthetic News Radio (with Keith Emnett) IBM Systems Journal, Vol. 39 Nos. 3-4, pp. 646-659, 2000.

Everywhere messaging (with Natalia Marmasse, Stefan Marti, Nitin Sawhney, and Sean Wheeler) IBM Systems Journal, Vol. 39 Nos. 3-4, pp. 660-677, 2000.

Location-aware information delivery with comMotion (with Natalia Marmasse), Proceedings of the Second International Symposium on Handheld and Ubiquitous Computing, pp. 157-171, Springer, 2000.

Nomadic Radio: Scalable and contextual notification for wearable audio messaging (with Nitin Sawhney), Proceedings of CHI 1999.

Speaking and listening on the run: Design for wearable audio computing (with Nitin Sawhney), Proceedings of International Symposium on Wearable Computing, 1998.

Audio Hallway: A virtual acoustic environment for browsing, Proceedings of UIST 1998.

Dynamic Soundscape: Mapping time to space for audio browsing (with Minoru Kobayashi), Proceedings of CHI 1997.

CLUES: Dynamic personalized message filtering (with Matt Marx), Proceedings of CSCW 1996.

Using acoustic structure in a hand-held audio playback device (with Deb Roy), IBM Systems Journal, Vol 35, Nos. 3 and 4, 1996.

Mailcall: Message presentation and navigation in a nonvisual environment (with Matt Marx), Proceedings of CHI 1996

AudioStreamer: Exploiting simultaneity for listening (with Atty Mullins), short paper, CHI 1995.

Multimedia momadic services on today's hardware, IEEE Network, September/October 1994.

Putting people first: Specifying proper names in speech interfaces (with Matt Marx), proceedings of UIST 1994.

Chatter: A conversational learning speech interface (with E. Ly) AAAI Workshop on Intelligent Multi-Media Multi-Modal Systems, 1994.

Voice Communication with Computers: Conversational Systems. New York: Van Nostrand Reinhold. 1994.

Capturing, structuring, and representing ubiquitous audio (with D. Hindus and C. Horner), ACM Transactions on Information Systems, Vol. 11, No. 4, October 1993.

Speech Recognition Architectures for Multimedia Environments, (with E. Ly and B. Arons), Proceedings of the 1993 AVIOS Conference, September 1993.

Phoneshell: the Telephone as Computer Terminal, Proceedings of the ACM Multimedia Conference, August 1993.

Voicenotes: A Speech Interface for a Hand-Held Voice Notetaker (with L. Stifelman, B. Arons, and E. Hulteen), Proceedings of INTERCHI'93, April 1993.

From Desktop Audio to Mobile Access: Opportunities for Voice in Computing, book chapter in Advances in Human-Computer Interaction, Vol. 4, H.R. Hartson and D. Hix editors. 1992.

Ubiquitous Audio: Capturing Spontaneous Collaboration (with D. Hindus), Proceedings of CSCW'92, November 1992.

Integrating Audio and Telephony in a Distributed Workstation Environment (with S. Angebrannt, R. Hyde, D. Luong, and N. Siravara), Proceedings of the Summer 1991 USENIX Conference, June 1991.

Augmenting a Window System with Speech Input (with M. Ackerman and D. Hindus), Computer, IEEE Computer Society, Vol. 23, No. 8, August 1990.

Observations on Using Speech Input for Window Navigation (with D. Hindus, M. Ackerman, and S. Manandhar), Proceedings, Human-Computer Interaction, Interact '90, IFIP, August 1990.

Phonetool: Integrating Telephones and Workstations (with S. Casner), Proceedings, GLOBECOM '89, IEEE Communications Society, November 1989.

Desktop Audio (with B. Arons), UNIX Review, October 1989.

Synthetic Speech for Real Time Direction-Giving (with J. Davis), IEEE Transactions on Consumer Electronics, IEEE, September 1989.

The Back Seat Driver: Real Time Spoken Driving Instructions (with J. Davis), Proceedings, IEEE Vehicle Navigation and Information Systems Conference, IEEE, Toronto, Canada, September 1989.

A Voice and Audio Server for Multimedia Workstations (with B. Arons, C. Binding, K. Lantz), Proceedings, Speech Tech 1989

An Audio and Telephone Server for Multi-media Workstations (with M. McKenna), Proceedings, Second IEEE Conference on Workstations, IEEE, Palo Alto, CA., 1988.

Employing Voice Back Channels to Facilitate Audio Document Retrieval, Proceedings, ACM Conference on Office Information Systems (COIS), Santa Clara, CA, 1988.

Conversational Telecommunications Environments, Proceedings, Second International Conference on Human-Computer Interaction, 1987.

Understanding Speech Without Recognizing Words, Proceedings, American Voice Input/Output Society Conference, AVIOS, 1987.

A Robust Parser and Dialog Generator for a Conversational Office System (with B. Arons and C. Simmons), Proceedings, American Voice Input/Output Society Conference, AVIOS, Palo Alto, CA, 1987.

Integrated Messages and Network Services for a Personal Workstation, IEEE Workshop on Telematics and Message Handling Systems, IEEE, 1986.

Voice Interaction in an Integrated Office and Telecommunications Environment, Proceedings, American Voice Input/Output Society Conference, AVIOS, San Francisco, CA, 1985.

Voice Communication with Computers, book chapter in Advances in Human-Computer Interaction, H. R. Hartson ed., 1985.

Speech Synthesis Gives Voiced Access to an Electronic Mail System, Speech Technology, Vol. 2, No. 3, Aug/Sept 1984.

A Conversational Telephone Messaging System (with B. Arons), IEEE Transactions on Consumer Electronics, IEEE, Vol CE-30, August 1984.

Phone Slave: A Graphical Telecommunications Interface (with B. Arons), Proceedings, Society for Information Display International Symposium, SID, San Francisco, CA, June 1984.

Input/Display Registration in a Stereoscopic Workstation, Displays, April 1984.

Remote Access to Voice and Text Messages, Proceedings, American Voice Input/Output Society Conference, AVIOS, Washington D.C., 1984.

Fuzzy Fonts: Analog Models Improve Digital Text Quality, Proceedings, National Computer Graphics Association Conference, Chicago, IL, 1983.

Greyscale Fonts Designed From Video Signal Analysis, Society of Applied Learning Technology, Houston, TX, 1983.

Spatial Input/Display Correspondence in a Stereoscopic Computer Graphic Work Station, Proceedings, ACM/SIGGRAPH, Detroit, MI, 1983.

A Programmable Virtual Vocabulary Speech Processing Peripheral (with W. Bender), Proceedings, American Voice Input/Output Society Conference on Voice Data Entry Systems Applications, AVIOS, 1983.

The Intelligent Voice Interactive Interface (with E. A. Hulteen), Proceedings, Human Factors in Computer Systems, National Bureau of Standards/ACM, Gaithersburg, MD, 1982.

Interactive Three-Dimensional Computer Space, Proceedings, SPIE Conference on Processing and Display of Three-Dimensional Data, SPIE, San Diego, CA, 1982, Vol. 367.

Speech Communications, a Systems' Approach, Proceedings, American Voice Input/Output Society Conference on Entry Systems Applications, 1982.

Voice Interaction: Putting Intelligence into the Interface, Proceedings, IEEE International Conference on Cybernetics and Society, IEEE, Seattle, WA, 1982.

The Intelligent Ear: A Graphical Interface to Digital Audio, Proceedings, IEEE International Conference on Cybernetics and Society, IEEE, Atlanta, GA, 1981.

Soft Typography, Information Processing 1980, IFIPS, S. Lavington ed., North-Holland Publishing Co., 1980.
