

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC.
Petitioner

v.

ZENTIAN LIMITED,
Patent Owner

Inter Partes Review Case No. IPR2023-00037
U.S. Patent No. 10,971,140

**DECLARATION OF CHRISTOPHER SCHMANDT
IN SUPPORT OF PETITION FOR *INTER PARTES* REVIEW OF
U.S. PATENT NO. 10,971,140**

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS	8
A.	EDUCATIONAL BACKGROUND AND PROFESSIONAL EXPERIENCE.....	8
II.	METHODOLOGY; MATERIALS CONSIDERED.....	11
III.	OVERVIEW AND LEGAL STANDARDS	13
A.	PERSON OF ORDINARY SKILL IN THE ART.....	14
B.	OBVIOUSNESS.....	14
C.	ANALOGOUS ART	20
D.	CLAIM CONSTRUCTION	21
IV.	LEVEL OF A PERSON OF ORDINARY SKILL	21
V.	OVERVIEW OF THE TECHNOLOGY	22
A.	SPEECH RECOGNITION.....	22
B.	FEATURE VECTORS	30
C.	ACOUSTIC MODELS	34
D.	HIDDEN MARKOV MODELS	35
E.	PRIOR ART SPEECH RECOGNITION SYSTEMS	43
VI.	OVERVIEW OF THE '140 PATENT	47
VII.	OVERVIEW OF THE PRIOR ART	48
A.	OVERVIEW OF JIANG	48
B.	OVERVIEW OF CHEN.....	48
C.	OVERVIEW OF LUCKE.....	49
D.	OVERVIEW OF ROBINSON	49
E.	OVERVIEW OF WRENCH	50
VIII.	OPINIONS REGARDING GROUND 1.....	50
A.	MOTIVATION TO COMBINE <i>JIANG</i> AND <i>CHEN</i>	50
B.	MOTIVATION TO COMBINE <i>CHEN</i> AND <i>JIANG</i>	52
C.	OBVIOUSNESS OF MODIFYING ELECTRONIC HARDWARE AND SOFTWARE TO BE CIRCUITRY	52
D.	CLAIM 1 IS RENDERED OBVIOUS BY <i>JIANG</i> IN VIEW OF <i>CHEN</i>	53
1.	<i>Claim 1</i>	53
a)	Claim 1(Pre): “A speech recognition circuit comprising”	53

b)	Claim 1(a): “one or more clusters of processors, each of the one or more clusters of processors comprising: a plurality of processors;”	54
c)	Claim 1(b): “and [one or more clusters of processors, each of the one or more clusters of processors comprising:] an acoustic model memory storing acoustic model data”	59
d)	Claim 1(c): “wherein each of the plurality of processors is configured to compute a probability using the acoustic model data in the acoustic model memory”	62
e)	Claim 1(d): “[wherein] the speech recognition circuit is configured to generate an initial score for an audio sample”	70
f)	Claim 1(e): “[wherein] the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.” ..	77
2.	<i>Claim 2: “...wherein the probability is an input to an evaluation of a state transition of a model of states”</i>	<i>83</i>
3.	<i>Claim 5: “...further comprising: a buffer for storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors”</i>	<i>85</i>
4.	<i>Claim 7: “...the one or more clusters of processors comprises a first cluster of processors and a second cluster of processors; the first cluster comprises a first acoustic model memory; and the second cluster comprises a second acoustic model memory that is distinct and separate from the first acoustic model memory”</i>	<i>87</i>

IX. OPINIONS REGARDING GROUND 2..... 89

A.	CLAIMS 1–3, 5, AND 7-8 ARE RENDERED OBVIOUS BY <i>JIANG</i> IN VIEW OF <i>CHEN</i> AND <i>LUCKE</i>	89
1.	<i>Claim 1</i>	89
a)	Claim 1(e): “[wherein] the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.” ..	89
2.	<i>Claim 5: “...further comprising: a buffer for storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors”</i>	<i>91</i>

X.	OPINIONS REGARDING GROUND 4.....	93
A.	CLAIM 4 IS RENDERED OBVIOUS BY <i>JIANG</i> IN VIEW OF <i>CHEN</i> AND <i>ROBINSON</i>	93
1.	<i>Claim 4: “...wherein the probability is computed from a Gaussian mixture model and one or more feature vectors.” ...</i>	93
XI.	OPINIONS REGARDING GROUND 6.....	94
A.	CLAIM 4 IS RENDERED OBVIOUS BY <i>JIANG</i> IN VIEW OF <i>CHEN</i> AND <i>WRENCH</i>	94
XII.	CONCLUSION	95

CLAIM LISTING

Claim 1:

Claim 1[Pre] A speech recognition circuit comprising:

1(a): one or more clusters of processors, each of the one or more clusters of processors comprising: a plurality of processors;

1(b): and an acoustic model memory storing acoustic model data,

1(c): wherein each of the plurality of processors is configured to compute a probability using the acoustic model data in the acoustic model memory,

1(d): wherein: the speech recognition circuit is configured to generate an initial score for an audio sample;

1(e): and the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.

Claim 2

The speech recognition circuit of claim 1, wherein the probability is an input to an evaluation of a state transition of a model of states.

Claim 3

The speech recognition circuit of claim 2, wherein the model is a Hidden Markov Model.

Claim 4

The speech recognition circuit of claim 2, wherein the probability is computed from a Gaussian mixture model and one or more feature vectors.

Claim 5

The speech recognition circuit of claim 1, further comprising:
a buffer for storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors.

Claim 6

The speech recognition circuit of claim 5, further comprising:
a search controller coupled to the at least one of the plurality of processors of the one or more clusters of processors, the search controller capable of controlling the at least one of the plurality of processors of the one or more clusters of processors to initiate speech recognition processing in at least a subset of processors of the one or more clusters of processors.

Claim 7

The speech recognition circuit of claim 1, wherein:
the one or more clusters of processors comprises a first cluster of processors and a second cluster of processors;
the first cluster comprises a first acoustic model memory; and
the second cluster comprises a second acoustic model memory that is distinct and separate from the first acoustic model memory.

Claim 8

The speech recognition circuit of claim 1, wherein:

the one or more clusters of processors comprises a first cluster of processors;

and

the first cluster comprises a first acoustic model coupled to all processors of the first cluster.

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 several years ago 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 recognition 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 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 built some of the earliest speech-based personal assistants for managing messages, calendar, contacts, etc. (Conversational Desktop 1985, Chatter 1993, MailCall 1996). We built quite a few systems employing speech recognition 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 the course of my research, I built a number of speech recognition client/server distributed systems, with the first being in 1985. Much of the initial motivation for a server architecture was that speech recognition required expensive digital signal processing hardware that we could not afford to put on each computer, so a central server with the required hardware was used. Later versions of the speech recognition server architecture allowed certain computers to perform specialized tasks serving a number of client computers providing voice user interfaces, either on screens or over telephone connections.

7. 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).

8. In my faculty position I taught graduate level courses in speech technology and user interaction design, and 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; MATERIALS CONSIDERED

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

10. For this work, I have been asked to review U.S. Patent No. 10,971,140 (“’140 Patent”) (Ex. 1001) including the specification and claims, and the ’140 Patent’s prosecution history (“’140 File History”) (Ex. 1002). In developing my opinions relating to the ’140 Patent, I have considered the materials cited or discussed herein, including those itemized in the Exhibit Table below.

Exhibit	Description
Exhibit 1001	U.S. Patent No. 10,971,140 (“’140 Patent”)
Exhibit 1002	Prosecution History for the 10,971,140 Patent (“’140 File History”)
Exhibit 1004	U.S. Patent No. 6,374,219 to Jiang (“Jiang”)
Exhibit 1005	U.S. Patent No. 5,428,803 to Chen et al. (“Chen”)
Exhibit 1006	Hsiao-Wuen Hon, <i>A survey of hardware architectures designed for speech recognition</i> , Carnegie Mellon University, 1991 (“Hon”)
Exhibit 1007	Ph.D. Thesis of Mosur Ravishankar (“Ravishankar”)
Exhibit 1008	U.S. Patent Application Publication No. 2001/0053974 to Lucke et al. (“Lucke”)
Exhibit 1009	U.S. Patent No. 5,983,180 to Robinson (“Robinson”)
Exhibit 1010	U.S. Patent No. 5,036,539 to Wrench Jr., et al. (“Wrench”)

Declaration of Christopher Schmandt
Patent No. 10,971,140

Exhibit 1011	Frederick Jelinek, <i>Statistical Methods for Speech Recognition</i> , The MIT Press, 1997 (“ <i>Jelinek</i> ”)
Exhibit 1012	Christopher Schmandt, <i>Voice Communication with Computers</i> , Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)
Exhibit 1013	Lawrence Rabiner and Biing-Hwang Juang, <i>Fundamentals of Speech Recognition</i> , Prentice Hall PTR, 1993 (“ <i>Rabiner</i> ”)
Exhibit 1014	Richard Klevans and Robert Rodman, <i>Voice Recognition</i> , Artech House, 1997 (“ <i>Klevans</i> ”)
Exhibit 1015	U.S. Patent No. 7,120,582 to Young et al. (“ <i>Young</i> ”)
Exhibit 1016	John Holmes and Wendy Holmes, <i>Speech Synthesis and Recognition</i> , 2nd Edition, Taylor & Francis, 2001 (“ <i>Holmes</i> ”)
Exhibit 1017	U.S. Patent No. 4,926,488 to Nadas et al. (“ <i>Nadas</i> ”)
Exhibit 1018	U.S. Patent No. 6,182,037 to Maes (“ <i>Maes</i> ”)
Exhibit 1019	U.S. Patent Application Publication No. 2003/0220796 to Aoyama et al. (“ <i>Aoyama</i> ”)
Exhibit 1020	U.S. Patent No. 6,092,045 to Stubbley et al. (“ <i>Stubbley</i> ”)
Exhibit 1021	U.S. Patent No. 6,151,574 to Lee et al. (“ <i>Lee</i> ”)
Exhibit 1022	U.S. Patent No. 6,580,814 to Ittycheriah et al. (“ <i>Ittycheriah</i> ”)
Exhibit 1023	U.S. Patent No. 6,452,348 to Toyoda (“ <i>Toyoda</i> ”)
Exhibit 1024	William A. Wulf and C.G. Bell, <i>C.mmp—A multi-mini-processor</i> , Carnegie-Mellon University, 1972 (“ <i>Wulf</i> ”)
Exhibit 1025	Lee D. Erman, Richard D. Fennell, Victor R. Lesser, and D. Raj Reddy, <i>System Organizations for Speech Understanding: Implications of Network and Multiprocessor Computer Architectures for AI</i> , IEEE Transactions on Computers, Vol. C-25, No. 4, April 1976 (“ <i>Erman</i> ”)
Exhibit 1026	U.S. Patent No. 6,393,481 to Deo et al. (“ <i>Deo</i> ”)
Exhibit 1027	U.S. Patent No. 6,615,338 to Tremblay et al. (“ <i>Tremblay</i> ”)
Exhibit 1028	U.S. Patent No. 5,922,076 to Garde (“ <i>Garde</i> ”)
Exhibit 1029	Lawrence R. Rabiner, <i>A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition</i> , Proceedings of the IEEE, Vol. 77, No. 2, February 1989 (“ <i>Rabiner 89</i> ”)
Exhibit 1030	Intentionally Left Blank
Exhibit 1031	Intentionally Left Blank
Exhibit 1033	Declaration of June Ann Munford (“ <i>Munford</i> ”)

11. I have considered these materials from the viewpoint of a POSITA as of the priority date of the '140 Patent. For the purposes of this declaration, I have

been asked to assume that the priority date of the '140 Patent is February 4, 2002. I note that my opinions provided in this Declaration are made from the perspective of a POSITA as of this priority date of the '140 Patent, unless expressly stated otherwise. To the extent that I use any verb tense in this Declaration that is present tense (e.g., “a POSITA would understand” instead of “a POSITA would have understood”), such verb tense should be understood to be my opinion as of the '140 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

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

A. Person of Ordinary Skill in the Art

13. I understand that the '140 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 ("POSITA"). 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.

14. 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 POSITA.

B. Obviousness

15. 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 POSITA 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 POSITA. I further understand that a hypothetical POSITA is assumed to know and to have all relevant prior art in the field of endeavor covered by the patent in suit and all analogous prior art. I understand that obviousness in an IPR review proceeding is evaluated using a preponderance of the evidence standard, which means that the claims must be more likely obvious than nonobvious.

16. 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 POSITA. 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 POSITA can implement a predictable variation, the variation is likely obvious;
- If a technique has been used to improve one device, and a POSITA 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 POSITA will be able to fit the teachings of multiple patents together like pieces of a puzzle;
- A POSITA 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 POSITA 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.

17. 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.

18. 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.

19. 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.

20. 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.

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

C. Analogous Art

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

D. Claim Construction

23. I understand that the claim terms in this IPR will be construed according to their ordinary and customary meaning as understood in light of the claim language, the patent's description, and the prosecution history viewed from the perspective of a POSITA. 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

24. Based on my review and analysis of the '140 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a POSITA in the field of the '140 Patent at the time of the earliest possible priority date (February 4, 2002) would have been knowledgeable regarding the field of speech recognition technology. In my experience working in this field, most workers of ordinary skill in the art as of the earliest possible priority date of February 4, 2002 would have had a master's degree in computer science, computer engineering, electrical engineering or an equivalent, and at least four years of professional experience relating to speech recognition technology. *See, e.g.*, the '140 Patent (Ex. 1001) at 1:18-58 (describing the "Field of the Invention" and providing a "Description of the Related Art"). A person with less relevant practical experience but with additional education can also qualify as a POSITA in the field of the '140 Patent provided the additional education

focused on speech recognition technology. When I refer to the understanding of a POSITA, I am referring to the understanding of such a person as of February 4, 2002.

25. As of February 4, 2002, I had more than ordinary skill in the art. I am, however, familiar with the skills and knowledge possessed by those I would have considered to be of ordinary skill in the art as of that date. For example, I knew people working in the field while I built systems employing speech recognition in handheld mobile devices (ComMotion 1999, Nomadic Radio 2000, and Impromptu 2001, for example).

26. My opinions provided in this declaration would not change in view of minor modifications to this level of ordinary skill.

V. OVERVIEW OF THE TECHNOLOGY

27. I was asked to briefly summarize the background of the technology from the standpoint of a POSITA prior to February 4, 2002.

A. Speech Recognition

28. Speech recognition using computer systems was known and widely used well before the priority date of the '140 Patent. Some of the earliest speech recognition systems were developed in the 1950s and 1960s. *Rabiner* (Ex. 1013), 37-38. By the 1980s, speech recognition systems using Hidden Markov Models were well known and “widely applied in virtually every speech-recognition research laboratory in the world.” *Rabiner*, 39. In general, speech recognition refers to “the

technologies used to allow computers to recognize the words in human speech.”

Schmandt (Ex. 1012), 45. Indeed, speech was known to be used in “a variety of computing environments based on appreciating its role in human communication.

Speech can be used as a method of *interacting* with a computer to place requests or receive warnings and notices. Voice can also be used as the underlying *data* itself[.]”

Schmandt, 11. Further, at least as early as 1994 (publication date of *Schmandt*),

“[d]esktop workstations can already support both these speech functions.” *Id.*

29. One of the initial processes in a speech recognition system is “parameter extraction,” which entails “methods of extracting information from speech waveforms. Parameter or feature extraction consists of preprocessing an electrical signal to transform it into a usable digital form, applying algorithms to extract only speaker-related information from the signal[.]” *Klevans* (Ex. 1014), 12. Such extraction involves “digital signal processing (DSP) methods that are common to all computer speech systems.” *Klevans*, 12-13. *Klevans* teaches that “the sound wave created by an individual’s speech is transduced into an analog electrical signal via a microphone. The electrical signal is sampled and quantized, resulting in a digital representation of the analog signal.” *Klevans*, 13. Following such analog to digital processing, “conditioning” may be performed on the signal, which includes bandpass filtering to remove additive noise and spectral flattening to compensate for

“the effect of the vocal tract on the excitation signal created by the vibrating vocal folds.” *Id.*

30. In general, it was known that speech recognizer systems included three basic components: (1) “speech **representation** that is computationally efficient for pattern matching” such as “Hidden Markov Models” that “are a statistical representation based on the various ways words or phonemes may be pronounced.” *Schmandt*, 45-46. I discuss Hidden Markov Models in more detail at Section V.D below. (2) A “set of **templates** or **models**, which are descriptions of each word to be recognized, in the representation of speech used by the recognizer. The templates describe the words in the recognizer’s vocabulary, i.e., those words that the recognizer can identify. They are reference models against which an input can be compared to determine what was spoken.” *Schmandt*, 46. (3) A “**pattern matching** algorithm to determine which template is most similar to a specimen of speech input. This element of the speech recognizer must determine word boundaries, locate the most similar template, and decide whether the difference between the input and the selected template is minor enough to accept the word.” *Id.* In general, a speech recognizer captures incoming speech and converts it “to the chosen internal representation (see Figure 7.1). The pattern matcher then selects the template that most closely matches the input or rejects the utterance if no template is a close

enough match.” *Id.* The functional elements of an exemplary speech recognizer are depicted below in Figure 7.1.

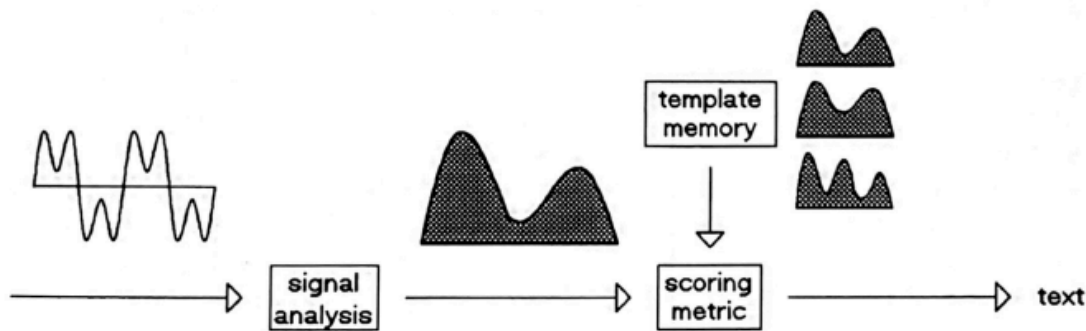


Figure 7.1. The functional elements of a speech recognizer. The user’s speech is digitized and converted to the recognizer’s internal representation. The captured speech is then compared with the words stored in the recognizer’s template memory. The pattern matching algorithm then determines which template, if any, is the closest match.

Schmandt, 46 (“Figure 7.1”).

31. When discussing speech recognition, one important topic is phonemes, which are speech sounds that form the words in a language.

The **phonemic** layer is concerned with the string of phonemes of which the words are composed. Phonemes are the speech sounds that form the words of any language. Phonemes include all the sounds associated with vowels and consonants. A grunt, growl, hiss, or gargling sound is not a phoneme in English, so it cannot be part of a word; such sounds are not referred to as speech. At the phoneme layer, while talking we are either continuously producing speech sounds or are silent. We are not silent at word boundaries; the phonemes all run together.

Schmandt, 19. *Schmandt* further discloses that “[p]honemes are important small units, several of which make up most syllables.” *Schmandt*, 23. Further, *Schmandt*

discloses that a “**phoneme** is a unit of speech, the set of which defines all the sounds from which words can be constructed in a particular language. There is at least one pair of words in a language for which replacing one phoneme with another will change what is spoken into a different word. Of course, not every combination of phonemes results in a word; many combinations are nonsense.” *Schmandt*, 23-24.

32. It was also known to utilize spectrograms for “characterizing the speech signal and representing the information associated with the sounds” “via a spectral representation.” *Rabiner*, 50. *Rabiner* teaches that “[p]erhaps the most popular representation of this type is the sound spectrogram in which a three-dimensional representation of the speech intensity, in different frequency bands, over time is portrayed.” *Id.* *Rabiner* teaches that the “spectral intensity at each point in time is indicated by the intensity (darkness) of the plot at a particular analysis frequency.” *Id.*

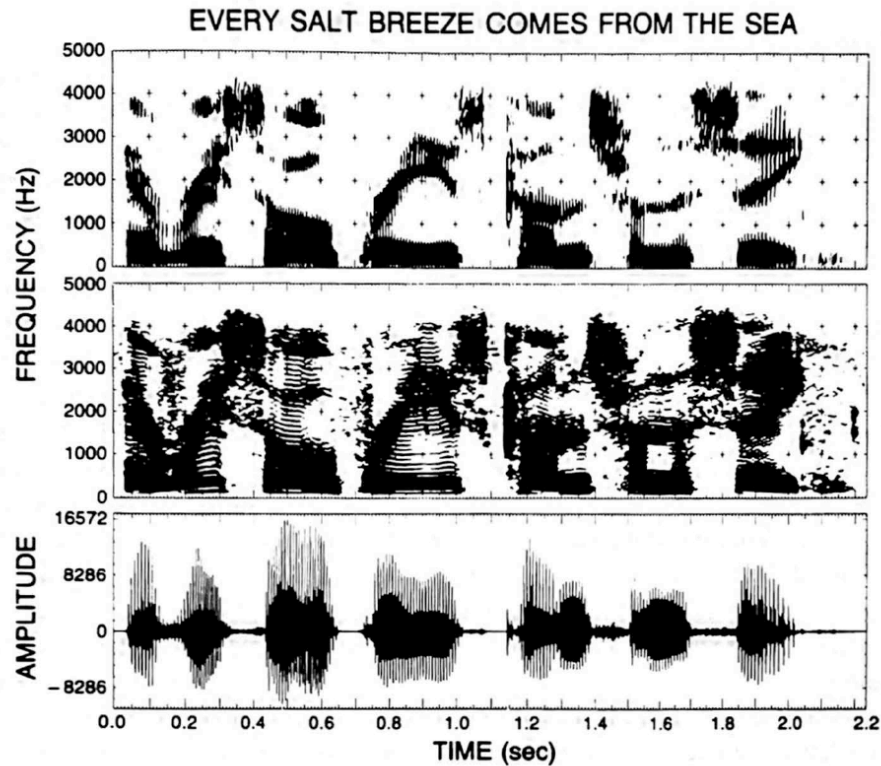


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance “Every salt breeze comes from the sea.”

Rabiner, 50 (“Figure 2.8”).

33. One speech recognition technique known well before the priority date of the '140 Patent was the statistical pattern-recognition approach to speech recognition. *Rabiner* (published in 1993) teaches the statistical pattern-recognition technique includes four main steps: feature measurement, pattern training, pattern classification, and decision logic. *Rabiner*, 82. *Rabiner* teaches the first step is “[f]eature measurement, in which a sequence of measurements is made on the input signal to define the ‘test pattern.’ For speech signals the feature measurements are usually the output of some type of spectral analysis technique, such as a filter bank

analyzer, a linear predictive coding [LPC] analysis, or a discrete Fourier transform (DFT) analysis.” *Id.* The second step, pattern training, involves creating reference patterns or models that characterize “the statistics of the features of the reference pattern.” *Id.* During the pattern classification step, an “unknown test pattern is compared with each (sound) class reference pattern and a measure of similarity (distance) between the test pattern and each reference pattern is computed.” *Id.* *Rabiner* further teaches that to “compare speech patterns (which consist of a sequence of spectral vectors), we require” “a local distance measure, in which local distance is defined as the spectral ‘distance’ between two well-defined spectral vectors[.]” *Id.* The fourth step is decision logic, “in which the reference pattern similarity scores are used to decide which reference pattern (or possibly which sequence of reference patterns) best matches the unknown test pattern.” *Id.*

34. It was also known to use lexical trees to organize vocabulary before the priority date of the ’140 Patent. For example, *Young* teaches using a “lexical tree 500 that is initialized before processing begins. The lexical tree 500 represents the active vocabulary 230 based on the phonetic relationships between words in the vocabulary.” *Young* (Ex. 1015), 11:37-41. *Young* further teaches that the “lexical tree includes a root node 505 that represents new words entering the lexical tree. From the root node 505, the tree expands to a group 510 of nodes that correspond to phonemes with which words start.” *Young*, 11:41-44. *Young* teaches that “[e]ach

node in the group 510 represents a phoneme that appears at the beginning of one or more words” using an example where “a node 610 corresponds to all words in the vocabulary that start with the phoneme ‘H’. Together, the nodes in the group 510 include representations of the starting phoneme of every word in the vocabulary.”

Young, 11:47-53. Figs. 5 and 6 of *Young* are depicted below.

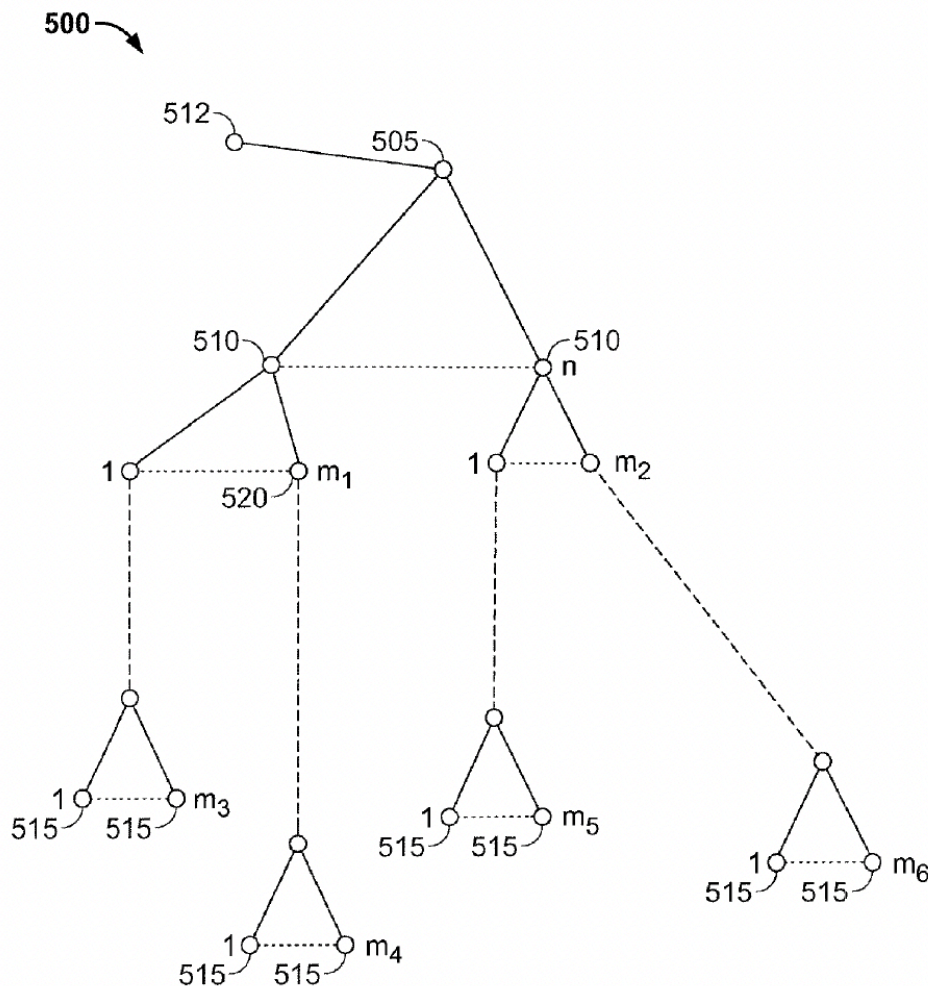


FIG. 5
(Prior Art)

Young, Fig. 5.

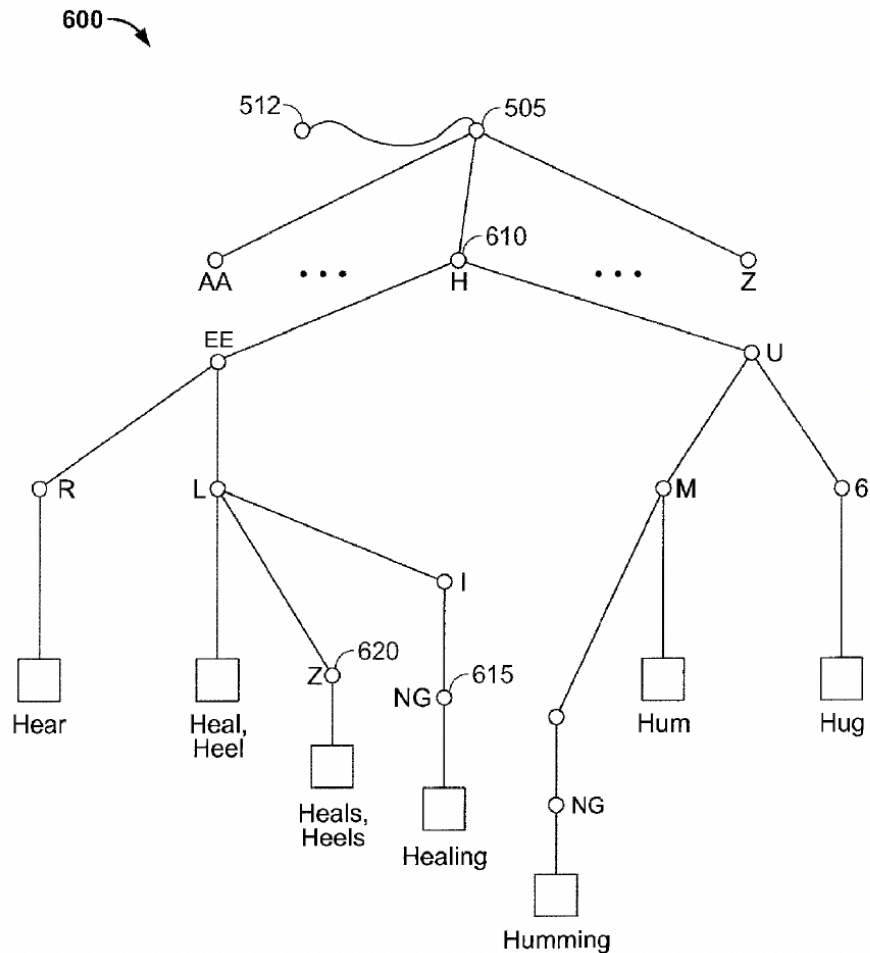


FIG. 6
(Prior Art)

Young, Fig. 6.

B. Feature Vectors

35. Using feature vectors in speech recognition systems was well known before the priority date of the '140 Patent. For example, *Holmes* teaches that in “calculating a distance between two words it is usual to derive a short-term distance that is local to corresponding parts of the words, and to integrate this distance over

the entire word duration. Parameters representing the acoustic signal must be derived over some span of time, during which the properties are assumed not to change much.” *Holmes* (Ex. 1016), 13. *Holmes* further teaches that in “one such span of time the measurements can be stored as a set of numbers, or **feature vector**, which may be regarded as representing a point in multi-dimensional space. The properties of a whole word can then be described as a succession of feature vectors (often referred to as **frames**), each representing a time slice[.]” *Id.*

36. In general, feature vectors are representations of speech where each dimension of the vector corresponds to “some prescribed acoustic feature.” *Nadas* (Ex. 1017), 1:15-23 (reproduced below). For example, *Nadas* teaches an n-dimensional space where dimensions correspond to acoustic features.

In speech processing, speech is generally represented by an n-dimensional space in which each dimension corresponds to some prescribed acoustic feature. For example, each component may represent a amplitude of energy in a respective frequency band. For a given time interval of speech, each component will have a respective amplitude. **Taken together, the n amplitudes for the given time interval represent an n-component vector in the n-dimensional space.**

Id. (emphasis added). Additionally, *Nadas* teaches that “[w]hen an unknown speech input is uttered, for each time interval, a value is measured or computed for each of the n components, where each component is referred to as a feature. The values of all of the features are consolidated to form an n-component feature vector for a time

interval.” *Nadas*, 1:32-37. *Nadas* teaches that “the components of the feature vectors (and prototype vectors)” may “correspond to well-known (1) Cepstral coefficients, (2) linear predictive coding coefficients, or (3) frequency band-related characteristics.” *Nadas*, 10:15-19. *Nadas* further teaches that a “speech type of event, such as a word or a phoneme, is characterized by a sequence of feature vectors in the time period over which the speech event was produced.” *Nadas*, 1:50-52. Further, *Jelinek* teaches that the “current continuous speech recognizers use as acoustic processor outputs vectors of cepstral coefficients [3].” *Jelinek* (Ex. 1011), 61.

37. *Maes* teaches using a well-understood acoustic front-end to extract feature vectors. *Maes* (Ex. 1018), 5:6-10 (“[A]n acoustic front-end 110 which is, itself, well-understood in the art, is used to sample utterances in an overlapping fashion and to extract feature vectors 120 therefrom, preferably as MEL cepstra, delta and delta-delta coefficients along with normalized log-energies.”). In another example, *Aoyama* teaches a feature extraction unit that analyzes audio input and extracts information to generate a feature vector. In this example, *Aoyama* teaches performing a Mel Frequency¹ Cepstrum Coefficient (MFCC) analysis and extracting

¹ I note that the Mel scale is a frequency representation which approximates the frequency response of the human auditory system, i.e., how humans hear frequency.

a feature vector utilizing a linear predictive coefficient (LPC) or Cepstrum coefficient. *Aoyama* (Ex. 1019), [0138]-[0139], [0150]-[0152], [0156]-[0160].

38. *Stubley* teaches converting speech samples into feature vectors is advantageous because feature vectors are more compact and provide better speech recognition performance than the speech samples themselves.

The purpose of the signal processing component is to convert the utterance, which is a series of samples of speech, into a more compact series of feature vectors. Besides being more compact, the feature vectors provide better speech recognition performance since they tend to remove many speaker-dependent effects and improve the robustness of the system to noise.

Stubley (Ex. 1020) at 11:30-36.

39. *Lee* teaches input speech is analyzed and represented by feature vectors that correspond to phonemes that are characterized by hidden Markov models (HMMs).

[I]nput speech is analyzed in signal frames, represented by feature vectors corresponding to phonemes making up individual words. The phonemes are characterized by hidden Markov models (HMMs), and a Viterbi algorithm is used to identify a sequence of HMMs which best matches, in a maximum likelihood sense, a respective concatenation of phonemes corresponding to an unknown, spoken utterance.

Lee (Ex. 1021) at 1:24-31 (emphasis added).

C. Acoustic Models

40. It was also known before the priority date of the '140 Patent to use acoustic models in speech recognition systems. For example, *Jelinek* teaches “acoustic modeling” includes “a statistical *acoustic model* of the speaker’s interaction with the acoustic processor.” *Jelinek*, 21. *Jelinek* teaches one type of acoustic model is a “*phonetic acoustic model*” that “is based on an intuitive linguistic concept.” *Jelinek*, 30. For example, *Jelinek* teaches creating a phonetic dictionary for vocabulary and creating a correspondence between words and sequences of “symbols from a predetermined phonetic alphabet[.]” *Id.*

41. In another example, *Schmandt* teaches using a set of templates or models as a representation of speech used by the speech recognizer.

A set of **templates** or **models**, which are descriptions of each word to be recognized, in the representation of speech used by the recognizer. The templates describe the words in the recognizer’s vocabulary, i.e., those words that the recognizer can identify. They are reference models against which an input can be compared to determine what was spoken.

Schmandt at 46.

42. A further example of using acoustic models is taught by *Stubley*. *Stubley* teaches “comparing a series of observations representing unknown speech, to stored models representing known speech[.]” *Stubley*, 1:5-10, 1:60-62. *Stubley* teaches that “the models comprise representations of elements of speech[.]” *Stubley*

at 2:51-52. Further, *Stublely* teaches in one example “the models comprise groups of representations of phonemes. This enables the models to be stored and used more efficiently.” *Stublely*, 2:40-42.

43. In another example, *Ittycheriah* teaches storing a number of reference models in a memory accessible to a processor. *Ittycheriah* (Ex. 1022) 3:34-35 (“A reference model store 112 is connected to the model construction module 110 for storing the L reference models[.]”), *Ittycheriah*, 4:30-32 (“Further, the reference model store 112 and the rank/distance store 120 may be electronic computer read/write memory or any other suitable memory.”).

44. In a further example, *Rabiner 89* teaches characterizing signals in terms of signal models. *Rabiner 89* teaches a “signal model can provide the basis for a theoretical description of a signal processing system which can be used to process the signal so as to provide a desired output.” *Rabiner 89*, 1 (Ex. 1029). Further, *Rabiner 89* teaches that such signal models include “statistical models in which one tries to characterize only the statistical properties of the signal[.]” with one example being a “hidden Markov model (HMM)” *Id.* Thus, one particular type of acoustic model is a Hidden Markov Model, which I discuss in detail below.

D. Hidden Markov Models

45. Before the priority date of the '140 Patent, Hidden Markov Models were used as a technique to “make pattern matching more powerful for dealing with

variations in speech patterns.” *Schmandt* at 54. Indeed, *Jelinek* teaches that the “usual acoustic model employed in speech recognizers” is “the *hidden Markov model*[.]” *Jelinek*, 21. In general, a Hidden Markov Model (HMM) is “a two-stage probabilistic process that can be used as a powerful representation of speech. A Hidden Markov Model is a well-behaved mathematical construct[.]” *Schmandt*, 57. Further, it was known that HMMs consisted of states. “An HMM consists of a number of internal states; the model passes from an initial state to the final state as a step-by-step process generating an observable output at each step (state transition). For example, the states may correspond to phonemes in a word with the observable output corresponding to the presence or absence of a number of acoustic features.” *Schmandt* at 57-58. Additionally, “[a]t each step, the model can either move to a new state or stay in the current one. The model is ‘hidden’ in that we cannot observe the state directly but only its output. From the series of observable outputs, we can attempt to guess when the model was in each state.” *Schmandt*, 58. “Alternatively, we can say whether some series of outputs was likely to have been generated by a particular HMM.” *Id.*

46. *Schmandt* teaches a “pedagogic example” of a “box-and-bowl setup” to explain Hidden Markov Models.

For example, consider the arrangement shown in Figure 7.10, in which a box and a bowl are each full of black balls and white balls. The box has many more black balls than white while white balls dominate the bowl. Starting with

the box, we remove one ball from it at random and pass it to an observer in another room who cannot see what is being done. Then a normal six-sided die is tossed. If the result is less than four, then the source of the next ball is the bowl. If the die is greater than three, then the next ball will be selected from the box. The cycle is then repeated. Whenever we select from the box, a die throw of less than four shifts attention to the bowl. Once we start selecting from the bowl, we continue selecting from it unless the die shows a one, at which point we are finished. This model is likely to spend a majority of its cycles selecting from the bowl since a state transition from the box has a probability of $1/2$ of shifting to the bowl, but a bowl transition has a probability of only $1/6$ of terminating and $5/6$ of continuing with the bowl. We should expect this arrangement to initially present more black balls than white to the observer (while it is in the box state) and then to produce more white balls. We should also expect more white balls overall. But keep in mind that this is a probabilistic process: Not all black balls come from the box and not all selections come from the bowl.

Id. Schmandt further teaches that the “box-and-bowl setup is a Hidden Markov Model; it is characterized by a number of *states* and associated probabilities for each *transition* from any state. The box and bowl are each a state; the rules about throwing the die define the transition properties.” Schmandt, 58-59. Schmandt also teaches that “[p]robability is expressed as a number between zero and one. If an event occurs with probability $1/2$, we would expect it to happen once in two trials. A probability of $1/6$ implies that we can expect the event once out of six trials. A smaller probability indicates that an event is less likely to occur.” Schmandt, 58.

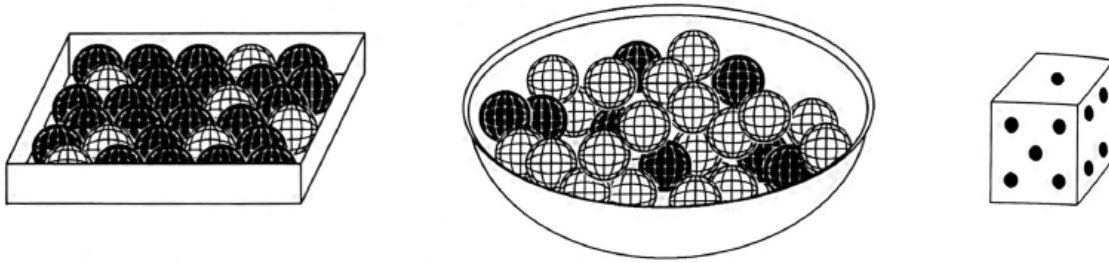


Figure 7.10. A box and a bowl full of colored balls illustrate a Hidden Markov Model.

Schmandt, 59 (“Figure 7.10”).

47. Further, *Schmandt* teaches a “more general HMM” as depicted in Figure 7.11 “A Markov Model” below. *Schmandt* teaches that “in each state $S(n)$ there is some probability $P(n,n)$ of remaining in that state and some probability $P(n, n+1)$ of transitioning to the next state.” *Schmandt*, 59. Additionally, *Schmandt* discloses that “[i]n some models it may be possible to bypass a state or a series of states as indicated by the dashed arc labeled P_{13} in the figure. The sum of the probabilities of all the arcs leaving a state is one; the model accounts for every possible path through its states.” *Schmandt*, 59.

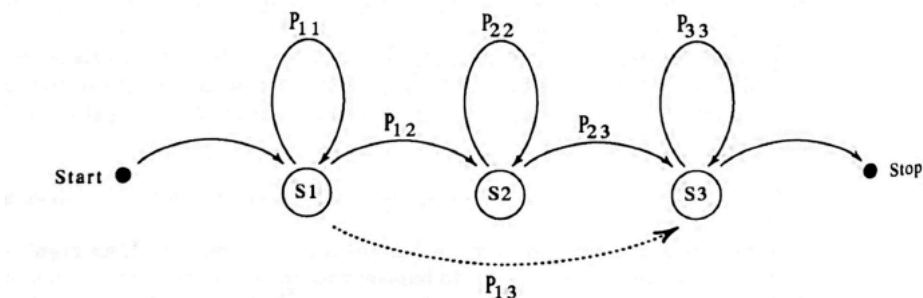


Figure 7.11. A Markov Model.

Schmandt, 59 (“Figure 7.11”).

48. *Schmandt* teaches that the example of the box-and-bowl setup and the more general HMM model apply to speech recognition because an HMM's internal states can represent acoustic segments.

An HMM can be used to represent a word with internal states representing characteristic acoustic segments, possibly phonemes or allophones. The output of a state is a frame or vector of acoustic parameters or features; this output is probabilistic to allow for variability in pronunciation and hence differences in the acoustic representation. The duration of an acoustic segment is a function of the number of steps in which the model is in the state corresponding to the segment. Staying in the same state, i.e. lengthening a phone, depends on the probability associated with the transition from that state to itself (P_{11} in Figure 7.11). Arcs such as P_{13} may be included to indicate that an intermediate state **S2** is optional during pronunciation, such as the second syllable in 'chocolate.'

In terms of the basic recognizer components, the 'templates' consist of a set of HMMs with one HMM associated with every word. The acoustic representation of an input specimen of speech is *not* a collection of states but rather a set of acoustic parameters corresponding to the possible observed sequences from the template HMMs. For the HMM-based recognizer, classification consists of determining which HMM has the highest probability of producing the observed acoustic sequence. The Viterbi algorithm is a commonly used method of solving this classification problem[.]

Schmandt, 59-60 (emphasis added).

49. *Schmandt* further teaches that "Hidden Markov Models are a powerful representation of speech and consequently are used in many speech recognition systems currently under development." *Schmandt*, 60. *Schmandt* also teaches that

HMMs can be used to “represent not only the phones of each word but also the probabilities of transitioning from one word to another.” *Id.*

50. *Rabiner* teaches that an observation vector in an HMM can be “LPC-derived weighted cepstral coefficients and weighted cepstral derivatives[.]” *Rabiner*, 231. For example, *Rabiner* teaches that the observation vector in the HMM can be a “38-component vector consisting of 12 cepstral coefficients, 12 delta cepstrum coefficients, 12 delta-delta cepstral coefficients, delta log energy, and delta-delta log energy.” *Rabiner*, 280; *see also Rabiner* at 301-302 (teaching spectral vectors with 38 components).

51. Another example of using an HMM for speech recognition well before the priority date of the '140 Patent is taught by *Nadas*. *Nadas* (issued in 1990) teaches an acoustic coprocessor 200 for speech recognition and “contemplates a probabilistic embodiment in which each prototype—rather than identifying a vector—corresponds to a probabilistic finite state machine PFSM (or Markov model).” *Nadas* (Ex. 1017), 9:68-10:4.

52. In another example, *Aoyama* teaches using an HMM as the acoustic model for speech recognition and said HMM model is stored in an acoustic model memory unit.

The matching unit 92, referring to the acoustic model memory unit 93, the dictionary memory unit 94 and the grammar memory unit 95 in utilizing the feature vector from the feature extraction unit 91 as occasion

demands, speech recognizes the voice (input speech) entered into the microphone 51 based on such as the Hidden Markov model (HMM) law.

More specifically, **the acoustic model memory unit 93 memorizes acoustic model** (e.g., including the standard pattern to be used in DP (Dynamic Programming) matching, other than HMM) showing the acoustic feature on the sub-words such as phoneme, syllable, and phoneme series in the audio language for identifying the speech. **Here, since the speech recognition is conducted based on the Hidden Markov Model law, the HMM will be used as the acoustic model.**

Aoyama at [0139]-[0140] (emphasis added); *see also Aoyama* at [0150]-[0151].

53. In yet another example, *Stubley* teaches using HMMs as acoustic models.

Advantageously, the models are represented as finite state machines with probability distribution functions attached. Continuously varying signals can be modeled as a series of stationary states, with a probability function at each state, indicating whether the signal moves to the next state or stays at the given state, after a given time interval. An example is the Hidden Markov Model (HMM). This enables the models to contain a representation of patterns with varying rates of change, yet a common underlying pattern.

Stubley, 2:29-37; *see also Stubley*, 7:23-35, 12:29-44 (“**The modelling paradigm used here is to construct models using hidden Markov models (HMMs)[.]**”) (emphasis added).

54. In another example, *Lee* teaches analyzing input speech represented by feature vectors that correspond to phonemes of words. *Lee* further teaches characterizing phonemes by Hidden Markov Models.

[I]nput speech is analyzed in signal frames, represented by feature vectors corresponding to phonemes making up individual words. The phonemes are characterized by hidden Markov models (HMMs), and a Viterbi algorithm is used to identify a sequence of HMMs which best matches, in a maximum likelihood sense, a respective concatenation of phonemes corresponding to an unknown, spoken utterance.

It is well known that each HMM comprises model parameters, e.g., mixture components which are characterized by Gaussian distributions. **In a learning phase in a speech recognition system, the HMMs are adapted to input speech by a user to adjust to the particular speech characteristics of the user, thereby increasing accuracy of the speech recognition.**

Lee, 1:24-38 (emphasis added).

55. In yet another example, *Toyoda* teaches a voice recognition module (sensor signal processor 31) that uses an HMM. *Toyoda* (Ex. 1023), 7:48-50. Furthermore, *Toyoda* teaches referencing an acoustic model memory that stores an acoustic model representing phonemes while a matching unit 42 recognizes voice input.

Using the feature parameters from the feature parameter extractor 41, **the matching unit 42 recognizes the voice input to the microphone 9 based on the continuous HMM model while referencing an acoustic model**

memory 43, a dictionary memory 44, and a grammar memory 45 as necessary.

The acoustic model memory 43 stores an acoustic model that represents an acoustic feature such as phonemes and syllables in a voice to be recognized. Since voice recognition is here carried out using the continuous HMM, an HMM is employed.

Toyoda, 7:61-8:5 (emphasis added).

E. Prior Art Speech Recognition Systems

56. Numerous speech recognition systems were well known before the priority date of the '140 Patent. For example, IBM's Tangora classifies acoustic input that "is then matched to words using discrete Hidden Markov Models of phoneme realization. A linguistic decoder then classifies the output of the acoustic front end producing a word for output." *Schmandt*, 64.

57. Another well-known speech recognition system were CMU's Sphinx and Sphinx II. Sphinx "operates in nearly real time using hardware multiprocessor peripherals to aid in the search process." *Id.* Further, "Sphinx makes extensive use of Hidden Markov Models: each phone is represented by an HMM, each word is a network of phones, and the language is a network of words." *Schmandt*, 64-65.

Further, *Ravishankar*² describes the Sphinx-II speech recognition circuit as including HMMs as “a set of states connected by transitions” expressly characterized as an “acoustic model,” explaining that HMMs were used as acoustic models in t prior to the ’140 Patent. *Ravishankar*, 12-13, 22, 24 (“Acoustic modelling in Sphinx-II is based on hidden Markov models...”). *Ravishankar* teaches that “organizing the HMMs to be searched as a phonetic tree instead of the flat structure of independent linear HMM sequences for each word is probably the most often cited improvement in search techniques.” *Ravishankar*, 17.

58. Yet another well-known speech recognition system was MIT’s SUMMIT. SUMMIT is a “phoneme-based connected speech recognizer[.]” *Schmandt*, 65. Transformed speech is segmented by “finding acoustic landmarks in the signal. Regions between the landmarks are grouped into segments on the basis of similarity. A set of phonetic classifiers assigns probabilistic phonetic labels to each of these portions of the speech signal.” *Id.* Further, “[w]ords are represented by

² I have reviewed the Declaration of June Ann Munford (Ex. 1033 (“*Munford*”)). *Munford* attests to *Ravishankar*’s public availability date and that Ex. 1007 (a scan done by June Ann Munford of a physical copy of *Ravishankar* from Carnegie Mellon University) is a true and correct copy of *Ravishankar*. *Munford* (Ex. 1033) at 1-15.

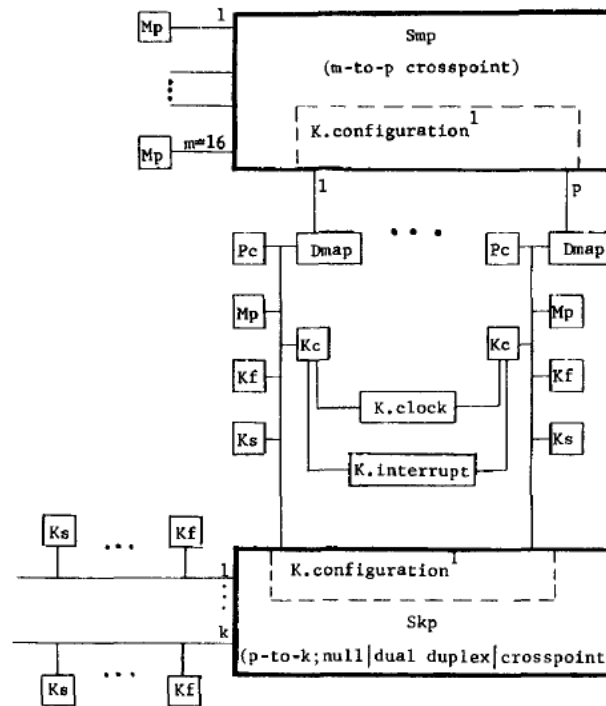
a network of phonemes; several networks may be used to encode alternate pronunciations. A pattern matching algorithm similar to dynamic programming matches the phoneme labels of the input against the word networks, deducing word boundaries in the process.” *Id.*

59. Many speech recognition systems prior to the '140 Patent recognized the advantages of using a computing platform having clusters of processors, employing clustered architectures for power, flexibility, scalability, and cost management. For example, *Hon* describes the MARS system, a speech recognition circuit having hardware organized in three levels: System, Cluster, and Processing Element. *Hon*, 24-25 (Ex. 1006). Each of the processing elements was a “16-bit custom designed microprogrammed VLSI processor” that communicated via a crossbar switch to form a pipeline. *Hon*, 24. Fifteen PEs connected via a crossbar formed a “cluster,” with the overall MARS system having up to 256 clusters. *Id.* *Hon* articulates how this clustered architecture provided advantages over other speech recognition computing platforms, stating “In contrast to the SRI-Berkeley’s search machine, MARS is scalable because it can be expanded by adding more clusters.” *Id.*

60. Other examples of speech recognition circuits employing clusters of processors can be found throughout the prior art. *Wulf* teaches a multi-processor computer system developed at CMU dubbed C.mmp using PDP-11 minicomputer

processors to recognize speech. *Wulf*, 1 (Ex. 1024). As seen in Figure 1 below,

C.mmp's included a plurality of processors Pc clustered to shared memory Mp:



where: Pc/central processor; Mp/primary memory; T/terminals;
 Ks/slow device control (e.g., for Teletype);
 Kf/fast device control (e.g., for disk);
 Kc/control for clock, timer, interprocessor communication

¹ Both switches have static configuration control by manual and program control

Figure 1—Proposed CMU multiminiprocessor computer/C.mmp

Wulf, 3-4, Fig. 1. *Erman* describes both the C.mmp system explained in *Wulf* and the HEARSAY speech recognition system that used a multiprocessor system to parallelize speech recognition processes using 16 processors clustered into two groups of 8 processors each administered by a single scheduler. *Erman*, 1-2, 5 (Ex. 1025). U.S. Patent No. 6,393,481 to Deo et al. teaches clusters of 4 general purpose

processors sharing access to data storage that may be used for speech recognition.

Deo, 9:43-58 (Ex. 1026). U.S. Patent No. 6,615,338 to Tremblay et al. teaches a clustered architecture in a VLIW processor including three DSPs operating as media functional units sharing memory to add, multiply-add, shift, compare, and perform other functions for a general function unit. *Tremblay*, Abstract, 6:51-67, Fig. 3 (Ex. 1027). *Tremblay* expressly envisions its clustered architecture being used for speech processing. *Tremblay*, 11:10-21, Fig. 3. U.S. Patent No. 5,922,076 to Garde teaches a digital signal processor designed to optimize performance for speech recognition using a cluster of four, six, or eight processors interconnected by an external bus to memory. *Garde*, 1:10-14, 1:66-2:6, 4:10-21, Fig. 1 (Ex. 1028). In summary, the '140 Patent did not invent the concept of clustered processors, nor did it originally conceive of the concept of employing clustered architectures for speech recognition. A POSITA at the time of the '140 Patent would have been familiar with systems using clustered processors for speech recognition and the advantages gained by utilizing a clustered processor architecture.

VI. OVERVIEW OF THE '140 PATENT

61. I have reviewed the '140 Patent (Ex. 1001). Descriptions of the '140 Patent that may be pertinent to my analysis are provided herein, and my descriptions of the '140 Patent are intended only as a summary of the '140 Patent and/or intended as being pertinent to the Challenged Claims of the '140 Patent. My opinions

provided herein are not intended as a discussion of my entire understanding of the '140 Patent or any other claims in any related patent to the '140 Patent. The '140 Patent describes a speech recognition system comprising clustered processors to perform speech recognition using an acoustic model stored in memory. *'140 Patent*, 1:18-20, 5:6-29, Fig. 2. The claimed circuit generates an initial score for an audio sample, used to determine which possible word paths in a lexical tree should be “pruned” from its search. *'140 Patent*, 6:21-35, 8:59-9:4.

VII. OVERVIEW OF THE PRIOR ART

A. Overview of Jiang

62. *Jiang* teaches a speech recognition system having a plurality of acoustic models that performs speech processing, acoustic model scoring, and word identification. *Jiang* at Abstract, 4:8-27, 4:51-5:3, 6:46-7:10, 7:30-8:51. *Jiang* teaches a speech recognition circuit utilizing an HMM acoustic model to compute probabilities and generate scores for searching a lexical tree. *Compare Jiang* (Ex. 1004), 2:1-15, 6:18-28, 7:29-54, 8:16-64 with *'140 Patent*, 1:30-44, 6:41-48, 8:64–9:29. Thus, in my opinion, *Jiang* is in the same field of endeavor as the '140 Patent and reasonably pertinent to a problem faced by the inventors of the '140 Patent.

B. Overview of Chen

63. *Chen* teaches an arrangement of processors into clusters, wherein each of the processors in the cluster has access to a common shared memory. *Chen*, 5:9-17, 9:5-19, Fig. 4. Like the '140 Patent, *Chen's* parallel processing computing

platform utilizes clustered processors to reduce improve computing power and performance. *Compare Chen* (Ex. 1005), 3:17-27, 5:9-17, 9:5-43, Fig. 4 with '140 Patent, 3:13-24, 5:6-41, Fig. 2. Thus, in my opinion, *Chen* is in the same field of endeavor as the '140 Patent and reasonably pertinent to a problem faced by the inventors of the '140 Patent.

C. Overview of Lucke

64. *Lucke* teaches a speech recognition circuit performing a pruning process for searching a lexical tree in which branches with an acoustic score lower than a threshold are “discarded” to reduce processing requirements. *Lucke*, 0027-0029, 0087. *Lucke*’s speech recognition circuit utilizes an HMM acoustic model to search its lexical tree. *Compare Lucke*, 0002-0005, 0008, 0025, 0079 (Ex. 1008) with '140 Patent, 1:30-44, 2:1-15, 6:41-48, 7:29-54. Thus, in my opinion, *Lucke* is in the same field of endeavor as the '140 Patent and reasonably pertinent to a problem faced by the inventors of the '140 Patent.

D. Overview of Robinson

65. *Robinson* teaches a speech recognition circuit that compares sampled speech with an HMM acoustic model, computing probabilities of matches. *Robinson*, 1:5-9, 8:28-67, 9:52-57. *Robinson*’s circuit computes feature vectors for comparison to a speech model. *Compare Robinson*, 8:28-55, 9:4-16, with '140 Patent, 4:52-67, 5:6-26. Thus, in my opinion, *Robinson* is in the same field of

endeavor as the '140 Patent and reasonably pertinent to a problem faced by the inventors of the '140 Patent.

E. Overview of Wrench

66. *Wrench* teaches a multiprocessor speech recognition circuit that includes an “initialization” program module for its controller that “performs the various tasks for **initializing** the [speech recognition subsystem] parameters.” *Wrench*, 6:40-50, 2:44-58. *Wrench*’s circuit provides a plurality of processors performing speech recognition administrated by the single master controller. *Compare Wrench*, 4:40-58, 6:38-7:7, Fig. 2 with *'140 Patent*, 5:46-52, 10:2-30, Fig. 2. Thus, in my opinion, *Wrench* is in the same field of endeavor as the '140 Patent and reasonably pertinent to a problem faced by the inventors of the '140 Patent.

VIII. OPINIONS REGARDING GROUND 1

A. Motivation to Combine *Jiang* and *Chen*

67. In my opinion, a POSITA would have been motivated and found it obvious to implement the speech recognition techniques taught by *Jiang* using the clustered processor architecture taught by *Chen*. As discussed above in my Background of the Technology, speech recognition circuits providing processors in a cluster with shared memory were extremely well-known in the art prior to the '140 Patent. As will be further discussed below regarding limitation 1(a), competing interests of increased computing power (allowing a more relaxed pruning threshold and thus fewer recognition errors while maintaining high speed and efficiency) with

the overall cost of speech recognition computing platforms would have motivated a POSITA to group a plurality of simple, commercially available processors into a cluster for recognizing unknown speech as found in *Chen* and numerous other prior art references including *Hon*, *Tremblay*, *Erman*, *Wulf*, *Garde*, and *Deo*. Specifically, a POSITA would have been aware of the need for maintaining high speed in a speech recognition circuit (to respond in a timely manner, emulating human speech interaction, rather than lagging progressively further behind), while also realizing that solitary, single-core processors sophisticated enough (for the time) to provide the necessary computing power would incur a high financial cost, leading the POSITA to seek less expensive speech recognition circuitry solutions to employ an HMM acoustic model to generate scores, as taught by *Jiang*. Combining *Jiang* and *Chen* would have required use of a known technique (e.g., *Chen*'s clustering of processors with memory) to improve a similar device (*Jiang*'s multiprocessor speech recognition circuit) in the same way, enabled *Jiang*'s speech recognition circuit to efficiently process branches of the lexical tree in parallel without significantly increasing its cost. Each of *Chen* and *Jiang* each contemplate hardware architectures with a plurality of signal processors; therefore, a POSITA, without undue experimentation would have been capable of simply substituting the more efficient clustered processing architecture of *Chen* for *Jiang*'s processing architecture to achieve a compatible improvement in processing speed and power.

B. Motivation to Combine *Chen* and *Jiang*

68. For similar reasons as those discussed for *Jiang* in view of *Chen*, a POSITA would have found it obvious, in my opinion, to configure a computing platform comprising clusters of processors as taught by *Chen* to perform the speech recognition techniques of *Jiang*. Before the '140 Patent, it was well known that calculating scores during a lexical tree search to prune branches unlikely to yield the correct word was a well-known technique to increase the efficiency of a speech recognition circuit, resulting in an increase in speed. A POSITA would have been capable of adding the hardware missing from *Chen* to utilize its computing platform for speech recognition, such as a microphone and A/D converter, widely-used, simple, and well-known elements in computer hardware at the time. Other modifications necessary, such as configuring *Chen's* circuitry to specifically recognize speech pursuant to *Jiang* would have required only software programming and well-known computing techniques and structures, and thus would have lead a POSITA to a reasonable expectation of success.

C. Obviousness of Modifying Electronic Hardware and Software to Be Circuitry

69. In my opinion, it would have been obvious to a POSITA to implement the computerized systems and methods taught by *Jiang*, *Chen*, *Lucke*, *Robinson*, and/or *Wrench* specifically using “circuitry,” as recited in the claims. A POSITA would have recognized that computers are electronic devices comprising at least one

or more computer processors, memory, a power source, and electrical connections therebetween and operating on electricity, thus teaching or at least rendering obvious “circuitry” performing the computerized steps taught by each of *Jiang*, *Chen*, *Lucke*, *Robinson*, and *Wrench*. To the extent any particular citation from one of these references describes computerized methods stored as software, a POSITA would have found it obvious to physically implement those computerized methods using electronic circuitry such as a processor core and associated memory. Thus, in my opinion, a POSITA would have found it obvious and been motivated to embody the various electronic components disclosed in the references as “circuitry” without undue experimentation.

D. Claim 1 Is Rendered Obvious By *Jiang* in View of *Chen*

1. Claim 1

a) Claim 1(Pre): “A speech recognition circuit comprising”

70. Though *Jiang* does not expressly use the word “circuit,” a POSITA would have recognized that *Jiang* discloses a circuit for performing speech recognition. *Jiang*, 1:15-18, 6:18-61, 4:51-5:3, Abstract, Figs. 1, 2, 7. *Jiang* describes in detail a number of computer-implemented structures and hardware components including a microphone, A/D converter, RAM, ROM, input and output devices, microprocessor-based electronics, PCs, ports, and program modules that would have comprised “circuitry” as understood by a POSITA. *Jiang*, 5:1-6:45,

Figs. 1-2. Further, a POSITA would have found it obvious to implement the structures and computerized methods described by *Jiang* using “circuitry,” especially since *Jiang* expressly teaches its invention may be employed in a variety of “computer system configurations[.]” *Jiang*, 4:60-62.

b) Claim 1(a): “one or more clusters of processors, each of the one or more clusters of processors comprising: a plurality of processors;”

71. As discussed above in the Background of the Technology, speech recognition circuits comprising clusters of processors were well-known in the art prior to the time of the '140 Patent. For example, *Hon* describes the MARS (Microprogrammable Accelerator for Rapid Simulation) multiprocessor machine developed in the 1980s that utilized up to 256 clusters of processors, each comprising 15 coupled processing elements to perform natural language processing. *Hon*, 24-25. *Hon* lauds the flexibility and scalability afforded to the MARS machine by its clustered architecture, which would have motivated a POSITA to implement clustered architecture in other speech recognition circuits. *Id.* Similarly, *Erman* described a speech recognition system that used a multiprocessor system to parallelize speech recognition processes using 16 processors clustered into two groups of 8 processors each administered by a single scheduler. *Erman*, 1-2, 5. Additional examples of speech recognition circuits discussed using clustered processors can be found in *Tremblay*, *Wulf*, *Garde*, and *Deo*. Based on this support,

I can conclude that the '140 Patent did not invent the concept of clustered processors, nor did it originally conceive of the concept of employing clustered architectures for speech recognition.

72. Notably, *Jiang* itself is a “multiprocessor” speech recognition system, which a POSITA would have recognized would have benefitted from a clustered architecture for the same reasons recited by *Hon*: scalability to any desired number of clusters and flexibility to particular speech recognition applications. *Jiang* 4:51-5:3 (describing “multiprocessor” embodiments), 6:39-42 (“...CPU 21 [] may include one **or more** processors...”); *Hon*, 24-25. Other examples I discussed above, including those described in *Tremblay*, *Erman*, *Wulf*, *Garde*, and *Deo*, each show that speech recognition was well-known to be advantageously performed on computing architectures employing a clustered-processor architecture prior to the priority date of the '140 Patent. A POSITA would have recognized that *Jiang*'s multiprocessing speech recognition system would have benefitted in scalability and flexibility from using a clustered-processor architecture, as taught by *Chen* and supported in the background art, particularly *Hon*.

73. *Chen* describes a “parallel processing computer system” comprising a plurality of “clusters” of processors 100a-d, each cluster including two or more processors 102a-d “that are symmetrically connected to a cluster shared memory 104a via a connection node 106a.”

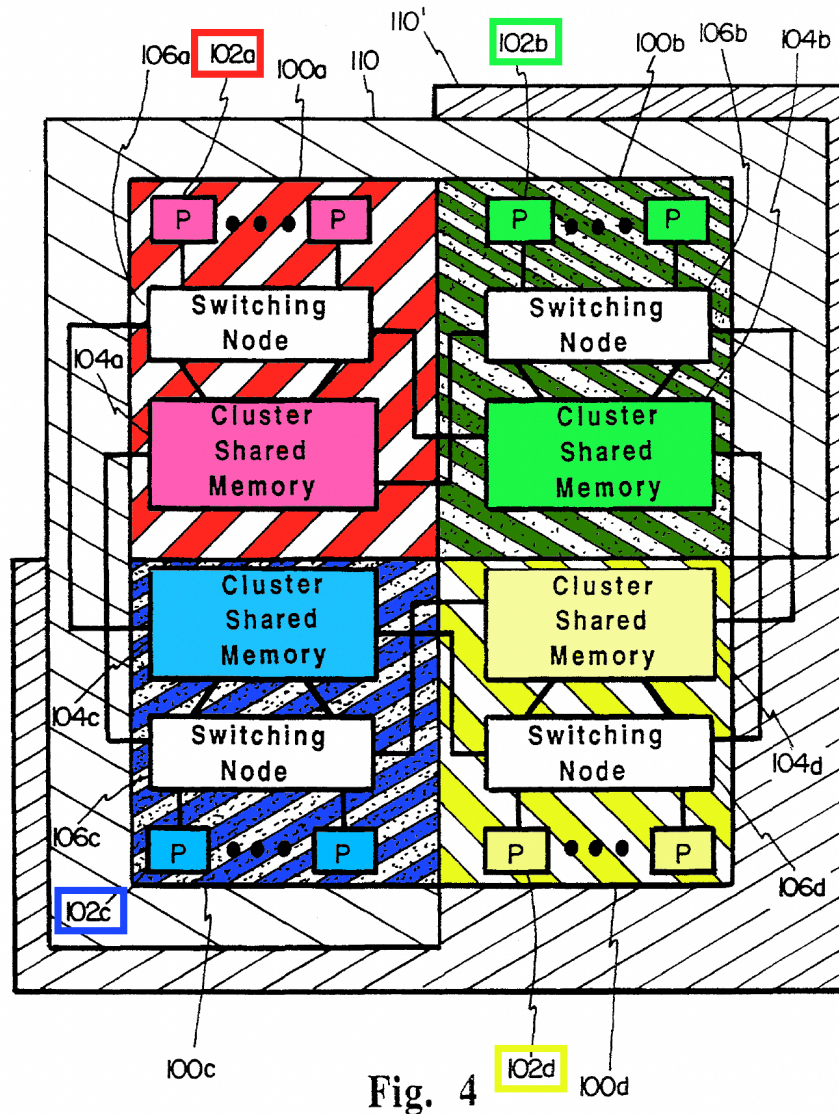


Fig. 4

Chen, Fig. 4 (annotated), 9:5-19. The clusters in this annotated Figure 4 have been color-coded to more easily indicate the separate clusters being taught. Each of the red (cluster a), green (cluster b), blue (cluster c), and yellow (cluster d) clusters include a respective plurality of processors 102a-d (i.e. “2 or more” processors) marked with the letter P and a respective cluster shared memory 104a-d. Much like *Jiang* is described as being formed of conventional, suitable processors and other

commercially available components, *Chen* is expressly described as utilizing commercially available microprocessors and memory chips. *Jiang*, 6:18-61; *Chen*, 10:14-35. *Chen* also expressly envisions that existing multiprocessor systems “might be retrofitted to take advantage of the unified parallel processing architecture of the present invention.” *Chen*, 10:21-25. For these reasons, a POSITA would have found it obvious to retrofit *Jiang*’s system of commercially available components “to take advantage of” *Chen*’s “unified parallel processing architecture.” *Jiang*, 6:18-61; *Chen*, 10:14-35.

74. In my opinion, a POSITA would have been motivated to retrofit *Jiang*’s system to utilize the clustered-processor hardware arrangement of *Chen* for improvements in processing speed and power. Processing resources (speed and power) are important in speech recognition circuits because, as explained by *Jiang*, the strictness to which a pruning threshold is set in a speech recognition system depends upon the computational resources available. *Jiang*, 2:19-46, 11:20-24. A higher (stricter) pruning threshold requires less computational resources, but results in a higher chance of recognition errors due to over pruning. That is, the “correct” word to be recognized gets pruned at some point in the recognition process when the branch on which that word is located has, temporarily, a score below the high pruning threshold, removing the correct word from consideration before its score has a chance to significantly increase and win out as the recognized word. In doing so,

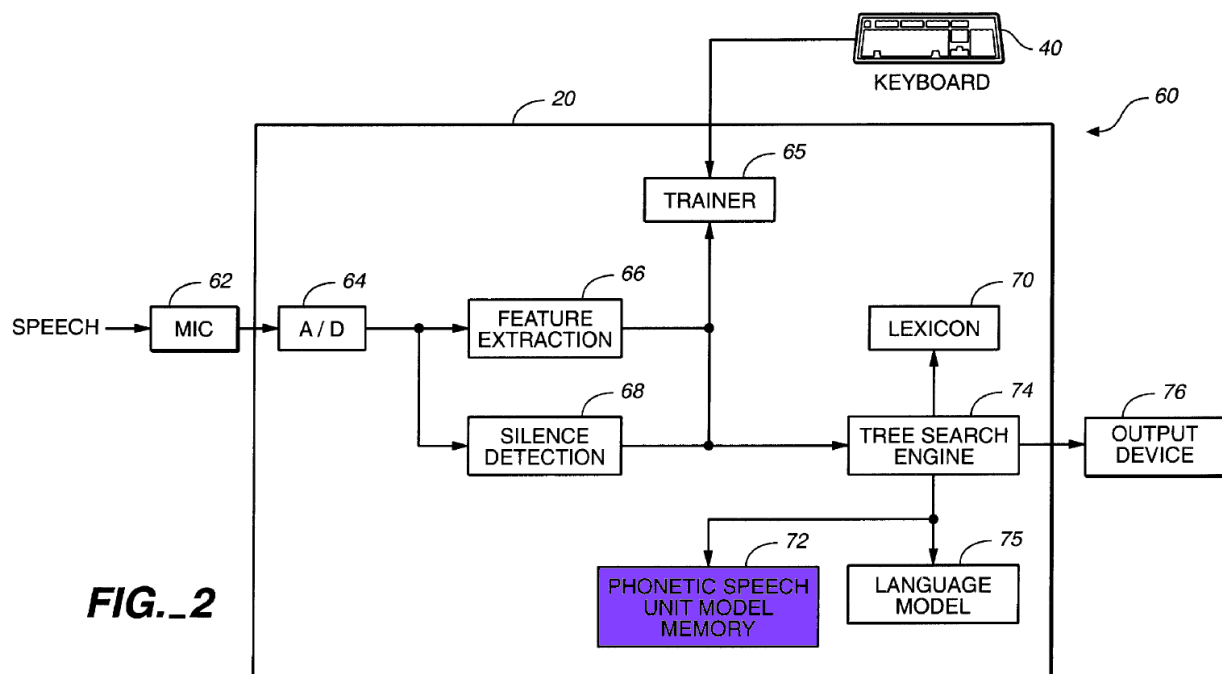
a simpler, less sophisticated, expensive, or advanced computing hardware architecture may be used for speech recognition, at the cost that, in some subset of recognition events, over pruning will cause a recognition error. A more sophisticated or advanced computing hardware architecture reduces over pruning errors by enabling a lower (more relaxed) pruning threshold, though the tradeoff is often one of expense. Very high performance processors, especially at the time of *Jiang* and *Chen* (prior to the '140 Patent) would have been expensive, especially for implementations wherein many speech recognizing machines are required, such as at a large business or telephone call center or cloud-based server farms performing speech recognition. *Chen* directly discusses the limiting role that the expense of hardware components can play in computing platforms, and contemplates reducing the cost of a computing platform as a potential goal of its invention at 2:30-33 and 4:29-40. Modifying *Jiang*'s speech recognition circuit to use clusters of processors as taught by *Chen* furthers *Jiang*'s objective of accurate speech recognition by providing a more powerful computing platform (enabling a more relaxed pruning threshold) while reducing the expense incurred by grouping commercially available, simple processors into clusters. *Jiang*, 11:16-24. A POSITA would have been motivated by *Jiang*'s description of the role of pruning thresholds and *Chen*'s express statement of retrofitting systems to utilize its "unified parallel processing architecture" to combine their teachings into a single, improved speech recognition

circuit that is both accurate and efficient. *Jiang*, 6:18-61, 2:19-46, 11:20-24; *Chen*, 10:14-35. A POSITA would have been motivated to employ the clustered processor architecture **in addition to** the pruning techniques taught by *Jiang* to result in the most powerful, efficient, and accurate system possible. A POSITA would have recognized both *Chen* and *Jiang* describe hardware architectures with multiple signal processors; therefore, *Chen's* improved clustered processing architecture would have been simply substituted for *Jiang's* to desirably result in improved processing speed and power, leading a POSITA to a reasonable expectation of success for the combination. *Jiang*, 4:51-65, 6:62-7:10; *Chen*, 4:4-24, 10:1-11.

c) Claim 1(b): “and [one or more clusters of processors, each of the one or more clusters of processors comprising:] an acoustic model memory storing acoustic model data”

75. As discussed above in the Background of the Technology, a POSITA would have understood that Hidden Markov Models were well-known “acoustic models” prior to the '140 Patent. The '140 Patent itself describes HMMs as representing acoustic models and states that the “**acoustic model** memory 23 stores a Hidden Markov Model for acoustically modelling words as lexical trees.” '140 Patent, 1:30-44, 6:41-48. Further, as discussed, *Jelinek* teaches that the “usual acoustic model employed in speech recognizers” is “the *hidden Markov model*,” while *Schmandt* teaches “Hidden Markov Models are a powerful representation of speech and consequently are used in many speech recognition systems currently

under development” and may be used to “represent not only the phones of each word but also the probabilities of transitioning from one word to another.” *Jelinek*, 21; *Schmandt*, 60. Further, *Ravishankar* describes HMMs as “a set of states connected by transitions” expressly characterized as an “acoustic model,” explaining that HMMs were used as acoustic models in the widely-known Sphinx-II speech recognition circuit prior to the ’140 Patent. *Ravishankar*, 12-13, 22, 24 (Ex. 1007) (“Acoustic modelling in Sphinx-II is based on hidden Markov models...”). Such an HMM-based phonetic model is taught by *Jiang* at least 2:53-60 and 7:30-47, and thus *Jiang* teaches an “acoustic model,” as claimed. *Jiang* is clear that its HMM acoustic model data is stored in its “model memory” 72.



Jiang, 7:29-54, 6:18-28, Fig. 2 (annotated).

76. As discussed for Claim 1(a), *Chen* teaches a speech recognition circuit comprising one or more clusters, each cluster 100a-d including two or more processors 102a-d that are symmetrically connected to a cluster shared memory 104a-d via connection nodes 106a-d:

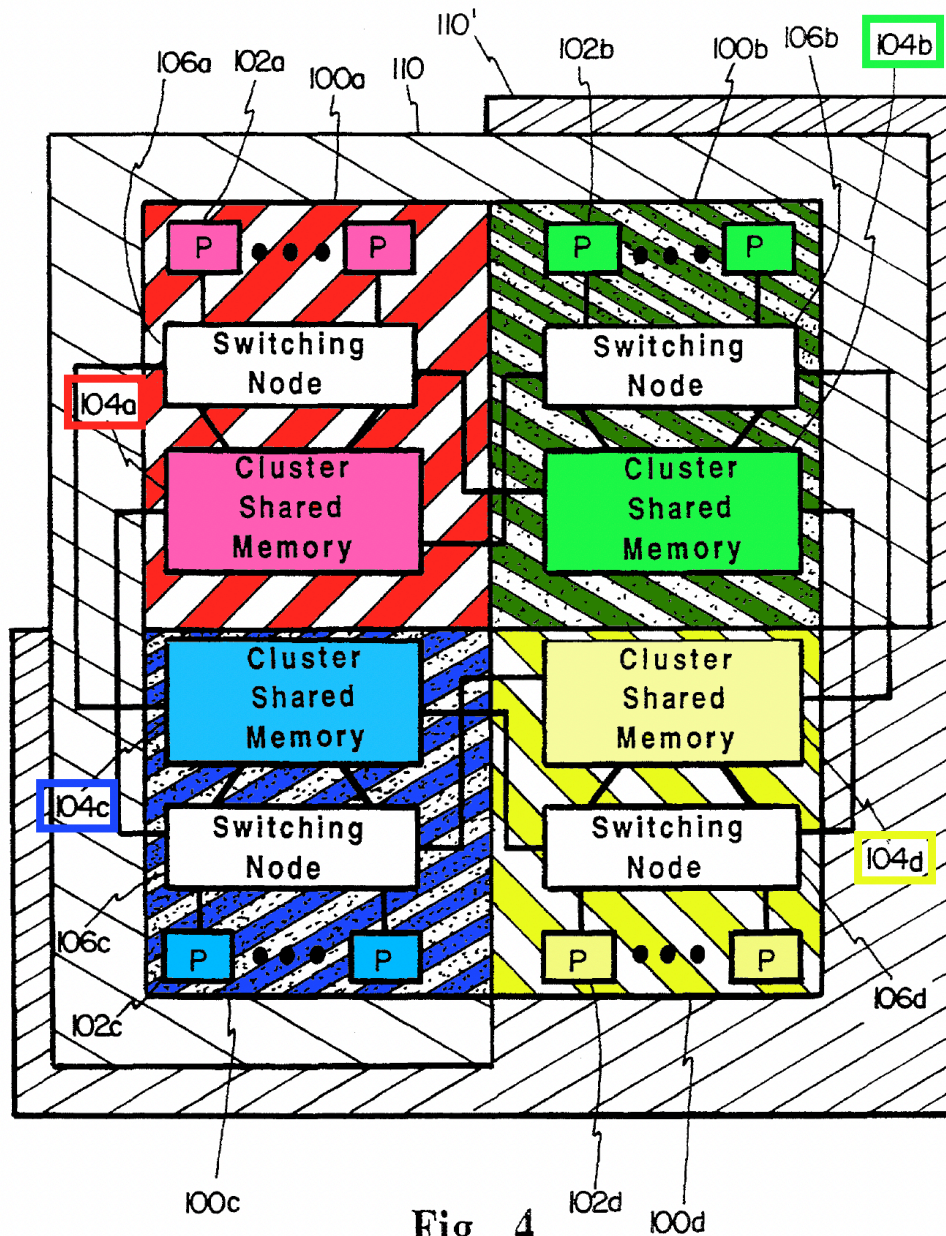


Fig. 4

Chen, Fig. 4 (annotated), 9:5-19. *Chen* characterizes this architecture as “physically separable clusters, each cluster having a common cluster shared memory that is symmetrically accessible by all of the processors in that cluster.” *Chen*, 5:9-17. Again, *Chen* explains that the clusters may be formed of commercially available components. *Chen*, 10:14-35. A POSITA would have been motivated to store at least a portion of the acoustic model data (HMM acoustic model from *Jiang*’s model memory 72) in the shared cluster memories 104a-d of each of the clusters 100a-d taught by *Chen* for access by processors 102a-d during speech recognition so that each of the processors had access to the acoustic model data necessary for performing speech recognition. My motivation to combine *Jiang* and *Chen* discussed above for limitation 1(a) is incorporated by reference for this limitation as well. Specifically, A POSITA would have been motivated to employ the clustered processor architecture **in addition** to the pruning techniques taught by *Jiang* to result in the most powerful, efficient, and accurate system possible. Such a combination would have allowed for a relaxed pruning threshold at a reduced financial cost.

d) Claim 1(c): “wherein each of the plurality of processors is configured to compute a probability using the acoustic model data in the acoustic model memory”

77. As discussed, *Jiang* teaches a speech recognition process in which circuitry is configured to compute probabilities using HMM acoustic model data in its acoustic model memory 72, modified to reside in the cluster shared memories

104a-d of each cluster of *Chen*. *Jiang* describes its part of its speech recognition process as follows:

In order to find the most likely sequence of states corresponding to a **sequence of acoustic vectors**, the Viterbi algorithm is employed. The Viterbi algorithm performs a computation which starts at the first frame and proceeds one frame at a time, in a time-synchronous manner. **A probability score is computed** for each state in the state sequences (i.e., the HMMs) being considered. Therefore, a **cumulative probability score is successively computed** for each of the possible state sequences as the Viterbi algorithm analyzes the acoustic signal frame by frame. By the end of an utterance, **the state sequence (or HMM or series of HMMs) having the highest probability score computed by the Viterbi algorithm provides the most likely state sequence for the entire utterance**. The most likely state sequence is then converted into a corresponding spoken subword unit, word, or word sequence.”

Jiang, 2:1-15. *Jiang* describes a speech recognition circuit computing a probability (expressly stated above) using the acoustic model data (HMM data) in its acoustic model memory (acoustic model memory 72, modified to cluster shared memories 104a-d of each cluster). Probabilities determined from extracted feature vectors are provided to *Jiang*’s tree search engine 74 to perform a lexical tree search. *Jiang*, 7:29-54. *Jiang*’s “tree search engine 74” is described as being implemented in CPU 21 (**which may include one or more processors**), which, as described above regarding limitation 1(a), would have been modified by a POSITA to employ the clusters of processors taught by *Chen*. *Jiang*, 6:39-42, Figs. 1-2. Thus, *Jiang*’s tree

search engine 74 (1) accesses HMM information stored in phonetic speech unit model memory 72 (*Jiang*, 7:30-36); and (2) computes a probability (*Jiang*, 2:1-15, 8:44-51).

78. In my opinion, *Jiang* teaches that speech is commonly modeled by HMMs, where each HMM comprises a plurality of states. Further, *Ravishankar* teaches the same. *Ravishankar* at 28 (“Most systems use hidden Markov models (HMMs) to represent the basic units of speech.”). For a spoken speech input having specific spectra, the spectra are converted into one or more feature vectors. *Jiang* at 1:58-67, 6:62–7:10. Each state of the HMM has an associated probability corresponding to a frame of spoken speech, where each frame may represent a plurality of feature vectors. *Jiang* at 1:21-24, 1:58-66, 6:62-7:10; *Ravishankar* at 28, 41. *Jiang* teaches the spoken utterance (i.e., speech) is encoded as one or more feature vectors reflecting the spectral characteristics of the speech. *Jiang* at 1:58-66, 6:62-7:10. To determine the entirety of the uttered speech unit (whether it be a subword unit, a word, or a sequence of words), a most likely *state sequence* is computed. In general, discussing utilizing HMMs, *Jiang* refers to a “probability score” being “computed for each state” (of the HMM), and a “cumulative probability score” successively computed, utilizing the Viterbi algorithm, for each possible state sequence. *Jiang* at 2:10-15.

79. Thus, the HMM includes a plurality of states, where *each state* includes a *probability* that the feature vector(s) are modeled accurately. *Jiang* teaches the system can be trained by providing speech input that is converted to codewords. HMMs are created by training module for each phoneme. The HMMs thus represent modeled speech associated with the phonemes. Therefore, a POSITA would have understood that *Jiang* contemplates training the system to generate HMMs that are representative of one or more phonemes. I note *Jiang* teaches “training” the speech recognition system to generate model HMMs. *Jiang* at 11:29-48. Some states will have a low probability because the HMM state is unlikely to represent the feature vector(s) representative of the spoken speech, whereas other states will have a higher probability. *Jiang* at 2:15-46 (discussing pruning state sequences that are unlikely to model the spoken speech). The higher the probability, the more likely the state is an accurate model of the feature vector(s). As the HMM is traversed, a *cumulative probability* is calculated for each successive state in the HMM. *Jiang* at 2:7-15. At the end of the utterance (and therefore, at the end of the HMM), the sequence of states (i.e., the “state sequence”) with the highest cumulative probability is the “most likely state sequence for the entire utterance.” *Jiang* at 2:10-15. This “most likely sequence” is then converted into the utterance, whether it be a spoken subword unit, word, or word sequence. *Id.*

80. A feature vector commonly corresponds to a frame of speech, where a frame of speech is commonly 10 ms in length. *Jiang* at 1:58-67, 6:62-7:1; *Ravishankar* at 28 (“Frames are typically spaced at 10msec intervals”). I note that a 10 ms length of speech is short, and therefore, most phonemes (and, by default, words) in the English language will comprise a plurality of frames. As is known, there are approximately 40-50 phonemes in the English language. *Jiang* at 3:4-6; *Ravishankar* at 42 (Section 3.1.2). Because a spoken 10 ms utterance will not usually include the entirety of a phoneme, the speech recognition process frequently divides the speech recognition process into multiple stages. First, a most likely phoneme of the approximate 40-50 phonemes is determined, and then, a most likely word formed by one or more phonemes is determined.

81. To recognize *phonemes*, the speech recognition system is trained, which creates acoustic or phonetic models, e.g., HMMs, representing the phonemes. Per *Jiang*:

One or more HMMs are created by training module 65 for each phoneme of each word in a desired vocabulary represented by the training data. The HMMs thus represent model output distributions associated with the phonemes in the vocabulary. The prefix tree is then formed, based upon the desired vocabulary, such that phonemes are structured to provide a leaf which is associated with each word in the desired vocabulary.

Jiang at 11:25-47, 7:30-8:51; *Ravishankar* at 28. Once the acoustic models, such as the HMMs, are created, the system can then use the HMMs to recognize phonemes.

These HMMs are then used in *Jiang's* prefix tree to identify words. Further, *Jiang* teaches:

Another **conventional technique** for further **reducing the magnitude of computation** required for speech recognition includes the use of a **prefix tree**. A **prefix tree** represents the lexicon of the speech recognition system as a **tree structure** wherein all of the words likely to be encountered by the system are represented in the tree structure.

In such a prefix tree, **each subword unit (such as a phoneme) is typically represented by a branch which is associated with a particular phonetic model (such as an HMM)**. The phoneme branches are connected, at nodes, to subsequent phoneme branches. All words in the lexicon which share the same first phoneme share the same first branch. All words which have the same first and second phonemes share the same first and second branches.

Jiang at 2:47-60, 7:30–8:51; *Ravishankar* at 33 (discussing that organizing HMMs as a phonetic tree (referred to as a lexical tree) to perform word searching significantly reduces the computation).

Organizing the HMMs to be searched as a phonetic tree instead of the flat structure of independent linear HMM sequences for each word is probably the most often cited improvement in search techniques in use currently. This structure is referred to as *tree-structured lexicon* or *lexical tree*. If the pronunciations of two or more words contain the same n initial phonemes, they share a single sequence of n HMM models representing that initial portion of their pronunciation. (In practice, most systems use triphones instead of just basephones, so we should really consider triphone pronunciation sequences. But the basic argument is the same.) Since the word-initial models in a non-tree structured Viterbi search are typically the majority of the

total number of active models, the reduction in computation is significant.

Ravishankar at 33. See *Ravishankar* at 67 (Section 4.2) (same), 68-69, 22 (discussing that a linear sequence of HMM models for each word in the vocabulary is “one to two orders of magnitude beyond the capabilities of modern workstations”). Thus, a subword unit, such as a phoneme, is represented as a branch in the prefix tree. Each branch is alone associated with a phonetic model, such as one or more HMMs. Therefore, the prefix tree employs a plurality of HMMs, where each branch of the prefix tree is associated with an HMM that was used to determine the particular phoneme branch.

82. *Jiang*’s speech recognition circuit is expressly described as determining the “most likely” phonemes based on stored HMMs. *Jiang*, 7:30-44. A POSITA, having read *Jiang*’s description of using the most likely phonemes determined from the HMMs is using the probability score computations based on HMMs (described at 2:1-15) “in a known manner” (*Jiang*, 8:51 – a known manner such as that described by *Ravishankar*, 33) to search a lexical tree (“prefix tree”) for determination of the speaker’s uttered word at 8:16-52. Thus, *Jiang* teaches computing a probability (i.e., the *Jiang* probability score) using the acoustic model data (i.e., the *Jiang* HMMs) in the acoustic model memory (i.e., *Jiang*’s phonetic speech unit model memory 72), teaching limitation 1(c).

83. *Chen* explains that its configuration allows “each” (or “every”) processor 102 to access the cluster shared memory and operate in parallel. *Chen*, 5:10-17, 5:33-41, 18:64-19:9, 6:4-26. Based on these disclosures, a POSITA would have recognized *Chen* teaches or at least renders obvious that each processor 102 in the clusters employed as the tree search engine 74 in the *Jiang+Chen* modified speech recognition circuit would have been configured to compute a probability using the acoustic model data in the acoustic model memory. Specifically, as I discussed for limitation 1(b), in the proposed system, each of *Chen*’s processors 102a-d in respective clusters 100a-d would have computed probability scores as taught by *Jiang* using the acoustic model data stored in shared cluster memories 104a-d of each of the clusters 100a-d. A POSITA would have recognized that it would have been a waste for any particular processor to *not* be configured to compute a probability using the acoustic model data in the acoustic model memory, and thus to maximize both computational and cost efficiency the POSITA would have configured the *Jiang+Chen* modified speech recognition circuit to compute a probability using the acoustic model data in the acoustic model memory. Again, my motivation to combine *Jiang* and *Chen* discussed above for limitation 1(a) is incorporated by reference for this limitation as well.

e) **Claim 1(d): “[wherein] the speech recognition circuit is configured to generate an initial score for an audio sample”**

84. Figures 3-6 of *Jiang* each illustrate a lexical tree. *Jiang*, 4:36-44 (providing an overview of the Figures), 8:5-15 (describing Fig. 3), 9:28-42 (describing Fig. 4), 10:34-50 (describing Fig. 5), 10:65-67 (describing Fig. 6), Figs. 3-6. Though *Jiang* uses the terminology “prefix tree,” a POSITA would have recognized that *Jiang*’s descriptions and Figures are describing a well-known lexical tree structure. *Jiang*, 2:47-52 (stating “A prefix tree **represents the lexicon** of the speech recognition system **as a tree structure** wherein all of the words likely to be encountered by the system are represented in the tree structure.”), 2:53-67 (describing in detail a well-known lexical tree structure). Lexical trees were well-known data structures in the art for organizing HMMs in a speech recognition circuit. For example, *Ravishankar* states that “Organizing [] HMMs to be searched as a phonetic tree instead of the flat structure of independent linear HMM sequences for each word is **probably the most often cited improvement in search techniques in use currently**. This structure is referred to as *tree-structured lexicon* or *lexical tree*. If the pronunciations of two or more words contain the same n initial phonemes, they share a single sequence of n HMM models representing that initial portion of their pronunciation.” *Ravishankar*, 33. Figure 3 illustrates a “prior art” lexical tree, without silence nodes. Figure 4 merely adds silence nodes to the prior art lexical tree

of Figure 3, but makes no changes otherwise. Figure 5 adds even more silence (context dependent) silence phone nodes to the lexical tree of Figure 4, but makes no further changes. Figure 6 illustrates the lexical tree of Figure 5 with particular branches (those stemming from a “T” phoneme) pruned. Given that each of these figures illustrate lexical trees and given that (other than the pruning of Figure 6) the only differences are silence nodes, it is my opinion that the Petition appropriately treats discussions from *Jiang* of the lexical trees illustrated in Figures 3-6 as a single concept, even though particular citations may be directed toward any particular lexical tree. The pruning of Figure 6 is discussed in the Petition directly, as well as below with regard to limitation 1(e).

85. Claim 1(d) recites an “audio sample,” which *Jiang* teaches. Specifically, *Jiang* teaches that a microphone provides an analog system to an A/D converter, which “samples” the signal at a rate of approximately 16kHz into “a plurality of digital samples.” *Jiang*, 6:46–7:1. In more detail, *Jiang* states:

Feature extraction module 66 divides the digital signal received from A/D converter 64 into **frames** which include **a plurality of digital samples**. Each frame is approximately 10 milliseconds in duration. The frames are then preferably encoded by feature extraction module 66 into a **feature vector** reflecting the spectral characteristics for a plurality of frequency bands. In the case of discrete and semi-continuous hidden Markov modeling, feature extraction module 66 also preferably encodes the feature vectors into one or more **codewords** using vector quantization techniques and a codebook derived from training data. Thus, feature extraction module 66 provides,

at its output the **feature vectors (or codewords)** for each spoken utterance.

Jiang, 6:62–7:10. A “silence detection module” is then applied to the samples provided by the A/D converter...to determine boundaries between words being uttered by the user.” *Jiang*, 7:16-29. A POSITA would have recognized this description from *Jiang*, in which an acoustic signal received via a microphone, “sample[d]” by an A/D converter, and divided into frames comprising a plurality of digital samples by the feature extraction module 66 is encoded into feature vectors (or later, codewords) by the feature extraction module 66, teaches the recited “audio sample[s].”

86. Turning to the File History of the ’140 Patent, I see that the Examiner issued a rejection regarding limitations 1(d) (“generate an initial score for an audio sample”) and 1(e) (“used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score”), stating that the concepts were not supported by the specification, and were thus New Matter. *’140 File History*, 89. Applicant responded to the contrary, stating that support for the limitations was given by the ’140’s description of a pruning process:

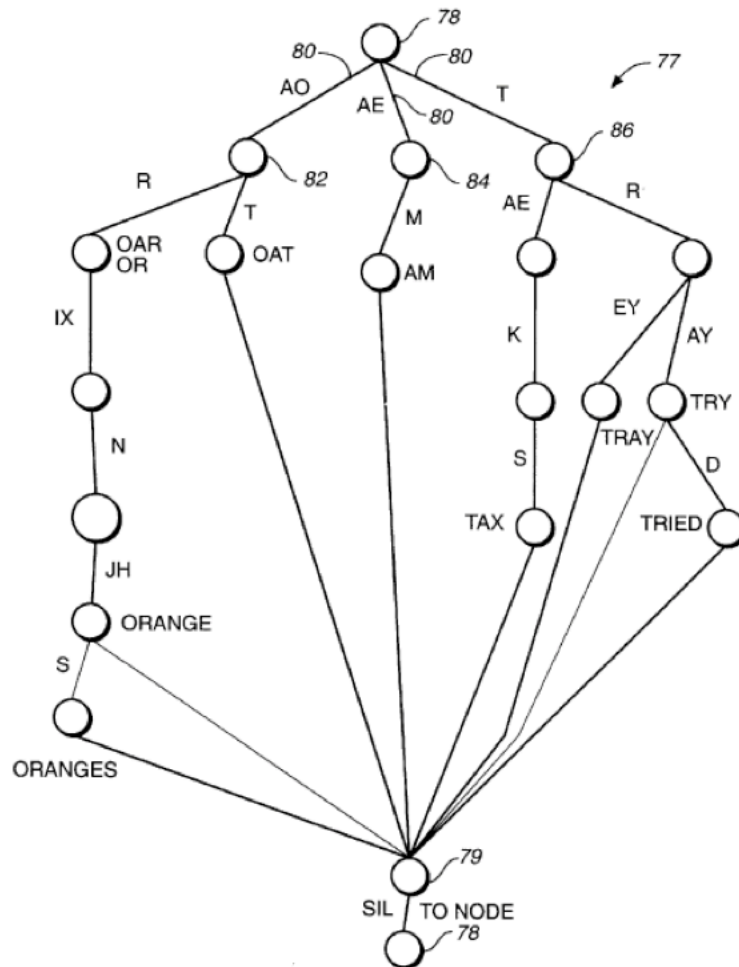
Here, one of ordinary skill would conclude that Applicant’s application clearly supports the full scope of the claims and there is possession of the claimed invention. For example, **Figure 4 of Applicant’s Specification illustrates a “process performed by a**

lexical tree processor to determine a temporary lexical tree score.” Applicant’s As-filed Specification at 11 (emphasis added). *See also* FIG. 4; Applicant’s As-filed Specification at 11-12 (describing the process for determining a temporary lexical tree score for a lexical tree in connection with Figure 4); FIG. 6 (illustrating steps S42 for temporary lexical tree score determination). **The temporary lexical tree score “can be used to prune out lexical trees from the processing which have a low score.”** Applicant’s As-filed Specification at 11 (emphasis added). **In other words, the process for processing the lexical tree branches having a temporary lexical tree score lower than a threshold would not continue to determine a “final score via processing a larger amount of model data than that was processed to generate the initial score,”** as originally recited in claim 10 (now canceled; the subject matter of which has been incorporated in amended independent claim 1). *See, e.g.*, Applicant’s As-filed Specification at 13 (“**Pruning** is applied to the lexical tree by the lexical tree processor to delete paths in order to keep the breadth of the search manageable The **pruning** applied locally by the lexical tree processor can be purely on the basis of a local **threshold**”).

’140 *File History*, 104; *see also* ’140 Patent, 7:46-8:44, 8:64–9:4, Figs. 4, 6. In response, the Examiner issued a Notice of Allowance, indicating that pruning portions of lexical trees based on an initial score is sufficient to teach limitations 1(d) and 1(e).

87. *Jiang* teaches, in detail, generating an initial score in a lexical tree search, which (as discussed below regarding claim 1(e)) is later used for pruning portions of the lexical tree scoring below a threshold. As discussed above, *Jiang* teaches that incoming speech is “sampled” into extracted feature vectors or

codewords. Thereafter, the sampled audio is provided to a tree search engine 74 configured to determine a user's utterance by searching a lexical tree, calculating the most likely string of phonemes spoken. *Jiang*, 7:29-54. This is a well-known process in speech recognition circuits, as expressly stated by *Jiang* when it refers to its tree search as being "all done in a known manner." *Jiang*, 8:51. *Jiang*'s lexical tree aligns with the description of a lexical tree provided by *Ravishankar*, as discussed above, teaching organizing HMMs as a lexical tree in which words having shared pronunciations share a portion of a branch. *Ravishankar*, 33. *Jiang* explains "as tree 77 is traversed from the input node 78 to the output node 79, a **score** is assigned to **each node** connected to a phoneme branch then under consideration by the speech recognition system. **The score is indicative of a likelihood that the particular phoneme being examined is the actual phoneme** indicated by the codeword received from the feature extraction module 66." *Jiang*, 8:16-24, 8:44-51.



Jiang, Fig. 3. In one example, a lexical tree includes an exemplary subset of words stemming from the initial phonemes (after silence) of AO, AE, and T. *Jiang*, Fig. 3, 8:5-43. The below citation from *Jiang* steps through a detailed example in which the word “orange” is recognized by *Jiang*’s circuitry:

For example, if the word ORANGE were input into system 60 by the user, feature extraction module 66 would likely divide the word ORANGE into codewords indicative of the phonemes represented as follows: AO, R, IX, N, JH. **As the tree search engine traverses tree 77, it preferably computes a score, for each phoneme branch**

considered in tree 77, wherein the score represents the likelihood that the particular phoneme encoded by the codeword corresponds to the phoneme for the branch under consideration. Thus, tree search engine 74 **computes a score for node 82** which indicates that the first codeword under consideration is highly likely to be represented by the AO phoneme corresponding to the branch under consideration. **Tree search engine 74 also preferably computes a score for each of the other nodes 84 and 86 in tree 77,** wherein the score is indicative of the likelihood that the codeword being analyzed is represented by the phonemes AE, and T. Under ideal circumstances, the score assigned to nodes 84 and 86 is lower than the score assigned to node 82.

Jiang, 8:25-43. Here, *Jiang* teaches computing “scores” as the lexical tree is traversed in a search for the most likely word. Scores are computed at each node and phoneme branch combination, including for each of nodes 84, 86 in tree 77. The node with the highest score (node 82), having the highest score, is determined to be the most likely, while (as will be discussed regarding claim 1(e) below) branches stemming from nodes having scores below a threshold may be “pruned” from the search. *Jiang*, 8:44-64. Pruning is done at each node in *Jiang*’s lexical tree, any of which provide support for an “initial score,” as claimed. To the extent that this score must be the *first* score generated for the audio sample, I note that *Jiang* expressly describes calculating scores for branches 82, 84, and 86 of tree 77, which (as seen in Fig. 3) are the first branches after silence received from the audio sample as compared to the “first codeword under consideration.” *Jiang*, 8:35-43, Fig. 3. For this reason, it would have also been obvious to a POSITA to begin calculating the

“scores” at the initial branch of the lexical tree, based on the description of *Jiang*.

Id.

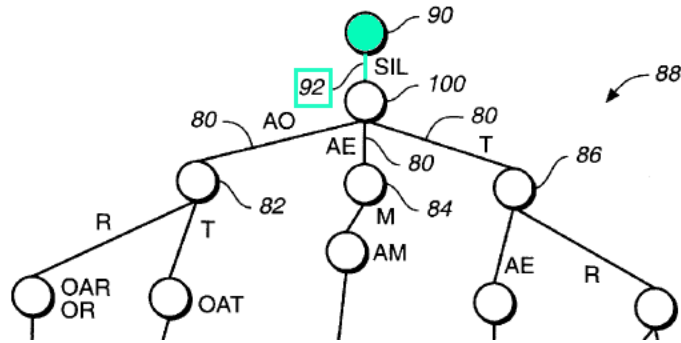
- f) **Claim 1(e): “[wherein] the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.”**

88. As I discussed above, during Prosecution of the ’140 Patent, Applicant argued that “pruning” a lexical tree based on a branch’s score falling below a threshold supports, and thus teaches, limitation 1(e): “The temporary lexical tree score ‘can be used to prune out lexical trees from the processing which have a low score.’...In other words, the process for processing the lexical tree branches having a temporary lexical tree score lower than a threshold would not continue to determine a ‘final score via processing a larger amount of model data than that was processed to generate the initial score[.]’” ’140 File History, 104, citing ’140 Patent, 7:57-63. *Jiang* teaches this limitation.

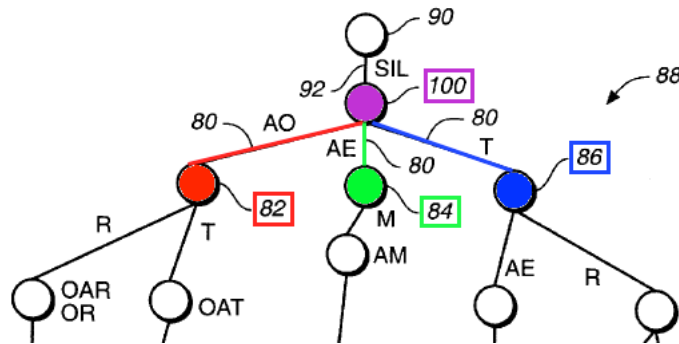
89. Specifically, *Jiang* teaches a “minimum threshold value” used when searching its lexical tree. *Jiang*, 2:39-46. Branches of the lexical tree whose initial scores are below the threshold are pruned, being no longer considered for further processing: “a pruning technique...[] is accomplished by **comparing, at each node, the scores assigned to that node with the largest score on any of the other nodes corresponding to the frame being considered. If the score at a particular node is**

sufficiently low compared to the maximum score for other corresponding nodes in tree 77, it is assumed that the branches leading to the node under consideration (the node with the lower score) will not likely be part of the completed most likely phoneme sequence (i.e., the phonemes in that sequence will not likely be part of the final word recognized by the system). Thus, **that branch is dropped (or pruned) from tree 77 and is no longer considered in further processing.**” *Jiang*, 8:52-64, *see also* 2:25-46. In some embodiments, pruning may be determined based on a difference from the maximum score for any node. *Jiang*, 10:51-11:15. “If the score for the node being compared is less than the maximum score by the threshold level, all subsequent branches connected to that node are pruned from the tree, thereby drastically reducing the search space for any given search.” *Id.* Further, as previously discussed, *Jiang* teaches that the threshold level used may be determined empirically in order to balance the potentially competing considerations of computational power and the error rate arising from the strictness of the pruning. *Jiang*, 11:16-24.

90. As mentioned above, *Jiang* describes lexical trees illustrated in Figs. 3-6. Specifically discussing Fig. 4 in detail, *Jiang* explains that the tree 88 begins with node 90 connected to branch 92 representing silence, highlighted below.



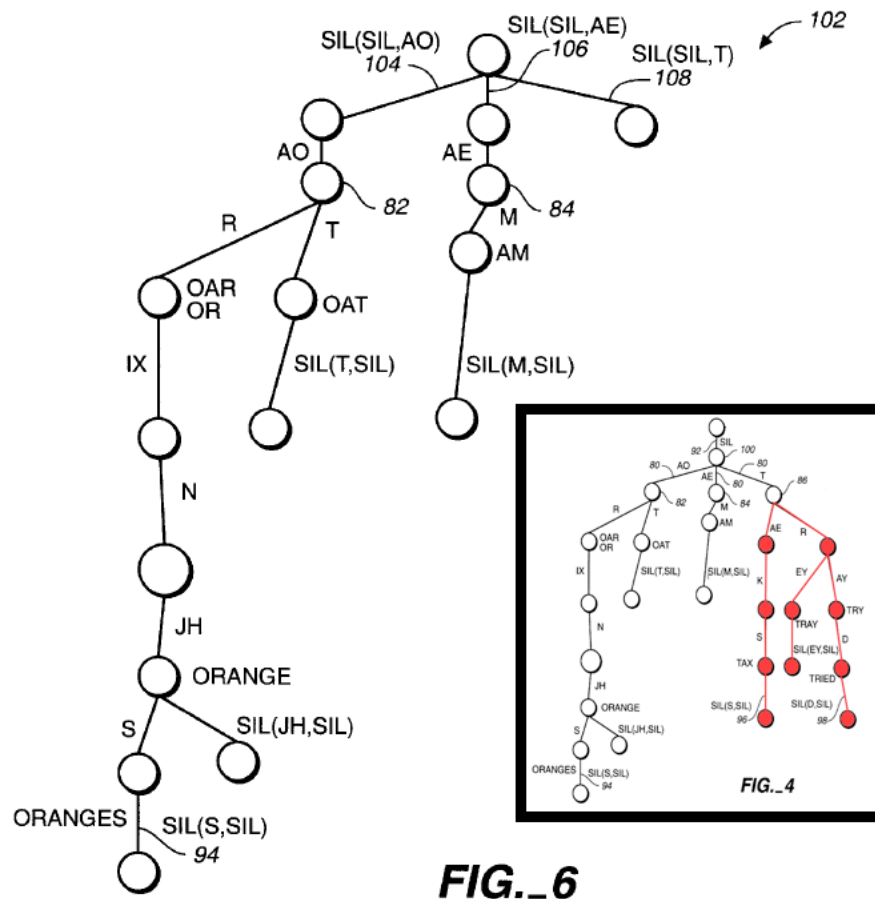
Jiang, 9:28-37, Fig. 4 (excerpted and annotated). The top node 100 then branches out for phonemes AO, AE, and T, which are respectively connected to nodes 82, 84, and 86 highlighted red, green, and blue below.



Jiang, Fig. 4 (excerpted and annotated). These branches are merely representative – lexical trees may have reflect more, fewer, or different branches based on the phonemes of words to be recognized. *Jiang*, 7:65-8:4, 11:8-15. *Jiang* continues: “If the score for the node being compared is less than the maximum score by the threshold level, all subsequent branches connected to that node are pruned from the tree, thereby drastically reducing the search space for any given search. *Id.* FIG. 6 illustrates that the score assigned to the node for silence branch 108 is sufficiently

low that the remainder of the branches of the tree have been pruned from the tree.”

Below is a reproduction of Figure 6 with an annotated version of Figure 4 inset, illustrating that the branches below a “T” phoneme have been pruned due to the initial score for this branch falling below a pruning threshold:



Jiang, Fig. 6 inset with Fig. 4 (annotated). In the interest of “computational savings,” *Jiang*’s speech recognition circuit has pruned branches from its lexical tree based on generated initial scores falling below an empirically determined threshold. *Jiang* expressly characterizes its process as “beginning” by assigning scores to nodes 104,

106, 108, the using those initial (“beginning”) scores for nodes 104, 106, 108 to determine which branches to continue processing to determine a final score for a final node representing a recognized word. *Jiang* 10:53-62 (discussing the tree traversal “beginning” by assigning scores to nodes 104, 106, 108), 10:67–11:3 (discussing assigning scores to each of nodes 104, 106, 108), 11:8-12, 8:52-64. Similarly, when comparing Figs. 5 and 6, a POSITA would have again recognized how *Jiang* teaches claim 1(e). Initial scores are generated to determine which branches to prune (again, as shown below, branches below “T” are pruned):

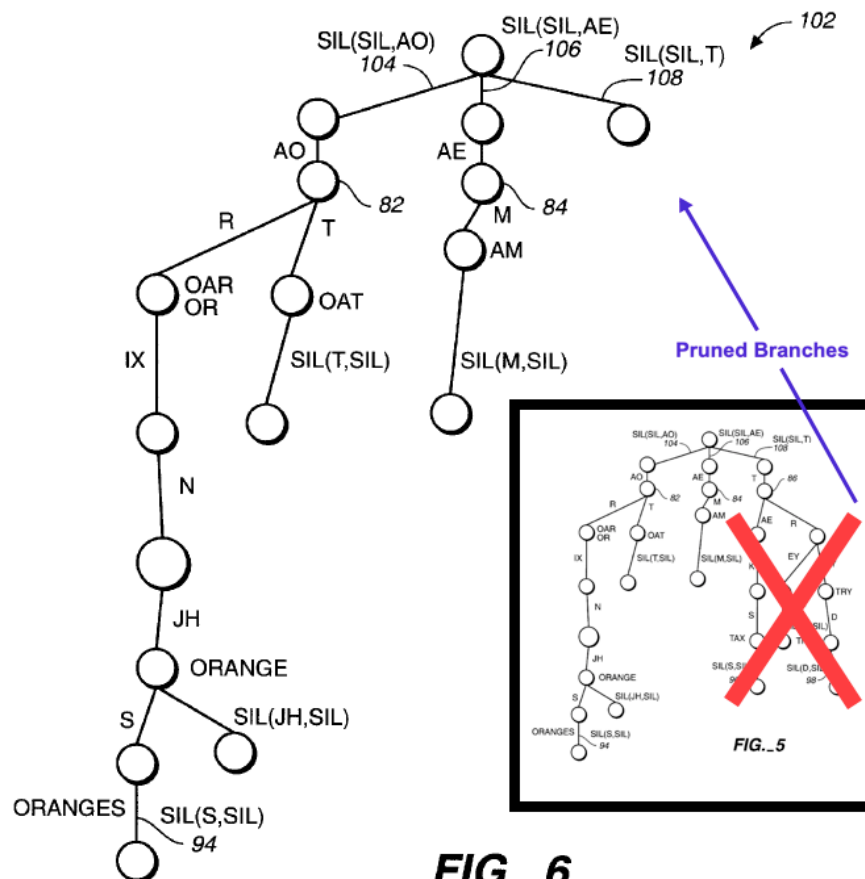


FIG. 6

Jiang, Fig. 6 inset with Fig. 5 (annotated to illustrate pruned branches). Whereas, branches associated with R, IX, N, and JH continue being processed to determine a final score indicating that the word uttered was “oranges,” while the branches stemming from node 108 (i.e., AE, K, S, R, EY, AY, and D as shown below node 88 in Fig. 5) are no longer processed to generate a final score.

91. A POSITA would have recognized that *Jiang*’s description of pruning teaches Claim 1(e) by the same reasoning used by Applicant in identifying the written description support of the limitation during prosecution. *Jiang* teaches lexical tree scores computed for each of a plurality of branches. Branches having a lexical tree score lower than a threshold would not continue to be processed to determine a “final score via processing a larger amount of model data than that was processed to generate the initial score[.]” *Compare Jiang*, 2:25-46, 8:52-64, with ’140 *File History*, 104. Searching is ceased for those branches falling below this “minimum threshold value,” resulting in reduced computation required and no larger amount of model data being utilized from the HMM acoustic model. *Jiang*, 2:25-46, 3:15-21, 8:52-64, 11:38-44. Conversely, for branches whose initial score is determined to be sufficiently high, *Jiang*’s circuit continues processing those branches via additional (i.e., “a larger amount”) of the HMM model data representing phonemes in the lexical tree than was used to generate the initial score. *Id.*

2. Claim 2: “...wherein the probability is an input to an evaluation of a state transition of a model of states”

92. *Jiang* teaches “the probability is an input to an evaluation of a state transition of a model of states[,]” as claimed. Specifically, *Jiang*’s discussion of using HMMs to cumulatively determine the most likely phonemes uttered teaches “the probability is input to an evaluation of a state transition of a model of states.” *Jiang*, 1:19-2:15. I note, here, that “the probability” in this claim has its antecedent basis in limitation 1(c), stating “wherein each of the plurality of processors is configured to compute **a probability** using the acoustic model data in the acoustic model memory.” *Jiang* expressly states that in order to find the most likely sequence of states corresponding to a sequence of acoustic vectors, a Viterbi algorithm is employed, which “**starts at the first frame and proceeds one frame at a time,**” computing a “**probability score for each state.**” *Jiang*, 2:1-7. “**A cumulative probability score is successively computed** for each of the possible state sequences as the Viterbi algorithm analyzes the acoustic signal frame by frame.” *Jiang*, 2:7-10. The state sequence of HMMs “having the highest probability score computed by the Viterbi algorithm provides the most likely state sequence for the entire utterance. **The most likely state sequence is then converted into a corresponding spoken subword unit, word, or word sequence.**” *Jiang*, 2:10-15. Here, a probability calculated for a given frame is input to the evaluation of the HMM (a model of states) to compute a cumulative probability frame by frame that the sequence of observed

states represents a particular corresponding subword unit, such as a phoneme. *Jiang*, 2:1-15, 1:41-48, 1:58-67. This teaching aligns with the description found in *Schmandt*, teaching “An **HMM** consists of a number of internal states; the **model** passes from an initial state to the final state as a step-by-step process generating an observable output at each step (**state transition**). For example, the states may correspond to phonemes contained in a word with the observable output corresponding to the presence or absence of a number of acoustic features.” *Schmandt*, 57-58.

93. Additionally, as discussed, *Jiang* teaches a speech recognition circuit performing a lexical tree search using the computed probabilities to determine the most likely word uttered. *Jiang*, 7:29-54, 8:5-51. Through its description of the lexical tree search using probabilities generated using an HMM, *Jiang* further teaches “the probability is an input to an evaluation of a state transition of a model of states[,]” as claimed. *Jiang*’s circuit searches a phoneme-based prefix tree, as illustrated in Figs. 3-6 and discussed in detail above regarding limitations 1(d)-1(e), to “determine a most likely state sequence for which the overall probability is maximum, given the transition probabilities between states and the observation probabilities. *Jiang*, 1:19-57, 2:1-24, 2:47-3:21, 7:12-15. Thus, since *Jiang*’s speech recognition circuit performs a process in which calculated probabilities representing the likelihood of phonemes are input to the evaluation of a word-level HMM model

to determine the most likely word uttered, *Jiang* again teaches “the probability is an input to an evaluation of a state transition of a model of states[,]” as claimed. *Jiang*, 1:19-27, 3:10-21 (expressly teaching “evaluating the probability” of phonemes), 7:29-46. In summary, through both its description of (1) probabilities being calculated frame by frame to determine the most likely phonemes uttered and (2) successively inputting probabilities of the likely phonemes uttered to search a lexical tree, *Jiang* teaches “the probability is an input to an evaluation of a state transition of a model of states[,]” at both the subword (phoneme calculation) and whole word (lexical tree search) levels.

3. Claim 5: “...further comprising: a buffer for storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors”

94. *Jiang*’s speech recognition process begins with incoming spoken utterances being captured by a microphone and converted to a “sequence of digital signals” which is provided to feature extraction module 66 via an A/D converter at a sample rate of approximately 16 kHz, which divides the digital signal into frames approximately 10 milliseconds in duration. *Jiang*, 6:46-7:10. The frames are then encoded by the feature extraction module 66 into a feature vector and encoded into codewords. *Id.* “Thus, feature extraction module 66 provides, at its output the feature vectors (or codewords) for each spoken utterance. The feature extraction module 66 preferably provides the feature vectors (or codewords) at a rate of one codeword

approximately every 10 milliseconds.” *Jiang*, 7:5-10, *see also* 1:58-67, Fig. 2. A POSITA would have recognized *Jiang*’s stream of incoming feature vectors (or equivalently, codewords derived therefrom) would have been provided from the feature extraction module to the processing unit to the tree search engine implemented in processing unit 21 for speech recognition via storage in memory connected via the system bus 23, teaching or at least rendering obvious the claimed “input buffer.” Using an input buffer to store and pass data to be processed between computer elements was well-known and widely used prior to the ’140 Patent and would have been well within the skill of a POSITA to implement in the circuit of *Jiang+Chen*. An input buffer would have been the structure expected by a POSITA for transmitting *Jiang*’s feature vectors (or codewords) from its feature extraction module 66 to its tree search engine implemented in processing unit 21. A POSITA would have known a buffer is a well-known memory structure into which a first computer module enters data at a first rate while a second module reads data at a second rate, which may differ from the first rate. A POSITA would have recognized buffers provide flexibility to systems such as the speech recognition circuit of *Jiang+Chen* and would have expected the *Jiang+Chen* circuit to have an input buffer. Specifically, for the combination described above in which *Jiang*’s processing unit 21 is modified pursuant to *Chen* to comprise one or more clusters of processors, in my opinion a POSITA would have recognized *Jiang* in view of *Chen*

teaches storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors, rendering obvious the claimed “buffer.” To the extent a POSITA would have had to modify the system of *Jiang+Chen* to include an “input buffer,” a POSITA would have been able to do so without undue experimentation and with a reasonable expectation of success because input buffers were so well known.

4. ***Claim 7: “...the one or more clusters of processors comprises a first cluster of processors and a second cluster of processors; the first cluster comprises a first acoustic model memory; and the second cluster comprises a second acoustic model memory that is distinct and separate from the first acoustic model memory”***

95. As I discussed above, *Jiang* in view of *Chen* teaches a first cluster of processors and second cluster of processors, each of which comprising a cluster shared memory, as color-coded red, green, blue, and yellow below in annotated Fig.

4.

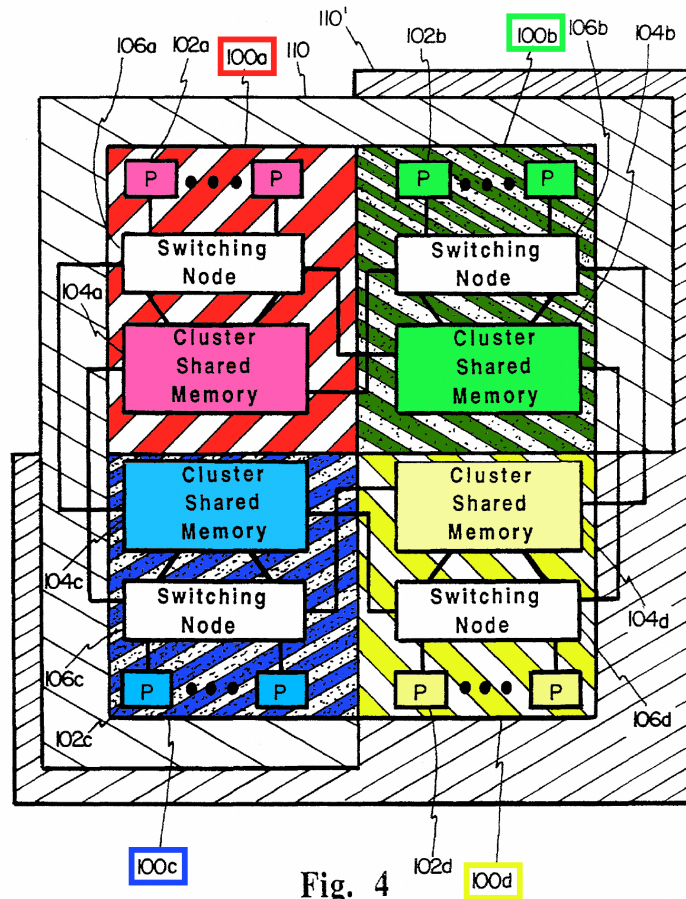


Fig. 4

Chen, Fig. 4 (annotated). Of course, a POSITA would have recognized particular applications of the modified system may have more or fewer than 4 clusters. In the proposed modified speech recognition circuit, each cluster 100a-d would have acoustic model data stored in its respective cluster shared memory 104a-d. *See*, discussion of limitations 1(a)-1(b). Thus, *Jiang* in view of *Chen* teaches a first cluster comprising a first acoustic model memory and a second cluster comprising a second acoustic model memory. Additionally, *Chen* teaches that some, but not all, the clusters are interconnected, stating that the processors 102a in cluster 100a can directly access memories 104a, 104b, and 104c of clusters 100a, 100b, and 100c,

respectively, but cannot directly or adjacently access the memory 104d of cluster 100d, while the processors 102d in cluster 100d can directly access memory 104d, 104b, and 104c, but cannot directly or adjacently access the memory 104a. *Chen*, 9:5-19, 9:30-43. That is, *Chen* teaches or at least renders obvious at least a first cluster (e.g., cluster 100a) comprising a first acoustic model memory (e.g., memory 104a) and a second cluster (e.g., cluster 100d) comprising a second acoustic model memory (e.g., memory 104d) that is “separate and distinct” from the first acoustic model memory (104a). *Id.* Further, *Chen*’s clustered computing architecture is described as providing multiple processors in “**physically separable clusters**, each cluster having a common cluster shared memory that is symmetrically accessible by all of the processors in that cluster.” *Chen*, 5:9-17. Thus, in my opinion, a POSITA would have recognized that *Chen* teaches or at least renders obvious clusters are “separate and distinct” from one another.

IX. OPINIONS REGARDING GROUND 2

A. Claims 1–3, 5, and 7-8 Are Rendered Obvious By *Jiang* in View of *Chen* and *Lucke*

1. Claim 1

a) Claim 1(e): “[wherein] the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.”

96. Like *Jiang*, *Lucke* teaches a speech recognition circuit (*Lucke*, 0002-0005, Figs. 1,7) (Ex. 1008) using HMM acoustic models (*Lucke*, 0008, 0025, 0079)

used to drive a lexical tree search (*Lucke*, 0017-0023, Fig. 4). *Lucke*'s circuit calculates "acoustic scores [] using a series of features on the basis of acoustic models which are connected along a series of branches...extending from the root node" of the word network tree to the "end node" representing a recognized word. *Lucke*, 0025, 0070. Also, like *Jiang*, *Lucke* prunes its lexical tree of branches whose initial generated score is below a threshold for processing efficiency. *Lucke*, 0027-0029, 0087. *Lucke*, however, states expressly that branches of its lexical tree for words having an acoustic score "**greater** than a predetermined threshold value" are selected to remain in consideration and "**the processing continues.**" *Lucke*, 0027-0029, 0081-0082. The continued processing utilizes a larger amount of the HMM acoustic model data to continue performing calculations of hypothesized words (and, subsequently through iterations, word sequences) until its matching section determines "from among such word sequences, a word sequence having the greatest final score...as the speech recognition result." *Lucke*, 0029, 0085.

97. A POSITA would have been motivated to *continue* processing branches of the lexical tree having a sufficiently *high* initial score as taught by *Lucke* in the speech recognition circuit of *Jiang+Chen*, already expressly configured to *cease* processing particular branches of a lexical tree based on a *low* initial score. *Lucke*, 0027-0029, 0081-0082. Continuing the processing would have been the naturally expected functionality of the system to a POSITA in order for the circuit to

successfully determine a final recognized word, leading to a reasonable expectation of success. A POSITA would have accomplished any modification necessary to the system via minor software revisions, well within the skill of a POSITA at the time of the '140 Patent. A POSITA would have recognized that the system of *Jiang+Chen* modified via *Lucke* to continue processing branches of its lexical tree whose initial scores were above the pruning threshold teaches the claimed limitation of “the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.”

2. Claim 5: “...further comprising: a buffer for storing one or more feature vectors coupled to at least one of the plurality of processors of the one or more clusters of processors”

98. *Lucke* describes a “buffer [] temporarily stores the features supplied from the feature extraction section[.]” *Lucke*, 0076-0077. *Lucke*’s matching section calculates scores using “the series of the necessary features stored in the buffer.” *Id.* A POSITA would have recognized that *Lucke*’s buffer storing features supplied from a feature extraction section operating on incoming speech data teaches the claimed “buffer for storing one or more feature vectors[.]” A POSITA would have understood, based on *Lucke*’s description of matching section 4 (which contains the input buffer), *Lucke*’s input buffer would have been coupled to its CPU 102 to provide the extracted features for speech recognition. *Lucke*, 0076-0077, 0111, Fig.

7. A POSITA would have been motivated to combine *Lucke's* buffer into the system as a well-known, operable method of storing the feature vectors, allowing each coupled processor of the *Jiang+Chen* circuit to access the feature vector as the particular processor becomes available for a next operation and providing additional flexibility in processor scheduling. As mentioned above, a POSITA would have recognized buffers provide flexibility to systems such as the speech recognition circuit of *Jiang+Chen* and would have expected the *Jiang+Chen* circuit to have an input buffer. A POSITA would have recognized the value of this flexibility to the circuit, allowing the rate of recognition by the processors to be disassociated from the rate of incoming frames of speech, which may vary. A POSITA would have had a reasonable expectation of success for the combination of *Lucke's* input buffer into the system of *Jiang+Chen* because each of *Jiang* and *Lucke* teach typical speech recognition circuits that extract feature vectors via frames derived from sound input to a microphone and supplied to an A/D converter. *Compare Jiang 6:46-7:10 with Lucke, 0005-0006.*

X. OPINIONS REGARDING GROUND 4

A. Claim 4 is Rendered Obvious By *Jiang* in View of *Chen* and *Robinson*

1. Claim 4: “...wherein the probability is computed from a Gaussian mixture model and one or more feature vectors.”

99. Like *Jiang*, *Robinson* describes a speech recognition circuit that compares extracted feature vectors to a speech model comprising HMMs. *Robinson*, 1:5-9, 8:28-55, 9:4-16 (Ex. 1009). Specifically, *Robinson* describes a phoneme probability calculator that relates the “acoustic observations” of speech to “linguistic states such as phonemes” to compute the “**probability** of seeing an acoustic **vector**, $O(t)$ at time (t) given the state $S(t)$.” *Robinson*, 8:56-9:16. A POSITA would have recognized that this “acoustic vector” derived from and representing features of incoming speech is a “feature vector” as described by the ’140 Patent. Further, *Robinson* circuit specifically performs these probability computations using a model that is a “mixture of Gaussians,” which a POSITA would have recognized teaches the claimed “Gaussian mixture model.” Specifically, in the combined system of *Jiang+Chen+Robinson*, a POSITA would have been motivated to use the Gaussian Mixture Model of *Robinson* to compute the probabilities taught by *Jiang* which are input to an evaluation of a state transition of a model of states, as discussed above regarding claim 2. Gaussian mixture models were well-known and widely used prior to the ’140 Patent, as detailed in *Ravishankar*, providing a commonly-used framework for representing output probabilities in Hidden Markov Models.

Ravishankar, 28-29. Because Gaussian mixture models were one of “[s]everal parametric models [] used to approximate” the probability computation known at the time of the ’140 Patent, a POSITA would have found Gaussian Mixture Models obvious to try in the *Jiang+Chen* speech recognition circuit. *Robinson*, 9:4-16. A POSITA would have been able to combine *Robinson*’s Gaussian Mixture Models using simple programming techniques within their skill without undue experimentation, leading to a reasonable likelihood of success for the combination.

XI. OPINIONS REGARDING GROUND 6

A. Claim 4 is Rendered Obvious By *Jiang* in View of *Chen* and *Wrench*

100. While *Chen* describes a shared control mechanism 122 coupled to at least one of its plurality of processors that is utilized by a common parallel user interface to coordinate and communicate among processes executing in parallel in the clustered processing computer system, neither *Chen* nor *Jiang* teaches initiating speech recognition processing in at least a subset of processors via a search controller. *Chen*, 6:4-22, 18:43-63, 19:14-57. *Wrench* describes a speech recognition circuit comprising a plurality of processors under the control of a single “master processor” controller that acts as an “executive” to its plurality of speech recognition processors. *Wrench*, 4:40-58, 5:6-24, 6:38-7:32, Fig. 2 (Ex. 1010). *Wrench* explains that the “system executive” (master processor) schedules tasks, such as speech recognition tasks performed by the template processors. *Wrench*, 7:3-15. Thus, a

POSITA would have recognized that *Wrench*'s master processor teaches a controller coupled to a plurality of processors that "initiates speech recognition processing in at least a subset of processors," as claimed. *Wrench*'s master processor's uses an "initialization" program module that "performs the various tasks for **initializing** the RS parameters[,]" where (based on 2:20-34 and 4:44-48) RS represents the "speech recognition subsystem." *Wrench*, 6:40-50, 4:44-58. In my opinion, a POSITA would have been motivated incorporate *Wrench*'s master processor as part of the shared control mechanism taught by *Chen* to initiate the search process taught by *Jiang* as an operable mechanism for initializing speech recognition in its plurality of processors, ensuring smooth, cohesive operation of the plurality of processors, improving the functionality of the overall speech recognition circuit. This modification using *Wrench* in the circuit of *Jiang+Chen* would have been within the skill of a POSITA without undue experimentation, leading to a reasonable likelihood of success because it would have only required providing *Wrench*'s software modules to hardware already present in the circuit of *Jiang+Chen*.

XII. CONCLUSION

101. For the aforementioned reasons, in my opinion, to a POSITA, the claims of the '140 Patent are rendered obvious by the prior art.

102. I hereby declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct, and that all statements made

Declaration of Christopher Schmandt

Patent No. 10,971,140

of my own knowledge are true and that all statements made on information and belief are believed to be true. I understand that willful false statements are punishable by fine or imprisonment or both. *See* 18 U.S.C. § 1001. If called to testify as to the truth of the matters stated herein, I could and would testify competently.

Date: December 2, 2022

Respectfully submitted:

By:

A handwritten signature in black ink, appearing to read "Chris Schmandt", written over a horizontal line.

Chris 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

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

- *Facebook v. Windy City*, 2017
- *Facebook v. EVERYMD.COM LLC*, 2017
- *Comcast v. Promptu*, 2017
- *Express Mobile, Inc v. Big Commerce*, 2017
- *Facebook v. Hypermedia*, 2017
- *Microsoft v. Uniloc 2017 LLC*, 2019
- *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

Professional Experience -- Intellectual Property

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"
- 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

Most publications can be found at <http://www.media.mit.edu/speech/publications> or <http://living.media.mit.edu/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. Angebrannndt, 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.

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.
