
IN THE
UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE PATENT TRIAL AND APPEAL BOARD

AMAZON WEB SERVICES, INC.
AND AMAZON.COM SERVICES LLC,

Petitioners

- vs. -

ZENTIAN LIMITED

Patent Owner

Patent No. 10,062,377

Issued: August 28, 2018

Inventor: Guy Larri, et al.

Title: DISTRIBUTED PIPELINED PARALLEL SPEECH
RECOGNITION SYSTEM

Inter Partes Review No. IPR2023-01194

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 10,062,377**

Mail Stop Patent Board
Patent Trial and Appeal Board
P.O. Box 1450
Alexandria, VA 22313-1450

TABLE OF CONTENTS

	Page
I. INTRODUCTION	1
II. SUMMARY OF THE '377 PATENT	1
A. Patent Description	1
B. Summary of Unpatentability	2
C. Priority Date	4
D. Level of Skill of a POSITA.....	4
III. REQUIREMENTS FOR IPR UNDER 37 C.F.R. §42.104.....	5
A. Grounds for Standing (37 C.F.R. §42.104(a))	5
B. Identification of Challenges (37 C.F.R. §42.104(b))	5
C. Claim Construction (37 C.F.R. §42.104(b)(3)).....	6
IV. Showing of Analogous Prior Art.....	6
A. <i>Jiang</i>	6
B. <i>Smyth</i>	7
C. Nguyen	8
D. Boike.....	9
E. Baumgartner	9
V. GROUND 1: CLAIM 1 IS OBVIOUS OVER <i>JIANG</i> AND <i>SMYTH</i>	10
A. Claim 1(Pre)	10
B. Claim 1(a).....	10
1. Summary of Claim 1(a) Mapping.....	10
2. File History	11

TABLE OF CONTENTS
(Continued)

	Page
3. Jiang Alone Teaches Claim 1(a)	11
4. Jiang in Combination with Smyth Teaches Claim 1(a)	14
a) Motivation to Modify <i>Jiang</i> to Include <i>Smyth</i> 's First Programmable Device	16
C. Claim 1(b)	17
1. The '377 Patent's Discussion of "extracted" or "derived quantities"	17
2. Jiang's Teachings	18
D. Claim 1(c)	19
1. Summary of Claim 1(c) Mapping	19
2. '377 Patent's Discussion of "distance"	20
3. Jiang's Teachings	21
a) "a plurality of acoustic states of an acoustic model"	21
b) "a feature vector"	22
c) "calculat[ing] distances..."	23
4. Smyth's Teachings	28
a) "a second programmable device programmed to"	28
b) "a plurality of acoustic states"	31
c) "[plurality of speech states] of an acoustic model"	33

TABLE OF CONTENTS (Continued)

	Page
d) “[programmed to] calculate distances indicating the similarity...”	35
5. Motivation to Modify Jiang to Include Smyth’s Second Programmable Device	39
6. Motivation to Modify Jiang to Include Smyth’s Second Programmable Device Performing Distance Calculations.....	41
E. Claim 1(d).....	42
1. Jiang’s Teachings	42
2. Smyth’s Teachings	43
F. Claim 1(e).....	45
1. Jiang’s Teachings	45
a) “a third programmable device”.....	45
b) “programmed to identify spoken words in said digital audio stream using Hidden Markov Models and/or Neural Networks”	46
2. Smyth’s Teachings.....	48
a) “a third programmable device”.....	48
b) “programmed to identify spoken words in said digital audio stream”.....	49
c) “using Hidden Markov Models and/or Neural Networks”	51
d) Motivation to Modify <i>Jiang</i> with <i>Smyth</i> for Claim 1(e)	52

TABLE OF CONTENTS **(Continued)**

	Page
G. Claim 1(f)	56
H. Claim 1(g).....	59
I. Claim 1(h).....	60
VI. GROUND 2: CLAIMS 2-3 ARE OBVIOUS OVER <i>JIANG</i> , <i>SMYTH</i> , AND <i>NGUYEN</i>	62
A. Claim 2	62
B. Claim 3	65
VII. GROUND 3: CLAIM 4 IS OBVIOUS OVER <i>JIANG</i> , <i>SMYTH</i> , <i>NGUYEN</i> , AND <i>BOIKE</i>	66
VIII. GROUND 4: CLAIM 5 IS OBVIOUS OVER <i>JIANG</i> , <i>SMYTH</i> , <i>NGUYEN</i> , AND <i>BAUMGARTNER</i>	68
IX. GROUND 5: CLAIM 6 IS OBVIOUS OVER <i>JIANG</i> , <i>SMYTH</i> , <i>NGUYEN</i> , <i>BAUMGARTNER</i> , AND <i>BOIKE</i>	73
X. THE BOARD’S DISCRETION UNDER 35 U.S.C. §314(a) and §325(d).....	74
XI. CONCLUSION.....	75
XII. MANDATORY NOTICES UNDER 37 C.F.R. § 42.8(A)(1)	76
A. Real Party-In-Interest	76
B. Related Matters.....	76
C. Lead and Back-Up Counsel.....	76
D. Service Information	77

TABLE OF AUTHORITIES

	Page(s)
CASES	
<i>Apple Inc. v. Fintiv, Inc.</i> , IPR2020-00019 (PTAB, Mar. 20, 2021)	74
<i>General Plastic Industrial Co., Ltd. v. Canon Kabushiki Kaisha</i> , IPR2016-01357, Paper 19 (PTAB Sept. 6, 2017).....	74
<i>Phillips v. AWH Corp.</i> , 415 F.3d 1303 (Fed. Cir. 2005) (<i>en banc</i>)	6
STATUTES AND RULES	
35 U.S.C. §102	6, 7, 8, 9
35 U.S.C. §315(c)	1, 5
OTHER AUTHORITIES	
37 C.F.R. §42.100(b)	6
37 C.F.R. §42.104	5, 6
37 C.F.R. §42.122	5
83 Fed. Reg. 197 (Oct. 11, 2018).....	6

EXHIBIT LIST (37 C.F.R. § 42.63(E))

Exhibit	Description
1001	U.S. Patent No. 10,062,377 (“the ’377 Patent”)
1002	Prosecution History for the ’377 Patent (“’377 File History”)
1003	Declaration of Christopher Schmandt (“Dec.”)
1004	U.S. Patent No. 6,374,219 to <i>Jiang</i> (“ <i>Jiang</i> ”)
1005	U.S. Patent No. 5,819,222 to <i>Smyth</i> et al. (“ <i>Smyth</i> ”)
1006	U.S. Patent No. 6,959,376 to Boike et al. (“ <i>Boike</i> ”)
1007	U.S. Patent Application Publication No. 2002/0049582 to Baumgartner et al. (“ <i>Baumgartner</i> ”)
1008	Harry Newton, <i>Newton’s Telecom Dictionary</i> , 19 th Edition CMP Books, March 2003 (“ <i>Newton</i> ”)
1009	Kevin L. Kloker, <i>The Motorola DSP56000 Digital Signal Processor</i> , IEEE Micro. Vol. 6, Issue 6, pp. 29-48, December 1986 (“ <i>Kloker</i> ”)
1010	Christopher Schmandt, <i>Voice Communication with Computers</i> , Van Nostrand Reinhold, 1994 (“ <i>Schmandt</i> ”)
1011	U.S. Patent No. 6,499,012 to Peters et al. (“ <i>Peters</i> ”)
1012	Ph.D. Thesis of Mosur <i>Ravishankar</i> (“ <i>Ravishankar</i> ”)
1013	J. Mariani. <i>Recent Advances in Speech Processing</i> IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 429-440 May 1989 (“ <i>Mariani</i> ”)
1014	<i>Zentian, Ltd. v Apple Inc.</i> , WDTX Waco (6:22-cv-122) Complaint “ <i>Complaint</i> ”
1015	Lawrence Rabiner, <i>Fundamentals of Speech Recognition</i> , Pearson Education, 1993 (“ <i>Rabiner</i> ”)

Petition for *Inter Partes* Review of
U.S. Patent No. 10,062,377

Exhibit	Description
1016	Richard L. Klevans, <i>Voice Recognition</i> , Artech House Inc. 1997 (“ <i>Klevans</i> ”)
1017	U.S. Patent No. 7,120,582 to Young et al. (“ <i>Young</i> ”)
1018	John Holmes, <i>Speech Synthesis and Recognition</i> , 2 nd Edition, Taylor & Francis, 2001 (“ <i>Holmes</i> ”)
1019	U.S. Patent No. 4,926,488 to Nadas et al. (“ <i>Nadas</i> ”)
1020	Frederick Jelinek, <i>Statistical Methods for Speech Recognition</i> , The MIT Press, 1997 (“ <i>Jelinek</i> ”)
1021	Shigeru Katagiri, <i>Handbook of Neural Networks for Speech Processing</i> , Artech House Inc. 2000, (“ <i>Katagiri</i> ”)
1022	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, pp. 257-286, February 1989 (“ <i>Rabiner89</i> ”)
1023	U.S. Patent No. 6,182,037 to Maes (“ <i>Maes</i> ”)
1024	U.S. Patent Application Publication No. 2003/0220796 to Aoyama et al. (“ <i>Aoyama</i> ”)
1025	U.S. Patent No. 5,895,447 to Ittycheriah et al. (“ <i>Ittycheriah</i> ’447”)
1026	U.S. Patent No. 6,092,045 to Stubbley et al. (“ <i>Stubbley</i> ”)
1027	U.S. Patent No. 6,151,574 to Lee et al. (“ <i>Lee</i> ”)
1028	U.S. Patent No. 6,580,814 to Ittycheriah et al. (“ <i>Ittycheriah</i> ”)
1029	U.S. Patent No. 6,452,348 to Toyoda (“ <i>Toyoda</i> ”)
1030	U.S. Patent No. 6,195,634 to Dudemaine et al. (“ <i>Dudemaine</i> ”)
1031	U.S. Patent No. 5,893,058 to Kosaka (“ <i>Kosaka</i> ”)
1032	U.S. Patent Application Publication No. 2001/0010039 to Yang. (“ <i>Yang</i> ”)

Petition for *Inter Partes* Review of
U.S. Patent No. 10,062,377

Exhibit	Description
1033	U.S. Patent No. 5,754,878 to Asghar et al. (“ <i>Asghar</i> ”)
1034	U.S. Patent No. 6,041,400 to Ozcelik et al. (“ <i>Ozcelik</i> ”)
1035	Declaration of June Ann Munford (“ <i>Munford</i> ”)
1036	U.S. Patent No. 5,699,456 to Brown et al. (“ <i>Brown</i> ”)
1037	[RESERVED]
1038	U.S. Patent No. 4,567,606 to Vensko et al. (“ <i>Vensko</i> ”)
1039	U.S. Patent No. 6,182,203 to Simar, Jr. et al. (“ <i>Simar</i> ”)
1040	[RESERVED]
1041	[RESERVED]
1042	[RESERVED]
1043	[RESERVED]
1044	[RESERVED]
1045	[RESERVED]
1046	[RESERVED]
1047	U.S. Patent No. 6,879,954 to Nguyen et al. (“ <i>Nguyen</i> ”)
1048	[RESERVED]
1049	[RESERVED]
1050	[RESERVED]
1051	[RESERVED]
1052	U.S. Patent No. 4,718,094 to Bahl et al. (“ <i>Bahl</i> ”)
1053	Hermann Ney et al. <i>Dynamic Programming Search Strategies: From Digit Strings to Large Vocabulary Word Graphs</i> C.-H. Lee et al. (eds.), Automatic Speech and Speaker Recognition (“ <i>Ney</i> ”)

Petition for *Inter Partes* Review of
U.S. Patent No. 10,062,377

Exhibit	Description
1054	U.S. Patent No. 7,058,678 to Luick et al. (“ <i>Luick</i> ”)
1055	[RESERVED]
1056	[RESERVED]
1057	U.S. Patent No. 5,983,180 to Robinson (“ <i>Robinson</i> ”)
1058	Abdulmesih Aktas et al, <i>Multi-DSP and VQ-ASIC Based Acoustic Front-End for Real-Time Speech Processing Tasks</i> , First European Conference on Speech Communication and Technology EUROSPEECH ’89, Paris, France, September 1989 (<i>Atkas 1989</i>)
1059	Abdulmesih Aktas et al., <i>Speaker Independent Continuous Hmm-Based Recognition of Isolated Words on A Real-Time Multi-Dsp System</i> , Second European Conference on Speech Communication and Technology EUROSPEECH ’91, Genova, Italy, September 1991 (<i>Atkas 1991</i>)
1060	U.S. Patent No. 6,029,221 to Wu et al. (“ <i>Wu</i> ”)
1061	<i>Zentian, Ltd. v Apple Inc.</i> , WDTX Waco (6:22-cv-122) Scheduling Order (“ <i>Scheduling Order</i> ”)
1062	[RESERVED]
1063	[RESERVED]
1064	[RESERVED]
1065	[RESERVED]
1066	[RESERVED]
1067	U.S. Patent No. 7,697,641 to Haddadin et al. (“ <i>Haddadin</i> ”)
1068	U.S. Patent No. 4,745,544 to Renner et al. (“ <i>Renner</i> ”)
1069	U.S. Patent No. 4,908,781 to Levinthal et al. (“ <i>Levinthal</i> ”)
1070	U.S. Patent No. 6,603,990 to Zhang et al. (“ <i>Zhang</i> ”)

Petition for *Inter Partes* Review of
U.S. Patent No. 10,062,377

Exhibit	Description
1071	Declaration of Les Atlas, Ph.D.
1072	<i>Zentian Ltd. v. Amazon.com Services LLC, et al.</i> , No. 6:22-cv-123 (W.D. Tex.) Mar. 22, 2023 Amended Scheduling Order

I. INTRODUCTION

Petitioners Amazon.com Services LLC and Amazon Web Services, Inc. (“Petitioners”) request *Inter Partes* Review of Claims 1-6 (“Challenged Claims”) of USPN 10,062,377, assigned to Zentian Limited upon information and belief. ’377 *Patent* (EX1001). This Petition raises grounds of unpatentability identical to those in already-instituted IPR2023-00035. It neither adds to nor alters any arguments the Board has already considered, will not expand the instituted grounds, and is accompanied by a timely motion for joinder under 35 U.S.C. §315(c). Accordingly, IPR of the Challenged Claims should be instituted.

II. SUMMARY OF THE ’377 PATENT

A. Patent Description

The ’377 Patent is directed to a speech recognition system using multiple programmable devices to perform different steps of the recognition process. ’377 *Patent*, 5:63–6:6, 6:20–33. Figure 16 illustrates a speech recognition system comprising a front end 103, distance calculation engine 104, and search stage 106:

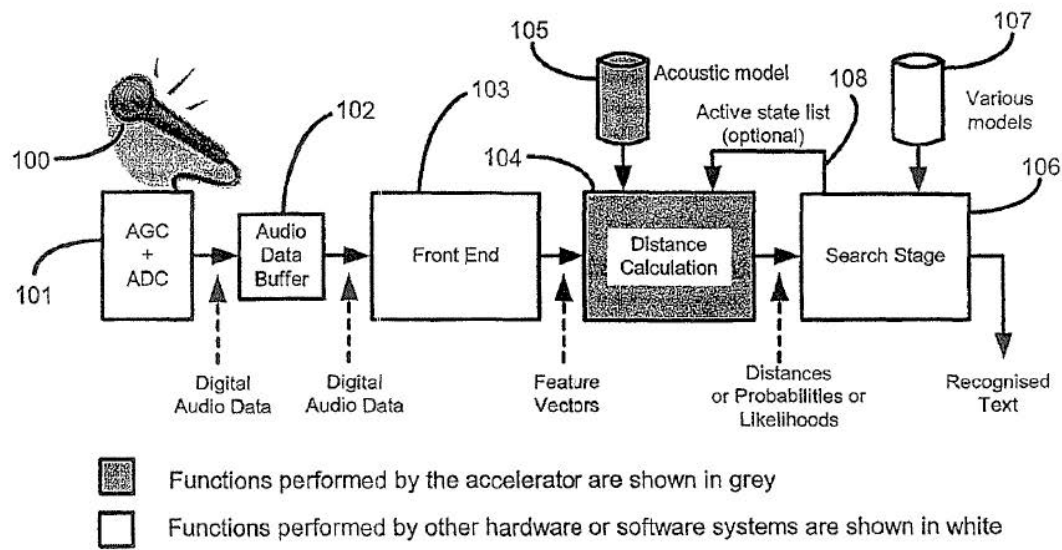


Figure 16

Id., Fig. 16. Claim 1 recites the functionality for each of the three components provided by a separate programmable device. *Id.*, Claim 1.

B. Summary of Unpatentability

Speech recognition commonly includes three general steps: (1) extracting feature vectors from the audio signal; (2) computing, for each feature vector, a likelihood indicating the similarity of the extracted feature vector to modeled spoken speech, where a common model is a Hidden Markov Model (HMM); and (3) searching a lexical tree using the indicated most likely spoken speech to identify

a word. *Dec.*, 97-98.¹ The '377 Patent's claims recite these three general steps, each of which is well-known in the art. The purported point of novelty is that each speech recognition step is performed by a distinct programmable device. *Dec.*, 97-98.

The below mapping relies on *Jiang* as teaching almost every limitation recited in Claim 1, including the three speech recognition steps. *Jiang* expressly teaches that feature vector extraction is performed on a first processor, and distance or likelihood calculation and word identification are performed by a “tree search engine 74” that may be its own dedicated processor. *Jiang*, 6:53-57 (disclosing feature extraction module 66 is a “conventional array processor”), 6:38-42 (disclosing tree search engine 74 may be a “dedicated speech recognition processor”).

Jiang does not expressly teach that each of the three steps is performed on a distinct programmable device; instead, *Jiang* teaches performing the steps on two distinct devices. *Dec.*, 97-98. However, *Jiang* expressly teaches the speech recognition system may comprise various computer configurations, including

¹ Submitted herewith is the Declaration of Les Atlas, Ph.D., who has reviewed, agrees with, and adopts the opinions in the Schmandt Declaration (IPR2023-00035, EX1003). EX1071, ¶¶ 25-26. For consistency with Apple's Petition, citations to Dr. Atlas's opinion are to paragraph numbers of EX1003 containing those opinions as expressed by Mr. Schmandt.

multiprocessor systems. In related art, *Smyth* expressly teaches performing the three claimed speech recognition steps on three distinct programmable devices.

Although *Jiang* is mapped for all of Claim 1's limitations except for three distinct programmable devices, *Smyth* is also relied on for some of the claimed functionality in view of Patent Owner's infringement contentions in the concurrent litigation.

C. Priority Date

The '377 Patent claims its earliest priority to a GB application filed September 14, 2004. '377 *Patent*, (30). For this Petition only, Petitioners apply September 14, 2004, as the priority date for the Challenged Claims.

D. Level of Skill of a POSITA

A POSITA at the time of the '377 Patent would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of speech recognition, or a bachelor's degree in the same fields with at least four years of experience in the field of speech recognition. Additional education or experience might substitute for the above requirements.

III. REQUIREMENTS FOR IPR UNDER 37 C.F.R. §42.104

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

Petitioners certify the '377 Patent is available for IPR and Petitioners are not barred or estopped from requesting IPR of the Challenged Claims. Because Petitioners file an accompanying motion for joinder within one month of institution in IPR2023-00035, this petition is timely under 35 U.S.C. §315(c) and 37 C.F.R. §42.122.

B. Identification of Challenges (37 C.F.R. §42.104(b))

In view of the prior art and evidence presented, the Challenged Claims are unpatentable and should be cancelled. 37 C.F.R. §42.104(b)(1). Based on the prior art identified below, IPR of the Challenged Claims should be granted. 37 C.F.R. §42.104(b)(2).

Proposed Grounds of Unpatentability	
<u>Ground 1</u>	Claim 1 is obvious under §103(a) over <i>Jiang</i> and <i>Smyth</i>
<u>Ground 2</u>	Claims 2–3 are obvious under §103(a) over <i>Jiang</i> , <i>Smyth</i> , and <i>Nguyen</i>
<u>Ground 3</u>	Claim 4 is obvious under §103(a) over <i>Jiang</i> , <i>Smyth</i> , <i>Nguyen</i> , and <i>Boike</i>
<u>Ground 4</u>	Claim 5 is obvious under §103(a) over <i>Jiang</i> , <i>Smyth</i> , <i>Nguyen</i> , and <i>Baumgartner</i>
<u>Ground 5</u>	Claim 6 is obvious under §103(a) over <i>Jiang</i> , <i>Smyth</i> , <i>Nguyen</i> , <i>Baumgartner</i> , and <i>Boike</i>

Sections V–IX identify where each element of the Challenged Claims is found in the prior art. 37 C.F.R. §42.104(b)(4). The exhibit numbers of the supporting evidence relied upon to support the challenges are provided above, and the relevance of the evidence to the challenges raised are provided in Sections V–IX. 37 C.F.R. §42.104(b)(5). Exhibits 1001–1072 are also attached.

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

In this proceeding, claims are interpreted under the same standard applied by Article III courts (i.e., the *Phillips* standard). 37 C.F.R. §42.100(b); 83 Fed. Reg. 197 (Oct. 11, 2018); *Phillips v. AWH Corp.*, 415 F.3d 1303, 1312 (Fed. Cir. 2005) (*en banc*). Petitioners apply the plain and ordinary meaning of all claim terms as understood by a POSITA for all terms not expressly construed below. Petitioners do not waive any argument in any litigation that claim terms in the '377 Patent are indefinite or additional terms need construction.

IV. SHOWING OF ANALOGOUS PRIOR ART

A. *Jiang*

Jiang was neither cited nor considered during prosecution of the '377 Patent. *Jiang* issued April 16, 2002, qualifying as prior art at least under 35 U.S.C. §102(b). *Jiang* (EX1004).

Because *Jiang*, like the '377 Patent, discloses a “system for recognizing speech based on an input data stream indicative of the speech [which] provides

possible words represented by the input data stream as a prefix tree[,]” *Jiang* is in the same field of endeavor as the ’377 Patent. *Compare, Jiang*, Abstract, with ’377 Patent, Abstract; *Dec.*, 99–100.

Because *Jiang*, like the ’377 Patent, is directed to the problem of performing speech recognition and discloses the use of various hardware configurations for speech recognition, *Jiang* is reasonably pertinent to a problem described by the ’377 Patent. *Compare Jiang*, 4:51–5:3, 6:46–7:10, 7:30–47, with ’377 Patent, 1:18–29, 2:21–24; *Dec.*, 99–100.

Therefore, *Jiang* is analogous prior art.

B. *Smyth*

Smyth was neither cited nor considered during prosecution of the ’377 Patent. *Smyth* issued October 6, 1998, qualifying as prior art under at least 35 U.S.C. §102(b). *Smyth* (EX1005).

Because *Smyth*, like the ’377 Patent, discloses a speech recognition system using multiple programmable devices for performing portions of the recognition task, *Smyth* “relates to speech recognition circuits and methods” and is in the same field of endeavor as the ’377 Patent. *Compare ’377 Patent*, 1:18–19, 12:56–63, 13:1–35, 24:5–11, with *Smyth*, 5:20–46, 5:58–6:4, 6:36–63; *Dec.*, 101–102.

Further, both *Smyth* and the ’377 Patent recognize the challenges with performing speech recognition tasks using limited computing resources;

therefore, *Smyth* is reasonably pertinent to a problem described by the '377 Patent.
Dec., 101–102.

Therefore, *Smyth* is analogous art.

C. Nguyen

Nguyen was neither cited nor considered during prosecution of the '377 Patent. *Nguyen* was filed April 22, 2002, qualifying as prior art under at least 35 U.S.C. §102(e). *Nguyen* (EX1047).

Because *Nguyen*, like the '377 Patent, discloses a speech recognition system using feature vectors, similarity calculation, and word search, and more specifically because both are concerned with improving known speech recognition systems, *Nguyen* “relates to speech recognition circuits and methods” and is in the same field of endeavor as the '377 Patent. *Compare* '377 Patent, 1:18-19, 12:56-63, 13:1-35, 24:5–11, *with* *Nguyen*, 2:21–29, 2:31–38, 3:11–23; *Dec.*, 103–104.

Further, both *Nguyen* and the '377 Patent recognize the challenges with performing speech recognition tasks using limited computing resources, and therefore, *Nguyen* is reasonably pertinent to a problem described by the '377 Patent.
Dec., 103–104.

Therefore, *Nguyen* is analogous art.

D. Boike

Boike was neither cited nor considered during prosecution of the '377 Patent. *Boike* was filed October 11, 2001, and issued October 25, 2005, qualifying as prior art under at least 35 U.S.C. §102(e). *Boike* (EX1006).

Like the '377 Patent, *Boike* is related to DSPs and employing multiple DSPs on a single integrated circuit. *Compare* '377 Patent, 37:60-62 (referring to a System on Chip design), Fig. 1A, *with Boike*, Abstract (referring to a single IC with multiple DSPs). Accordingly, *Boike* is in the same field of endeavor as the '377 Patent. *Dec.*, 105.

Because *Boike*, like the '377 Patent, is directed to the problem of dividing processing amongst distinct processors, *Boike* is reasonably pertinent to a problem described by the '377 Patent. *Dec.*, 105. *Compare* '377 Patent, 24:28–33, *with Boike*, 2:21–35.

Therefore, *Boike* is analogous art.

E. Baumgartner

Baumgartner was neither cited nor considered during prosecution of the '377 Patent. *Baumgartner* was published April 25, 2002, qualifying as prior art under at least 35 U.S.C. §102(b). *Baumgartner* (EX1007).

Like the '377 Patent, *Baumgartner* is related to accelerators for speech recognition systems. *Compare* '377 Patent, Abstract, Fig. 21, 30:60–31:5, *with*

Baumgartner, Abstract, Fig. 19, [0002]. Accordingly, *Baumgartner* is in the same field of endeavor as the '377 Patent. *Dec.*, 106.

The '377 Patent purportedly addressed the problem of distributing the computational load in large vocabulary speech recognition systems. *Dec.*, 106; '377 Patent, 1:30–47, 2:21–24. *Baumgartner* also addresses the problem of distributing the computational load in large vocabulary speech recognition systems. *Dec.*, 106; *Baumgartner*, [0005]. Accordingly, *Baumgartner* is reasonably pertinent to a problem described by the '377 Patent. *Dec.*, 106.

Therefore, *Baumgartner* is analogous art.

V. GROUND 1: CLAIM 1 IS OBVIOUS OVER *JIANG* AND *SMYTH*

A. Claim 1(Pre)

Jiang teaches a speech recognition system: “A speech recognition system recognizes speech based on an input data stream indicative of the speech.” *Jiang*, 4:8-9, 4:30-32, 1:15-19, 6:18-20, Abstract, Figs. 1-2.

B. Claim 1(a)

1. Summary of Claim 1(a) Mapping

Jiang, alone or in combination with *Smyth*, teaches or renders obvious Claim 1(a). *Jiang* teaches a feature extraction module 66 for calculating a feature vector, as claimed. *Jiang*, 6:31-36, 6:46–7:10. *Jiang's* feature extraction module 66 is a

programmable device, as the module 66 may be a hardware or software module in computer 20 having CPU 21.

To the extent Patent Owner contends *Jiang* does not disclose “a first programmable device[.]” *Smyth* also teaches a first programmable device used for speech recognition, namely a digital signal processor (DSP) providing frame generator 32 and feature extractor 33. *Smyth*, 5:49-53.

2. *File History*

During original examination, the Applicant admitted that DSPs and CPUs are programmable devices. (EX1002, 306-08). In response to the Examiner’s objection that “programmable devices” are unsupported in the ’377 Specification, the Applicant responded, “a DSP is a well-known programmable device, and a CPU is a well-known programmable device[.]” (EX1002, 301, 307). Per the Applicant, both DSPs and CPUs are programmable devices as recited in the claims.

3. *Jiang Alone Teaches Claim 1(a)*

Jiang teaches or renders obvious Claim 1(a).

Regarding the claimed “digital audio stream,” *Jiang* teaches a plurality of digital samples from an acoustic signal. *Jiang*, 6:62-64, 1:58-67. *Jiang* teaches “speech is input into [the speech recognition] system 60 in the form of an audible voice signal provided by the user to microphone 62.” *Jiang*, 6:46-48, 6:18-20 (referencing the speech recognition system 60). The 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. The stream of digital samples, i.e., the plurality of digital samples, are provided to a feature extraction module 66 as a “sequence of digital signals.” *Jiang*, 6:50-53. The sequence of digital signals is divided into frames of approximately 10ms in duration. *Jiang*, 6:62-65.

Because *Jiang* teaches converting audible (acoustic) speech into a plurality of digital samples, which are in turn divided into frames of 10ms acoustic data, *Jiang* teaches a “digital audio stream,” as claimed. *Dec.*, 107-108.

Regarding the claimed “calculat[ing] a feature vector from a digital audio stream,” *Jiang* teaches:

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 semicontinuous 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, 6:62–7:7². *Jiang* thus teaches the feature extraction module encodes the frames of digital audio samples into feature vectors, which are further encoded (using vector quantization techniques) into codewords. *Dec.*, 109-110. Per Mr. Schmandt, the encoding of the frames of digital audio samples into feature vectors teaches or renders obvious the claimed “to calculate a feature vector ...,” as *Jiang* expressly teaches a feature vector is encoded i.e., calculated. *Dec.*, 109-110.

Regarding the claimed “first programmable device,” *Jiang* teaches that in “a preferred embodiment, feature extraction module 66 is a conventional array processor....” *Jiang*, 6:53-57, Fig. 2, 6:31-36 (disclosing feature extraction module (along with other disclosed modules) “may be either hardware modules in computer 20, or software modules stored in any of the information storage devices disclosed in FIG. 1 and accessible by CPU 21 or another suitable processor”). *Jiang*, 6:31-36. *Jiang* further discloses that its “invention will be described, at least in part, in the general context of computer-executable instructions, such as program modules, being executed by a personal computer.” *Jiang*, 4:54-57. *Jiang* describes the program modules include routine programs performing tasks. *Jiang*, 4:57-60.

Therefore, per Mr. Schmandt, a POSITA would have understood that the disclosed modules, including feature extraction module 66, are *programmed*

² All emphases added unless otherwise stated.

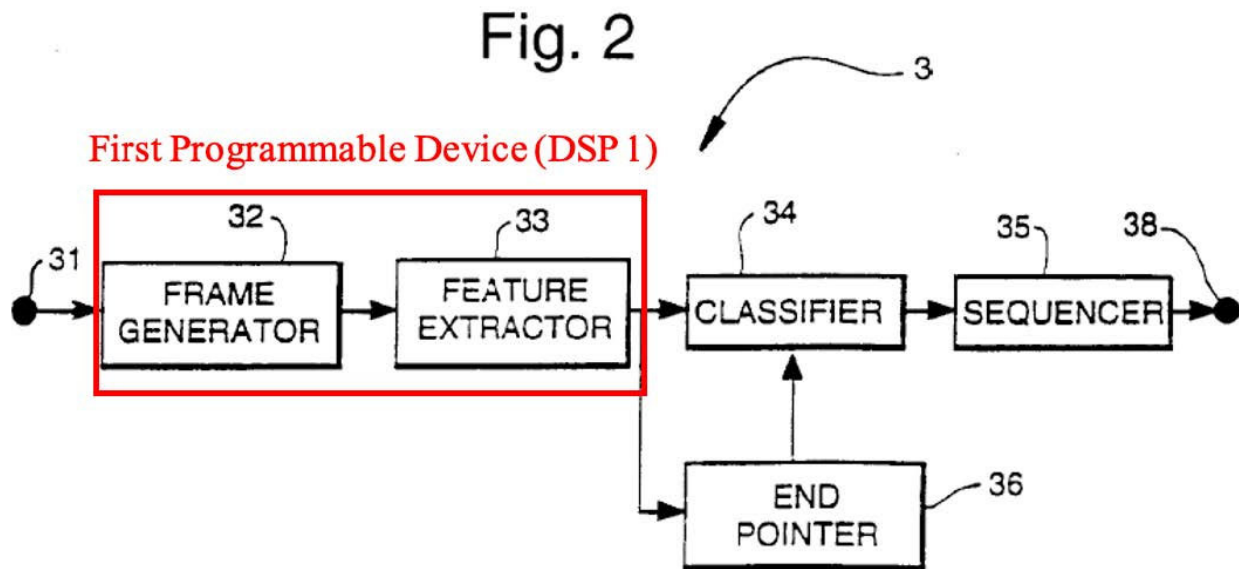
modules, such as programmed conventional array processor, for performing the disclosed tasks. *Dec.*, 111-114.

Therefore, *Jiang* teaches or renders obvious Claim 1(a). Per Mr. Schmandt, the acoustic signal output by the microphone, sampled by the A/D converter, divided into frames comprising a plurality of digital samples by the feature extraction module 66, and encoded into feature vectors (or codewords) by the feature extraction module 66, teaches the claimed calculating a feature vector from a digital audio stream. *Dec.*, 107-114. Because the feature extraction module 66 is a conventional array processor, and processors are programmed, *Jiang* further teaches a first programmable device.

4. *Jiang in Combination with Smyth Teaches Claim 1(a)*

As discussed in Section VI.B.3, *Jiang* teaches the programmed functionality of calculating a feature vector recited in Claim 1(a). To the extent Patent Owner asserts *Jiang* does not teach a first programmable device programmed to perform the functionality recited in Claim 1(a), *Smyth* teaches such.

Smyth teaches a “first programmable device” programmed to calculate a feature vector, as claimed, namely “a single suitably programmed digital signal processor (DSP) device...or similar device” providing frame generator 32 and feature extractor 33. *Smyth*, 5:49-53.



Smyth, Fig. 2 (annotated). *Smyth* expressly states a DSP is “programmed,” and a POSITA would have readily known DSPs are programmable. *Dec.*, 71-72, 115-119 (citing *Newton* (EX1008), *Kloker* (EX1009) establishing that a DSP is programmed).

Smyth also teaches that its programmable device is *programmed* to calculate a feature vector of a digital audio stream. *Smyth* discloses “an input 31 for receiving speech in digital form (either form [*sic*] a digital network or from an analog to digital converter).” *Smyth*, 5:2-3. *Smyth* refers to the inputted and digitized speech as “the succession of digital samples” that are then generated into frames of contiguous samples by frame generator 32. *Smyth*, 5:3-7, 5:19-29; *Dec.*, 117-119 (opining the succession of digital samples is a “digital audio stream” and further noting *Smyth*, 1:24-39 teaches digitizing speech data is “common” in speech recognition processing). Thus, *Smyth* discloses a “digital audio stream,” as claimed.

Smyth teaches the digital audio stream is passed through frame generator 32, which receives speech samples at the well-known 8 kHz rate, and forms frames corresponding to 16 ms increments of speech. *Smyth*, 1:30-31, 5:20-23. Feature extractor 33 then receives these frames of digitized speech and **“generates, in each case, a set or vector of features.”** *Smyth*, 5:31-33, 5:6-7 (“a feature extractor 33 for generating from a frame of samples a corresponding **feature vector**”). Because the DSP is “suitably programmed” to generate a feature vector by utilizing feature extractor 33 (*see Smyth*, 5:49-53), a POSITA would have understood the DSP is *programmed* to calculate the feature vector. *Dec.*, 117-119.

Because (1) *Smyth* teaches a DSP, (2) DSPs were well-known to be programmed, and (3) Applicant admitted during original examination that the claimed “first programmable device” was satisfied by a DSP, *Smyth* teaches a programmable device as recited in the claims, including Claim 1(a).

a) Motivation to Modify *Jiang* to Include *Smyth*’s First Programmable Device

A POSITA would have been motivated and found it obvious to modify *Jiang* to include *Smyth*’s first programmable device for calculating a feature vector. *Jiang* teaches its speech recognition system may be “practiced with other computer system configurations,” including a “multi-processor system” and “programmable consumer electronics.” *Jiang*, 4:60-65. Thus, a POSITA would have understood

Jiang contemplated modifications to its hardware architecture, including use of multiple processors. *Dec.*, 121-122.

Jiang also teaches a speech recognition system including feature extraction module 66 that is a program module of computer-executable instructions implemented in either hardware modules or software modules. *Jiang*, 4:55-57, 6:20-25, 6:31-32, 6:53-57. The feature extraction module 66 in *Jiang* modified to be a first programmable device as taught by *Smyth* (i.e., *Smyth*'s DSP calculating feature vectors) already performs the claimed calculating a feature vector and would entail mere hardware or software changes well known in the art. *Dec.*, 121-122. Per Mr. Schmandt, because *Jiang* already recognizes speech recognition may be performed with various “computer system configurations” (*Jiang*, 4:60-65), modifying *Jiang* to include a first programmable device for performing the feature vector calculation taught in *Jiang* would have had a reasonable expectation of success and would have yielded predictable results. *Dec.*, 121-122.

C. Claim 1(b)

1. *The '377 Patent's Discussion of “extracted” or “derived quantities”*

The '377 Patent does not use the terms “extracted” or “derived quantities,” except in repeating back the claim language. '377 Patent, 5:65–6:1, 6:20-25, 6:34-41, 6:51-54, 6:62-64. However, it describes calculating a feature vector “by splitting the audio signal into **spectral components**.... [T]he audio input signal is split into

12 frequency bands and one energy band, and a feature vector is constructed using these 13 components, plus their first and second derivatives[.]” *Id.*, 12:57-62. Per Mr. Schmandt, it is well-known that an audio signal includes spectral components or characteristics split among frequency bands. *Dec.*, 123-126, 109-111, 28-33, 35-41.

2. *Jiang’s Teachings*

Jiang teaches the calculated feature vector comprises extracted and/or derived quantities in the form of “spectral characteristics”: “[t]he frames are then preferably encoded by feature extraction module 66 into a **feature vector** reflecting **the spectral characteristics for a plurality of frequency bands.**” *Jiang*, 6:62–7:10, 6:53-57. Like the ’377 *Patent*, 12:57-62, *Jiang* also discloses computing a magnitude value for “each frequency band of a **frequency spectrum.**” *Jiang*, 6:53-57. *Jiang’s* feature extraction module 66 thus *extracts* quantities that are encoded into a feature vector and are indicative of the spectral characteristics of the audio signal. *Dec.*, 125-126.

Jiang further teaches the feature vector comprises extracted and/or derived quantities of the digital audio stream “during a defined audio time frame[.]” as claimed. *Jiang* teaches the digital signal is divided into frames of approximately 10 milliseconds in duration. *Jiang*, 6:62-65. These frames are then encoded by the feature extraction module 66 into a feature vector, where the feature vector

“reflect[s] the spectral characteristics[.]” *Jiang*, 6:65–7:1. Therefore, as is well-known in the art, each frame represents 10 milliseconds of speech input that is encoded as a feature vector. *Dec.*, 125-126.

Because *Jiang* teaches a frame of audio input of 10 ms in duration, i.e., a “defined audio time frame[.]” and further because *Jiang* teaches encoding, using the feature *extraction* module (which is mapped as the “first programmable device”), a feature vector of this frame that includes *extracted* quantities from the digital audio stream, *Jiang* teaches Claim 1(b).

D. Claim 1(c)

1. Summary of Claim 1(c) Mapping

Jiang in combination with *Smyth* teaches or renders obvious Claim 1(c). Both *Jiang* and *Smyth* teach the claimed functionality of calculating distances. *Smyth* teaches three distinct programmable devices, each in the form of a DSP. The proposed mapping for Claim 1(c) relies on (1) modifying *Jiang* to include the second programmable device of *Smyth*, where *Jiang* is relied on for teaching the claimed functionality of calculating distances; or additionally, or alternatively, (2) modifying *Jiang* to include the second programmable device of *Smyth*, where *Smyth* is relied on for teaching the second programmable device performing the claimed functionality of calculating distances.

2. '377 Patent's Discussion of "distance"

The '377 Patent describes "distances" as "probabilities" or "likelihoods":

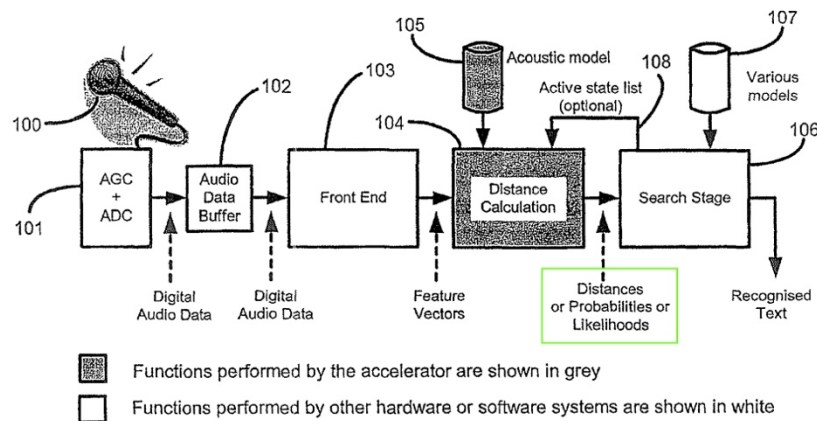


Figure 16

'377 Patent, Fig. 16 (annotated), 4:9-19 (also describing distances as representing "the likely correspondence of feature vectors to states in an acoustic model"). *Dec.*, 134-136. It describes "**a distance is an indication of the probability or likelihood** that a feature vector corresponds to a particular state. An important class of distance computation in speech recognition is the calculation of **output state probabilities** in recognisers using **Hidden Markov Models.**" '377 Patent, 4:38-42, 4:15-17, 13:24-26 ("The distance calculator stage of the recognition process computes a probability or likelihood that a feature vector corresponds to a particular state."). It describes "a wide variety of distance metrics used" that are "widely documented in the literature." *Id.*, 4:15-19.

Per Mr. Schmandt, the '377 Patent's description of a "distance" as a probability or likelihood of a feature vector compared to a particular state of an HMM is consistent with the well-known use of HMMs in speech recognition. *Dec.*, 134-136. Specifically, for each state of HMM, a distance (i.e., a probability or likelihood) is computed, such that a distance indicates the similarity of the spoken speech to the state sequence or path. *Id.*

3. *Jiang's Teachings*

Jiang teaches a processor "programmed to calculate distances ...," as claimed. The limitations of an acoustic model, plurality of acoustic states, and a feature vector are mapped first, for ease of reference.

a) "a plurality of acoustic states of an acoustic model"

Regarding the claimed "acoustic model," *Jiang* teaches utilizing HMMs as "phonetic speech unit models...which represent speech units to be detected by system 60." *Jiang*, 7:30-47. Per Mr. Schmandt, a POSITA would have understood *Jiang's* HMMs (which *Jiang* describes as "phonetic speech unit models") are "acoustic models," as claimed. *Dec.*, 137-138.

First, HMMs were well-known to be acoustic models. *Dec.*, 137-138 (*citing Ravishankar*, 19, 27-28). Second, *Jiang* expressly discloses the HMMs are **phonetic speech unit** models, and a POSITA would have understood **phonetic speech** is **acoustic**. *Dec.*, 137-138. Third, the '377 Patent describes HMMs as acoustic models.

'377 *Patent*, 38:29-31; *Dec.*, 137-138. Therefore, *Jiang* discloses an “acoustic model,” as claimed.

Jiang also teaches each HMM includes a “plurality of acoustic states[,]” as claimed: “A hidden Markov model includes a plurality of states” *Jiang*, 1:21-22, 1:43-48; *Dec.*, 137-139 (*citing Ravishankar*, 28 (“An HMM is a set of states connected by transitions”)). *Jiang*, 7:30-47 teaches a distance calculation process employing HMMs. *See* Section V.D.3.c. *Jiang* also teaches the discussion of speech recognition discussed at 7:30–8:51 is “all done in a known manner.” *Jiang*, 8:51. Therefore, per Mr. Schmandt, a POSITA would have understood *Jiang*’s distance calculation (discussed further in Section V.D.3.c) determines the similarity between a feature vector and a “plurality of states” (including acoustic states because modeling spoken speech) of an acoustic model, such as the HMM states discussed at *Jiang*, 7:30-45. *Dec.*, 137-139.

b) “a feature vector”

As discussed in Section V.B.3, *Jiang* teaches encoding the frames of spoken speech into a feature vector. *Jiang*, 6:62–7:10 (“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.”); *Dec.*, 140. *Jiang* also teaches performing an additional step for some embodiments of “encod[ing] the feature vectors into one or more codewords using vector quantization techniques and a

codebook derived from training data.” *Jiang*, 7:1-10. Per Mr. Schmandt, vector quantization is a well-known technique to reduce computation requirements and the search space by “represent[ing] the various cepstral features in an efficient manner” and clustering feature vectors into similar codewords. *Dec.*, 140-141 (*citing Brown*, 10:24-34, 3:53-65, and *Ravishankar*, 48-49, and further explaining that a codeword is an index into a codebook that stores a value representative of a cluster of data in its original form, i.e., a representative feature vector); *'377 Patent*, 35:36-37, 5:6-8 (discussing codebooks).

Jiang describes the speech recognition as determining the phoneme(s) represented by the *codewords*. *Jiang*, 7:38-43. A POSITA would have understood that because the codewords are vector quantized feature vectors, then *Jiang*’s teaching of determining the phonemes of the *codewords* is likewise determining the phonemes of the *feature vectors*. *Dec.*, 140-141. Thus, *Jiang*’s teaching of determining the likelihood that a codeword corresponds to a phoneme teaches determining the likelihood that a “feature vector” corresponds to the same phoneme, because the codewords are encoded feature vectors. *Id.*

c) “calculat[ing] distances...”

Jiang teaches its tree search engine 74 is programmed to “calculat[e] distances ...,” as claimed, namely a *score* for each phoneme branch of the prefix tree, where the “**score** represents the **likelihood** that the particular **phoneme** encoded

by the **codeword** corresponds to the phoneme for the branch under consideration.”

Jiang, 8:29-33. *Jiang* describes how the score for each phoneme branch is determined:

Upon receiving the codewords from feature extraction module 66, and the boundary detection signal provided by silence detection module 68, tree search engine 74 accesses information stored in the phonetic speech unit model memory 72. Memory 72 stores phonetic speech unit models, such as hidden Markov models, which represent speech units to be detected by system 60. In one preferred embodiment, the phonetic models stored in memory 72 include HMMs which represent phonemes. Based upon the HMMs stored in memory 72, tree search engine 74 determines a most likely phoneme represented by the codeword received from feature extraction module 66, and hence representative of the utterance received by the user of the system.

Jiang, 7:30-47; *Dec.*, 142.

Jiang provides a discussion in its Background of using HMMs to calculate a distance indicating a similarity between a feature vector and a plurality of HMM states, where the distance in *Jiang* is referred to as a “score” or “probability score.” *Jiang*, 1:19–2:15; *Dec.*, 145. For example, *Jiang* teaches utilizing HMMs to “model observed sequences of speech spectra, where specific spectra are **probabilistically associated with a state in an HMM.**” *Jiang*, 1:41-48. *Jiang* also teaches a “**probability score is computed for each state in the state sequences (i.e., the**

HMMs) being considered.” *Jiang*, 2:5-7; *see also* Section V.D.2 (discussing the ’377 Patent describing a “distance” as “an indication of the probability or likelihood that a feature vector corresponds to a particular state”). Thus, *Jiang*’s collective teachings indicate to a POSITA that *Jiang*’s “score [] indicative of a likelihood that the particular phoneme being examined is the actual phoneme indicated by the codeword received from the feature extraction module” comprises a “distance” as recited in Claim 1(c). *Jiang*, 8:20-23; *Dec.*, 142-145 (further explaining how *Jiang*’s Background provides additional teachings informing a POSITA that the calculated “score” is a probability or likelihood representative of a similarity or distance measurement). Therefore, *Jiang* teaches “calculating distances[,]” as claimed.

Jiang’s teaching of determining the likelihood a *codeword* is a particular phoneme would also have been understood by a POSITA to teach calculating distances indicating the similarity between a *feature vector* and HMM states. *Dec.*, 146-147. As discussed for Claim 1(a), *Jiang* teaches encoding the frames of spoken speech as feature vectors and, in some embodiments, encoding the feature vectors as codewords using vector quantization techniques. *Jiang*, 6:64–7:10, 1:58-67. The teachings at *Jiang*, 7:38-43 and 8:29-36 both refer to determining the likelihood of a phoneme represented by a *codeword*. As discussed for Claim 1(a), because the codeword is simply a vector quantized feature vector, a POSITA would have understood *Jiang*’s teachings of determining the likelihood the considered phoneme

is indicated by the *codeword* to likewise be determining the likelihood the considered phoneme is indicated by one or more *feature vectors*. *Dec.*, 109, 140-148. Thus, a score indicative of the likelihood that a codeword corresponds to the phoneme for the branch under consideration is a distance calculation (i.e., score indicative of likelihood) between a feature vector (i.e., codeword) and a plurality of predetermined acoustic states of an acoustic model (i.e., states of HMMs).

Jiang, 7:30-45 and 8:16-51 would also have been understood to teach determining the similarity between the feature vector (or codeword) and “a plurality of acoustic states of an acoustic model[.]” As quoted above, *Jiang* teaches a “**probability score is computed for each state in the state sequences (i.e., the HMMs)**” being considered.” *Jiang*, 2:5-7; *Dec.*, 144-147 (opining because *Jiang* teaches computing a probability score *for each state* of an HMM, *Jiang* teaches computing a probability score for a plurality of states of an acoustic model). HMMs are used to determine the most likely phoneme represented by the codeword and hence representative of the user’s utterance. *Jiang*, 7:37-42. HMMs include a plurality of states, and HMMs are acoustic models. See Section V.D.3.a. A score is computed for each phoneme. These scores are then used to indicate the likelihood that the codeword corresponds to the branch under consideration in the lexical (“prefix”) tree. *Jiang*, 8:16-42. Because *Jiang*, 7:30-45 teaches HMMs are used to determine the most likely phoneme from a feature vector, *Jiang* teaches the

calculated score indicates the similarity (referred to in *Jiang* as the probability or likelihood) between at least one feature vector (the at least one codeword) and the HMM states. *Dec.*, 144-147.

Jiang further teaches or renders obvious calculating distances (plural). As quoted above, *Jiang* teaches a “**probability score is computed for each state in the state sequences (i.e., the HMMs)** being considered.” *Jiang*, 2:5-7.

Further, *Jiang* teaches calculating a score for each respective node. *Dec.*, 148 (explaining that because *Jiang*, 8:29-33 teaches computing a score for **each** phoneme branch under consideration, then each phoneme branch is scored, resulting in calculating a plurality of distances, and discussing how *Jiang*’s threshold teachings at 11:1-15 further teach calculating more than one distance). Therefore, for each utterance (which is in turn encoded as a feature vector), a score is computed for each node indicating the likelihood the phoneme for the node is the phoneme encoded by the codeword. *Jiang*, 8:16-32; *Dec.*, 145-148. Because there are multiple nodes in the prefix tree (*see, e.g., Jiang*, Fig. 3), there are multiple scores, or distance calculations computed. *Dec.*, 145-148.

Jiang teaches a processor *programmed* to calculate distances because tree search engine (which may be a dedicated processor, per 6:39-42) determines the most likely phoneme discussed at 7:30-47 and 8:29-33. Because the dedicated processor performs the distance calculation, the processor is *programmed* to

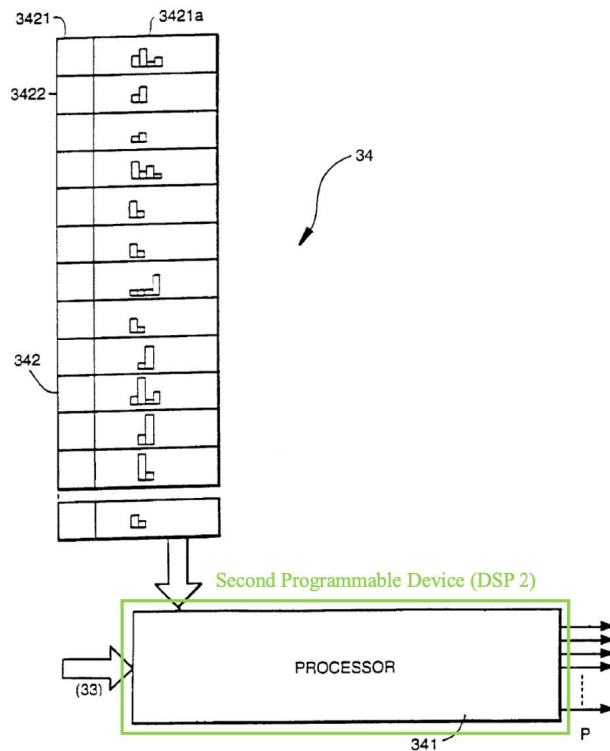
calculate distances. The proposed modification to *Jiang* divides the tree search engine's functionality into two processors (or DSPs), discussed at Section V.D.5.

4. *Smyth's Teachings*

a) “a second programmable device programmed to”

Smyth teaches “a second programmable device programmed to calculate distances[,]” as claimed, namely classifying (or classification or classifier) processor 341, which calculates state probabilities indicating the similarity that a feature vector corresponds to each acoustic state in state memory 342. *Smyth*, 5:55–6:4; *Dec.*, 149-156. *Smyth* teaches “classifying processor 341 may be a suitably programmed **digital signal processing (DSP)** device, may in particular be the same digital signal processing device as the feature extractor 33.” *Smyth*, 6:5-8. Figure 3, below, illustrates the classifier processor 341 as a “processor”:

Fig. 3



Smyth, Fig. 3 (annotated).

Per Mr. Schmandt, *Smyth* teaches three distinct DSPs for performing the three general speech recognition steps. *Dec.*, 150-155. *Smyth* teaches utilizing a first DSP for the frame generator 32 and feature extractor 33, a second DSP for the classifier processor 341 of classifier 34, and a third DSP (or microprocessor) for the parsing processor 351 of sequencer 35. *Smyth*, 5:49-53, 6:5-8, 6:47-51. Per Mr. Schmandt, *Smyth*, 5:49-53 is informing the reader of suitable exemplary DSP models, such as the Motorola DSP56000, for the single DSP for the frame generator and extractor. *Smyth*, 6:5-8 then discusses the DSP for the classifier processor 341, where the DSP

is also a suitably programmed DSP and may be the same DSP device as the feature extractor's DSP. Rather than repeating suitable model types (e.g., Motorola DSP56000), *Smyth* is informing the reader the DSP for the classifier processor 341 may be the same model of DSP device as for the single DSP for the frame generator and feature extractor. *Dec.*, 151-152. Mr. Schmandt notes *Smyth*, 6:47-50 further informs a POSITA's understanding, as the processing element for the parsing processor 351 is discussed as yet a different processor, namely a microprocessor (including identifying exemplary microprocessors by tradename) or the "same DSP device as is employed for any of the preceding processors." *Id.* Therefore, a POSITA would have understood *Smyth* teaches three distinct DSP devices of the same exemplary model types may be used, one DSP for *each* of the collective feature extractor and frame generator, the classifier processor 341, and the parsing processor 351.

Smyth teaches a second *programmable* device because DSPs were well-known programmable devices, and Applicant admitted a DSP was a programmable device. *Dec.*, 71-72, 93-96, 156.

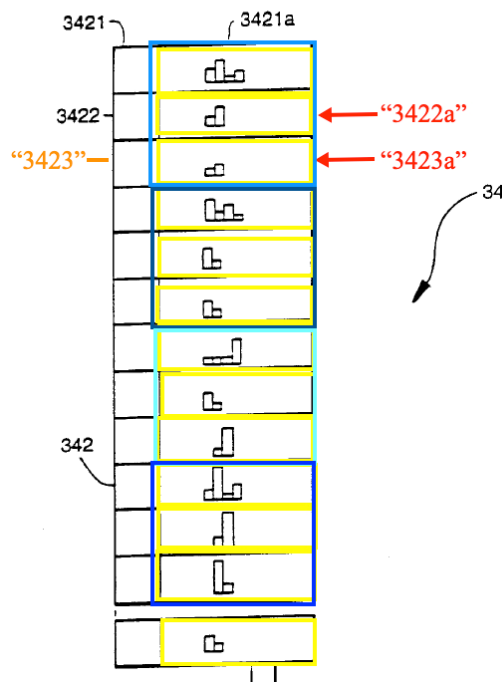
Section V.D.5 discusses a motivation to modify *Jiang* to include *Smyth's* distinct DSP for performing distance calculation.

b) “a plurality of acoustic states”

Smyth teaches the claimed “plurality of acoustic states[,]” namely speech states stored in state fields of state memory.

Smyth teaches classifier 34 comprises classifier processor 341 and state memory 342. *Smyth*, 5:55-57, Fig. 3. “The state memory 342 comprises a state field 3421, 3422, ...” *Smyth*, 5:58-59. Each state field stores a speech state. *Smyth*, 5:58-59 (“a state field 3421, 3422..., **for each of the plurality of speech states.**”). *Smyth*’s Fig. 3 identifies by reference numeral only two fields (3421, 3422) and one speech state (3421a). Per Mr. Schmandt, a POSITA would have understood the cells following state fields 3421-3422 are similarly state fields, and that each state field stores a speech state similar to speech state 3421a. *Dec.*, 157-158 (*citing Smyth*, 5:58-59 and noting *Smyth*’s use of ellipses that would also indicate similar state fields with respective speech states). Annotated Fig. 3 below depicts state memory 342, state fields 3421–3422 (with state field 3423 added and annotated orange), speech state 3421a (with speech states 3422a-3424a added and annotated red), and groups of three speech states corresponding to an allophone (annotations added in shades of blue):

Fig. 3



Smyth, Fig. 3 (annotated). *Smyth* does not identify element 3421a in its written description (although 3421a is identified in Fig. 3). However, *Smyth* teaches a “plurality of speech states” with each state field having a speech state. *Smyth*, 5:58-62. Consequently, a POSITA would have understood that element 3421a was a “speech state” of state field 3421, and added elements 3422a, 3423a were correspondingly “speech states” of respective state fields 3422. *Dec.*, 157-162; *Smyth*, 5:58-62. Therefore, *Smyth* teaches a plurality of states.

Smyth expressly teaches the speech states are “acoustic” states: “each allophone to be recognised by the recognition processor comprises 3 states, and accordingly 3 state fields are provided in the state memory 342 for each allophone.”

Smyth, 5:60-62. An allophone is well-known as an “acoustic” subword unit, namely alternative pronunciations of phonemes. *Dec.*, 159-162.

c) “[plurality of speech states] of an acoustic model”

Regarding the claimed “acoustic model,” *Smyth* teaches or renders such obvious. *Smyth* teaches a “template or model,” including a “plurality of model states,” that are used to determine a measure of similarity and recognition results for spoken speech. *Smyth*, 1:34-39, 5:7-9. *Smyth*’s Background discloses:

A **stored representation of the word or phrase**, known as a template or **model**, comprises a **reference feature matrix** of that word as previously derived from, in the case of speaker independent recognition, multiple speakers. The **input feature vector is matched with the model** and a **measure of similarity** between the two is **produced**.

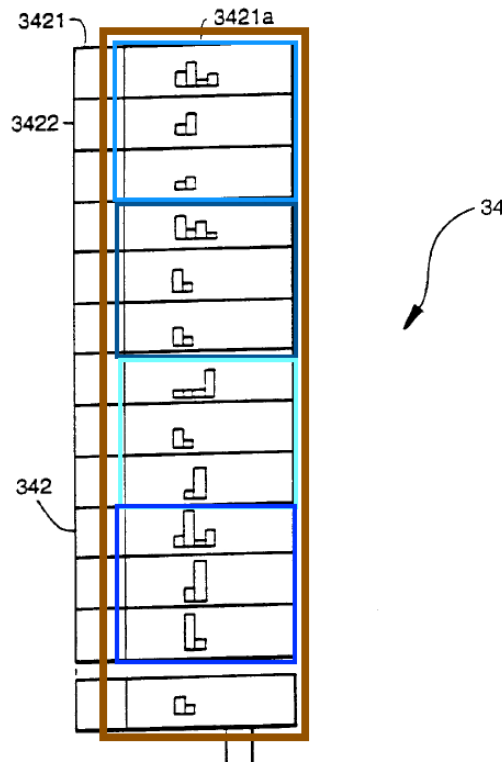
Smyth, 1:34-39. *Smyth* further teaches “classifier 34 receiv[es] the succession of feature vectors and operat[es] on each with a **plurality of model states**, to generate recognitions results[.]” *Smyth*, 5:6-9.

Mr. Schmandt explains that it was well-known to train the speech recognition system to generate models or templates of spoken input, as otherwise taught by *Smyth*, 1:34-39. *Dec.*, 163 (citing *Ney* (EX1053), p. 4 (teaching estimating probabilities of acoustic vectors “during the training phase of the recognition system”), *Jiang*, 11:25-29, *Ravishankar*, 19-20, 27-29 (discussing training acoustic

models in subword units, such as phonemes)). Per Mr. Schmandt, *Smyth*’s “plurality of model states” is an *acoustic model*, where reference to the states of the model (i.e., *Smyth*’s speech states at 5:58-62) is used to determine the likelihood that a feature vector corresponds to one of the speech states. *Dec.*, 163-165. For example, *Smyth* teaches calculating probabilities for each state in state memory 342, where each probability indicates this likelihood. *Smyth*, 5:63–6:4. Additionally, the recognition system “operat[es] on each [feature vector] with a plurality of model states,” further indicating the plurality of model states are an acoustic model used to determine probabilities. *Smyth*, 5:7-9. Thus, by *referencing* the states, i.e., the disclosed “plurality of model states,” feature vector probabilities are calculated to identify allophones. *Id.*

Smyth further teaches words are determined by stringing together identified allophones. *Smyth*, 6:15-19; *Dec.*, 164-165, 202. The stored representation of the word or phrase, i.e., the “template or model” per *Smyth*, 1:34-39, is the “string of allophones” forming the word. *Smyth*, 6:15-19. Annotated Fig. 3 below depicts *Smyth*’s allophones (annotated shades of blue) as a collection of acoustic states (annotated brown).

Fig. 3



Smyth, Fig. 3 (excerpt, annotated). This collection of speech states (where three speech states are provided for each allophone) teaches or renders obvious the claimed “acoustic model.” *Dec.*, 164-165.

d) “[programmed to] calculate distances indicating the similarity...”

Smyth discloses the second programmable device (i.e., classifier processor 341) is programmed to calculate distances indicating the similarity between a feature vector (i.e., the feature vector calculated in feature extractor 33) and a plurality of acoustic states (i.e., the speech states in state fields 3421, 3422, etc.) of an acoustic

model (i.e., the collection of allophones comprising speech states in state memory 342). *Dec.*, 166-169.

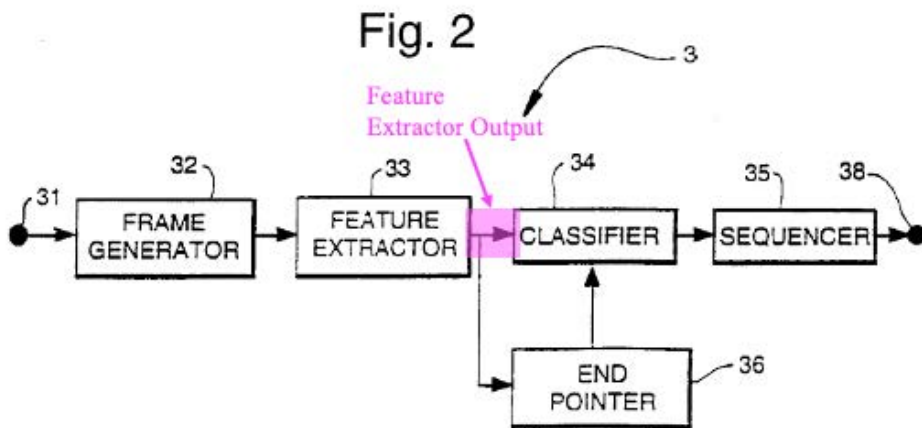
Smyth teaches calculating distances for indicating the similarity, as claimed.

Smyth teaches:

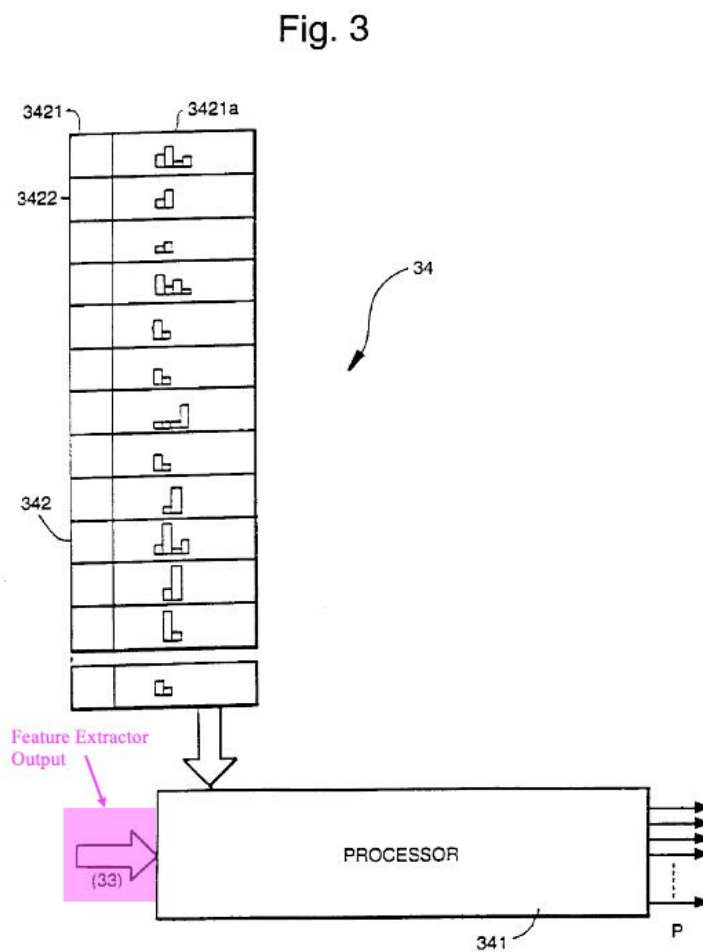
The **classification processor 34[1]** is arranged to **read each state field** within the memory 342 in turn, and ***calculate for each***, using the current input **feature coefficient set**, the ***probability*** that the **input *feature* set or *vector* corresponds to the corresponding state**. [¶]

Accordingly, the **output** of the classification processor is ***a plurality of state probabilities P, one for each state*** in the state memory 342, ***indicating the likelihood that the input feature vector corresponds to each state***.

Smyth, 5:63–6:4. *Smyth* further discloses “a classifier 34 **receiving the succession of feature vectors and operating on each with a plurality of model states**.” *Smyth*, 5:7-9, 5:55-57 (disclosing “the classifier 34 comprises a classifier processor 341 and a state memory 342”). *Smyth*’s Figs. 2-3 depict the output of feature extractor 33 as the input to classifier processor 341.



Smyth, Fig. 2 (annotated)



Smyth, Fig. 3 (annotated).

Per Mr. Schmandt, *Smyth* teaches or renders obvious the claimed “calculate[ing] distances.” *Dec.*, 166-169. *Smyth*’s calculated probabilities *P* is a “distance,” as claimed, because the probability indicates the likelihood, i.e., similarity, between the feature vector and the stored acoustic state. *Dec.*, 166-169; *Smyth*, 5:63–6:4. *Smyth* teaches calculating a *plurality* of probabilities (i.e., distances) for a “*plurality* of predetermined acoustic states” because a probability is calculated for “each state” of the acoustic model. *Dec.*, 166-169; *Smyth*, 6:1-4 (teaching calculating a plurality of probabilities “**one for each state** in the state memory”). The probability (i.e., distance) is for a “*plurality of predetermined acoustic states of an acoustic model*” per Sections V.D.4.b-c, which discusses *Smyth*’s model is an acoustic model comprising “a plurality of model states,” namely “speech states.” *Smyth*, 1:34-37, 5:9, 5:59; *Dec.*, 157-165.

Smyth’s classifier processor 341 (including in the form of a DSP, per *Smyth*, 6:5-8) is *programmed to* calculate distances because the processor 341 receives the feature vector from feature extractor 33 and calculates the probability indicating the likelihood the feature vector corresponds to the state. *Dec.*, 166-169; *Smyth*, 5:7-9 (discussing classifier receiving feature vectors and generating recognitions results by operating on a plurality of model states). Because the classifier processor 341 is a DSP (*Smyth*, 6:5-8), and DSPs are *programmed* devices (see Section V.D.2), *Smyth* teaches the classifier processor 341 (the claimed “second programmable device” in

the modified *Jiang*) is programmed to perform the step of calculating distances described in detail below. *Dec.*, 166-169.

5. *Motivation to Modify Jiang to Include Smyth's Second Programmable Device*

A POSITA would have found it obvious and been motivated to modify *Jiang's* tree search engine 74 to include a second processor/DSP, as taught by *Smyth*, or to otherwise perform *Jiang's* distance calculation using a processor/DSP distinct from the tree search engine's processor. *Dec.*, 170-182.

Jiang already teaches that its invention “may be practiced with other computer system configurations,” including “microprocessor-based or programmable consumer electronics.” *Jiang*, 4:60-65. *Jiang's* CPU 21, which includes tree search engine, “may include one or more processors.” *Jiang*, 6:39-42. *Jiang* repeatedly refers to distinct processors for performing distinct speech recognition functionality, indicating to a POSITA that *Jiang* envisions dedicated processors for performing specific speech recognition functionality. *See, e.g., Jiang*, 6:53-57 (stating feature extraction module 66 is a “conventional array processor”), 7:19-21 (stating silence detection module 68 “can either be implemented on the same, or a different processor as that used to implement the feature extraction module 66”); *Dec.*, 170-171.

It is known that calculating distances or probabilities is a computationally-intensive processing task. *Jiang*, 2:19-24; *Baumgartner*, [0003], [0005], [0007];

Dec., 171-173. *Jiang* reduces the computation needed via pruning, which “eliminates the need to compute the probability score for state sequences [i.e., words] that are very unlikely.” *Jiang*, 2:25-29. However, a known problem arises with pruning, as a threshold cut-off value must be chosen to “prune” state sequences that are below the minimum threshold value, which necessarily includes setting “a desired error rate.” *Jiang*, 2:43-46. Per Mr. Schmandt, pruning lessens the computational load but also potentially lessens the accuracy of the speech recognition. *Dec.*, 174-175 (citing *Jiang*, 2:25-46, 11:8-24). *Jiang* addresses the known issues with pruning by utilizing a lexical tree (i.e., the prefix tree). *Jiang*, 2:47-51; *Dec.*, 175 (citing *Ravishankar*, 33). But, even *Jiang* recognizes that utilizing lexical trees does not fully solve the issue of large computational loads (and decreased accuracy), as pruning suffers from setting a threshold level and consequently, having higher/lower number of hypotheses based on the level. *Jiang*, 11:8-24; *Dec.*, 175. *Jiang* thus gains efficiency by pruning the tree at a higher threshold level, but even *Jiang* recognizes the tree pruning is at the cost of increased accuracy. Accordingly, there is a need to balance efficiency while still having increased accuracy.

Modifying *Jiang* to decrease the computational load via a distinct processor for distance calculation helps achieve the balance between efficiency and accuracy. *Dec.*, 176-180. By employing distinct processors (e.g., DSPs as taught by *Smyth*), to

perform the various speech recognition steps, each DSP can be a dedicated processor, as urged by *Jiang*, 6:39-42. Providing discrete processors for feature extraction, distance calculation, and searching allows for segmentation of higher-load processing, resulting in a faster, more efficient, and more accurate speech recognition process, which is a goal of speech recognition. *Dec.*, 176-180 (*citing Ravishankar*).

There would have been a reasonable expectation of success in the proposed modification. Per Mr. Schmandt, *Jiang* suggests to a POSITA that the hardware and software architecture and configuration for performing the described functionality is amenable to design and use considerations, indicating to a POSITA the routine software change to the *Jiang* system and the additional processor would be well within *Jiang*'s contemplated solution. *Dec.*, 178-180 (*citing Jiang*, 4:60–5:3, which discusses that “those skilled in the art will appreciate that the invention may be practiced with other computer system configurations”).

6. *Motivation to Modify Jiang to Include Smyth's Second Programmable Device Performing Distance Calculations*

To the extent Patent Owner argues that *Jiang* does not teach calculating distances, a POSITA would have found it obvious and been motivated to modify *Jiang* to replace its determining probability scores for each phoneme branch with *Smyth*'s probability calculation that an input feature vector corresponds to a corresponding state. *Jiang*, 8:29-33; *Smyth*, 5:63-67; *Dec.*, 183. *Jiang* already

provides the background architecture for performing distance calculations, including utilizing HMMs to identify a “most likely phoneme represented by the codewords received from feature extraction module.” *Jiang*, 7:36-42. Therefore, implementing a distance calculation as taught by *Smyth* would have had a reasonable expectation of success, as adding a DSP performing *Smyth*’s distance calculation would have resulted in more efficient processing. *Dec.*, 170-183; Section V.D.5. Performing distance calculation to determine the most likely phoneme would have entailed use of a known technique, i.e., calculating distances by comparing a feature vector to a state of an acoustic model, such as an HMM, to improve similar devices, i.e., *Jiang*’s speech recognition system. *Dec.*, 183. Instead of performing *Jiang*’s “most likely phoneme” based on stored HMMs, it would have been obvious to substitute *Smyth*’s probability calculation for an input feature vector compared to a state, as such as substitution of one known element for another would have obtained predictable results. *Dec.*, 183. Because distance calculation was so well-known and commonly used (*Ravishankar*, 27-29), the simple substitution would have been readily implemented by a POSITA. *Id.*

E. Claim 1(d)

1. *Jiang*’s Teachings

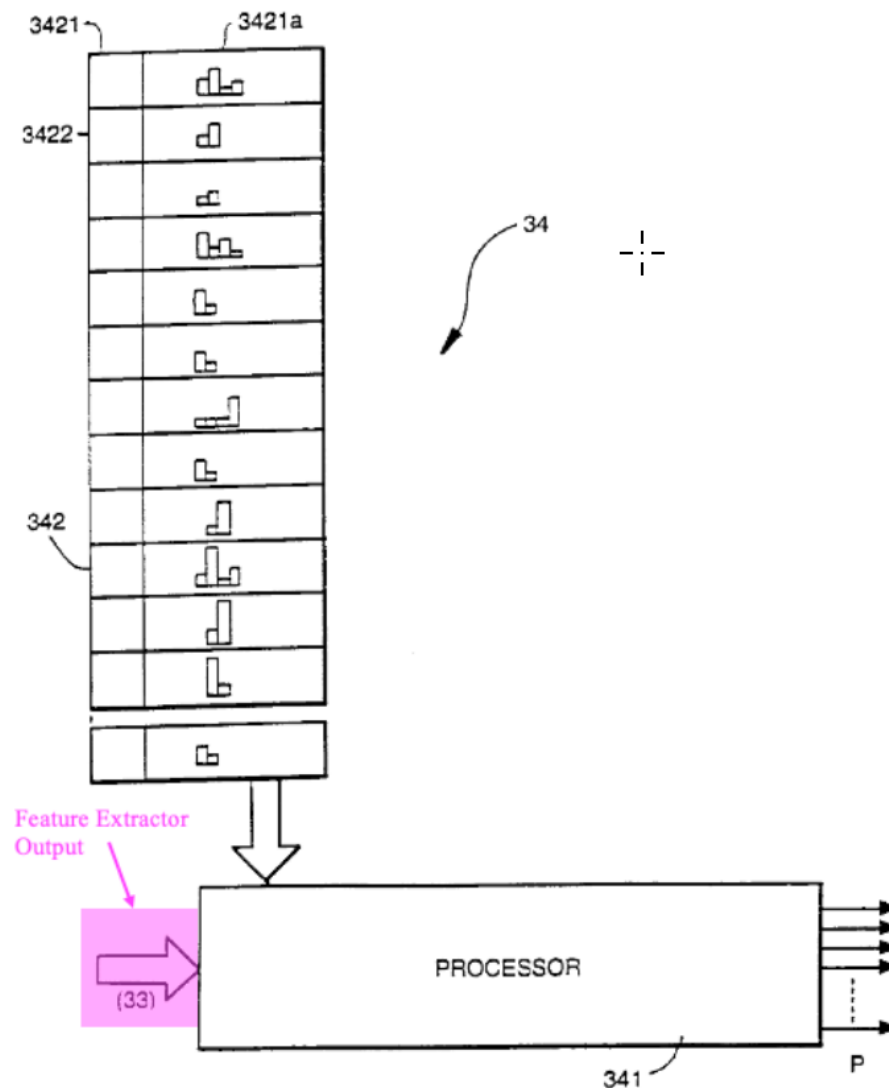
Jiang teaches Claim 1(d). *Jiang* teaches that “[u]pon receiving the codewords from feature extraction module 66, ..., tree search engine 74 accesses information

stored in the phonetic speech unit model memory 72.” *Jiang*, 7:30-34. The stored information includes HMMs, and “[b]ased upon the HMMs stored in memory 72, tree search engine 74 determines a most likely phoneme represented by the code word **received from feature extraction module 66**” *Jiang*, 7:34-42. Therefore, the feature vector (i.e., the codeword) is received by the tree search engine (modified to be the claimed “second programmable device,” per Section V.D) **after** it is calculated by the feature extraction module 66 (i.e., the claimed “first programmable device,” per Section V.B). *Dec.*, 184.

2. *Smyth’s Teachings*

Smyth teaches “classifier 34 receiv[es] the succession of feature vectors [from feature extractor 33] and operat[es] on each with a plurality of model states, to generate recognition results.” *Smyth*, 5:7-9.

Fig. 3



Smyth, Fig. 3 (annotated). Thus, *Smyth* teaches the classifier processor 341 performing the probability (i.e., distance) calculation on each *input* feature vector for each state (5:63-67) *after* the feature vector is calculated by the feature extractor 33 (part of the first programmable device). *Dec.*, 185. In the modified *Jiang* including *Smyth*'s second programmable device for calculating distances, the

distances would be received after *Jiang*'s feature extraction module extracts the feature vector. *Dec.*, 185.

F. Claim 1(e)

Jiang alone or in combination with *Smyth* teaches or renders obvious Claim 1(e). The '377 Patent describes what *Jiang* discloses—phonemes are determined using HMMs, and phonemes are used to identify words. *Compare* '377 Patent, 11:40-42, 3:12-14, 3:25-29, 4:33-44, with *Jiang*, 7:30-43, 8:16-42; *Dec.*, 186-208. Additionally, *Smyth* discloses its “third programmable device” using HMMs for word identification.

1. *Jiang*'s Teachings

a) “a third programmable device”

Jiang teaches the claimed “third programmable device,” namely tree search engine 74 modified to only perform the word search. *Jiang* discloses tree search engine is “preferably implemented in CPU 21 (which may include one or more processors) or may be performed by a dedicated speech recognition processor employed by personal computer 20.” *Jiang*, 6:39-42. Therefore, tree search engine includes, at the least, a processor (i.e., either the processor of CPU or the dedicated speech recognition processor). *Dec.*, 187-188. Per Mr. Schmandt, a processor is well-understood to be “programmable” for performing the disclosed tasks, e.g., identifying spoken words. *Dec.*, 71-72, 111-120, 156, 187-188. And, as noted in the

mapping for Claim 1(a), Applicant admitted a CPU is programmable. *See Section VI.B.3.* Because tree search engine is implemented in CPU 21 or includes a dedicated processor, a POSITA would have understood engine 74 is implemented in a programmed device or that such is obvious. *Dec.*, 187-188.

Per the mapping for Claim 1(c), the functionality of the tree search engine is divided into two processors, with a first processor performing the distance calculations recited in Claim 1(c), and a second processor performing the word search recited in Claim 1(e). *See Sections V.E.2, V.E.6.* Thus, the claimed “third programmable device” is met by *Jiang’s* tree search engine modified to only perform the word recognition recited in Claim 1(e).

A POSITA would have found it obvious and been motivated to modify tree search engine 74 to include two distinct programmable devices for the reasons provided in Section V.D.5.

b) “programmed to identify spoken words in said digital audio stream using Hidden Markov Models and/or Neural Networks”

Jiang teaches identifying spoken words *using HMMs*, as claimed, where the HMMs are used to determine a score for a node and associated phoneme branch of the prefix tree. The score indicates the likelihood the phoneme being examined for the phoneme branch is the actual phoneme indicated by the codeword. *Jiang*, 8:16-

24, 7:30-43; *Dec.*, 189-194. As discussed for Claim 1(a), the codeword is indicative of the encoded “digital audio stream,” as claimed.

Jiang discloses utilizing a lexical tree, namely a “prefix tree,” to “arrive at the **word** most likely indicative of the utterance of the user.” *Jiang*, 7:65–8:4; *Dec.*, 189-194 (opining *Jiang*’s prefix tree is a lexical tree). *Jiang* teaches:

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**).

Jiang, 2:53-67; *Dec.*, 34, 49-59, 127-133, 189-194 (discussing organizing HMMs into a lexical tree is known). Per Mr. Schmandt, *Jiang* is teaching ***word identification*** is performed using the prefix tree, where the prefix tree comprises ***a plurality of branches each utilizing one or more HMMs for phoneme scoring.*** *Dec.*, 189-194.

The words from a lexicon 70, which is all words known to the system, are presented to tree search engine in the form of the prefix tree. *Jiang*, 7:65–8:4. Tree search engine 74 is *programmed* to perform the word identification, namely traversing the prefix tree and outputting the word or word sequence. *Jiang*, 7:61–8:15. The tree search engine uses (1) the stored lexicon 70; (2) the most likely phonemes calculated at 7:30-47; and (3) the prefix tree to “determine **a word** which

most likely represents the codewords received by feature extraction module 66.”
Jiang, 7:47-54, 8:16-24; *Dec.*, 189-194.

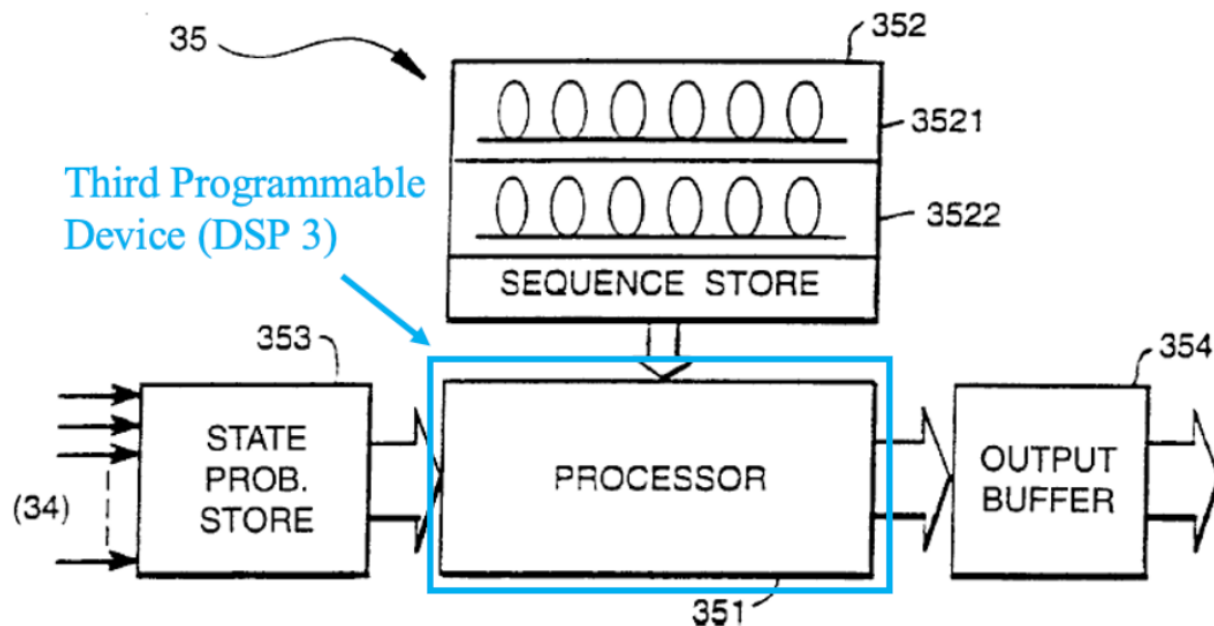
Therefore, the tree search engine (i.e., the claimed “third programmable device”) is programmed to identify spoken words (“to determine a word”) in the codewords (where the codewords are the encoded claimed “digital audio stream”). Because the tree search engine uses a prefix tree comprised of phoneme branches “associated with a particular phonetic model (such as an HMM),” *Jiang* teaches or renders obvious using HMMs to identify spoken words. *Jiang*, 2:53-56; *Dec.*, 189-194.

2. *Smyth’s Teachings*

a) “a third programmable device”

Smyth teaches a “third programmable device,” as claimed, namely parsing (or parser) processor 351 implemented as a DSP or microprocessor. *Dec.*, 195-196. Parsing processor 351 calculates the most likely path of states, compares this path with stored state sequences, and identifies **words** and phrases. *Smyth*, 6:36-60. *Smyth* teaches “parsing processor 351 may, for example, be a microprocessor ..., or may alternatively be a DSP device.” *Smyth*, 6:47-50. *Smyth’s* Figure 4 identifies parsing processor 351 as a processor.

Fig. 4



Smyth, Fig. 4 (annotated). As discussed for Claim 1(a), a POSITA would have understood a DSP is a programmable device. *Dec.*, 107-122, 195-200.

b) “programmed to identify spoken words in said digital audio stream”

Smyth teaches parsing processor 351 “identif[ies] spoken words in said digital audio stream,” as claimed. *Smyth* discloses a digital audio stream of spoken speech. *Smyth*, 5:1-5, 4:22-53 (discussing inputting spoken speech into the system). *Smyth* further teaches parsing processor 351 recognizes words and phrases in the digital audio stream. *Smyth*, 6:56-63. *Smyth* teaches a word recognition process wherein parsing processor 351 reads, at each frame of speech, both the newest state probabilities (i.e., the claimed calculated distances) from classifier processor 341

and the historic state probabilities stored in state probability memory 353. *Smyth*, 6:36-39. Parsing processor 351 then calculates the most likely path of states to date and compares this most likely path with each of the known state sequences stored in state sequence memory 352. *Smyth*, 6:39-42. Each known state sequence in state sequence memory 352 corresponds to a word, phrase, or other speech sequence to be recognized. *Smyth*, 6:52-53. *Smyth* teaches parsing processor 351 outputs a probability score at each frame for each known state sequence. *Smyth*, 6:54-55. Then, when the end of an utterance is detected, *Smyth* teaches “a label signal indicating the most probable state sequence is output from the parsing processor 351 to the output port 38, to indicate that the corresponding name, word or phrase has been recognised.” *Smyth*, 6:56-61. Thus, *Smyth* teaches identifying words in a digital audio stream. *Dec.*, 197-200.

Finally, *Smyth* teaches parsing processor 351 is *programmed to* identify the spoken words. *Smyth* teaches “parsing processor 351 is arranged to” perform the word identification. *Smyth*, 6:36. Further, *Smyth* teaches “parsing processor 351...may be a DSP device.” *Smyth*, 6:47-50. Given *Smyth*’s express disclosures that the other two DSPs (i.e., the claimed first and second programmable devices) are “suitably programmed,” a POSITA would have understood *Smyth* to teach the third DSP being similarly programmed. *Dec.*, 149-156. Additionally, it was well

known that DSPs are programmable devices. *Dec.*, 71-72, 107-122, 149-156, 195-200.

c) **“using Hidden Markov Models and/or Neural Networks”**

Smyth teaches identifying spoken words using Hidden Markov Models, as claimed. *Smyth* teaches “[t]he **word representation models may be Hidden Markov Models (HMMs)**, templates, or any other suitable word representation model.” *Smyth*, 3:15-17. *Smyth* also teaches parsing processor 351’s (i.e., the claimed third programmable device) word identification “calculation employs the well known Hidden Markov Model method.” *Smyth*, 6:43-44. *Smyth* teaches “[t]he parsing processor 351 [(i.e., the third programmable device)] is arranged to...calculate the most likely path of states to date over time, and to compare this with each of the state sequences stored in the state sequence memory 352.” *Smyth*, 6:36-42. *Smyth* explains the plurality of state sequence fields correspond “to a **word** or ph[r]ase sequence to be recognised consisting of a string of allophones.” *Smyth*, 6:15-19, 6:26-50. Because the parsing processor uses HMMs for word identification, the parsing processor is “*programmed to identify spoken words in said digital audio stream using Hidden Markov Models and/or Neural Networks,*” as claimed. *Dec.*, 201-202; *Smyth*, 6:45-46 (teaching parsing processor 351 performs the HMM processing).

d) Motivation to Modify *Jiang* with *Smyth* for Claim 1(e)

A POSITA would have found it obvious and been motivated to modify *Jiang* to include *Smyth*'s use of HMMs for word identification. *Dec.*, 203-208. Employing *Smyth*'s HMMs for word identification in the prefix tree of *Jiang* would have achieved a more accurate threshold scoring (*Jiang*, 11:6-15), which in turn would realize a more precise pruning of the lexical tree than *Jiang* currently provides. With a more precise pruning, more efficient and accurate recognition are achieved. *Id.*

As background for why the modification results in more efficient and accurate recognition, it is helpful to understand how *Smyth* advantageously uses HMMs for word identification. *Smyth* teaches that a state sequence corresponds to a word or phrase to be recognized. *Smyth*, 6:53-56, 6:15-19. "The states of the sequence are a plurality of groups of three states each relating to a single allophone." *Smyth*, 6:24-26, 5:58-62, 6:15-19 (a word is a string of allophones); *Dec.*, 160-165, 204-205 (explaining that an allophone is another way to pronounce a phoneme). The states of the sequence are modeled by an HMM. *Smyth*, 6:43-47. Therefore, the state sequence comprises a plurality of states, where each three-state grouping relates to a single allophone (or phoneme), and the state sequence is modeled by HMMs. *Dec.*, 203-208.

Smyth teaches that for each state in the state sequence, two probabilities are output—a repeat probability and a transition probability to the following state.

Smyth, 6:20-24. Transition probabilities provide helpful information in speech recognition, such as a likely duration of the speech corresponding to the state. *See Robinson* (EX1057), 11:12-22 (discussing difference between an acoustic probability (i.e., the occurrence of the phoneme at the particular state) and a transition probability (i.e., the probability the phoneme is transitioning from one phoneme to the next)); *Jiang*, 1:24-27 (defining a “transition probability” as the “probabilit[y] that an observation will transition from one state to the next”). Recall that each state of a state sequence represents an allophone (akin to a phoneme), and the word is a string of allophones. *Smyth*, 6:15-19. Therefore, when modeling a word using HMMs, as taught by *Smyth*, the transition probability informs the likelihood the spoken speech is transitioning to the next state, in particular the next allophone (or phoneme). *Dec.*, 203-208; *Smyth*, Fig. 5 (illustrating transition probabilities in an HMM). Because the HMM includes *transition probabilities* for each recognized phoneme, the HMM is used for word identification with increased accuracy. *Dec.*, 203-208.

It would have been obvious to modify *Jiang* to include HMMs for word identification in its prefix tree. Per Mr. Schmandt, *Jiang* suggests using HMMs for word identification:

This corresponding HMM is thus associated with the observed sequence. **This technique can be extended**, such that if each distinct

sequence of states in the HMM is associated with a sub-word unit, such as a phoneme, then **a most likely sequence of sub-word units can be found.**

Jiang, 1:49-53. Mr. Schmandt opines that *Jiang* is explaining that an HMM can be used to identify the most likely sequence of states for a *phoneme*, and that the “technique can be extended” to find the most likely sequence of sub-word units (i.e., phonemes). The sequence of sub-word units would then be a word. *Dec.*, 203-208.

Jiang also states: “the most likely state sequence is then converted into a corresponding spoken subword unit, **word**, or word sequence,” further indicating *Jiang* uses HMMs for word identification. *Jiang*, 2:14-16. Indeed, *Jiang* even teaches “a phonetic **HMM tree search** can be performed in order to find the proper phoneme,” further suggesting utilizing HMMs in the tree search for at least phoneme identification. *Jiang*, 7:45-46. Mr. Schmandt opines that these collective teachings strongly suggest that *Jiang* contemplates using HMMs for performing word identification, including using HMMs within a lexical tree search. *Jiang* also already recognizes the importance of transition probabilities in HMMs: “a transition probability is defined for each transition from **each state to every state**, including transitions to the same state.” *Jiang*, 1:22–2:15, pinpoint at 1:21-24.

Jiang thus expressly suggests (1) using HMMs for word identification by finding the “most likely sequence of sub-word units”; and (2) *employing* transition

probabilities. *Jiang*, however, does not expressly provide details on using HMMs for word identification or how to incorporate transition probabilities in the prefix tree in its teachings at 8:16-42. *Smyth* provides further details supplementing *Jiang*'s express suggestion of using transition probabilities for word identification and performing HMM tree searches.

By utilizing HMMs for word identification in the *Jiang* prefix tree, transition probabilities can be used in addition to *Jiang*'s distance calculations (i.e., "scores" discussed at 8:16-37) to identify the word more accurately. *Dec.*, 203-208. In *Jiang*'s prefix tree and without utilizing HMMs, each node does not have a transition probability. Instead, the tree has (1) a first score for the observation at the node (e.g., likelihood that the encoded phoneme in the codeword corresponds to the phoneme for the branch under consideration); and (2) a second score based on the first score and first scores for nodes further up the tree. *Jiang*, 8:44-51; *Dec.*, 203-208.

In the modified *Jiang*, for each node in the tree, a score indicating the most likely phoneme is determined (i.e., the claimed distance calculation), as taught by *Jiang*, 8:16-37, and a transition probability is determined, as taught by both *Jiang*, 1:21-24 and *Smyth*, 6:20-24. By utilizing HMMs for word identification, and because HMMs include a transition probability for each state, the transition probability from one node of the prefix tree to the next node is advantageously used to inform the word identification and to prune branches that are

unlikely. *Dec.*, 203-208. The transition probability for each node in the modified *Jiang* prefix tree better informs the recognition analysis, as it allows a more likely path (or sequence) to be chosen and pruning earlier in the recognition process less likely paths. *Dec.*, 203-208. This allows for an even higher *threshold* for pruning the prefix tree, thereby saving computational time and expense and improving overall speech recognition accuracy. *Dec.*, 203-208; *Jiang*, 11:6-15.

There would have been a reasonable expectation of success in modifying *Jiang*. First, the modification is employing a known technique of using HMMs to model words. Given the well-known prevalence at the time of utilizing HMMs for speech recognition, applying *Smyth*'s word HMM techniques to *Jiang* would have been seen as continuing the optimization of speech recognition. From a hardware perspective, additional memory would need to be utilized for the word HMMs, i.e., the stored state sequence memory 352 in *Smyth*. Adding additional memory to *Jiang*'s system, however, would have had minimal scaling and monetary costs, especially when weighed against the increased accuracy and resulting processing efficiency. *Dec.*, 203-208.

G. Claim 1(f)

Claims 1(f)-1(g) are similar, in that both recite the word identification using one or more distances. Claim 1(f), however, recites the distances “that were calculated by the second programmable device,” and Claim 1(g) recites the distances

“calculated from a first feature vector.” The below mapping for Claim 1(f) maps the word identification using one or more distances, which is also incorporated by reference for the Claim 1(g) mapping for the similar limitation, but further maps the distances “calculated by the second programmable device.”

Jiang teaches “**the information** received by tree search engine 74 **based on its accessing of the phonetic speech unit models** in memory 72 **is used in searching lexicon 70 to determine a word which most likely represents the codewords** received by feature extraction module 66...” *Jiang*, 7:49-54. As discussed in Section V.D.3.c, a probability or likelihood score is computed for each node associated with a phoneme branch of the prefix tree used for word identification. *Jiang*, 8:16-24, 8:29-33. Per Mr. Schmandt, the “information...based on [] accessing of the phonetic speech unit models,” quoted at *Jiang*, 7:49-54, includes, at the least, the likelihood score for each node discussed at *Jiang*, 8:16-42 as computed using HMMs discussed at *Jiang*, 7:30-47. *Dec.*, 209-213. That is, *Jiang* teaches calculating a score indicating a likelihood that the phoneme being examined in the prefix tree is the actual phoneme indicated by the codeword (i.e., the feature vector) received from the feature extraction module. *Jiang*, 8:16-24. This calculation of the score utilizes the HMM states of *Jiang*, 7:34-43. *Dec.*, 209-213. Thus, the information received by tree search engine based on accessing the phonetic speech unit models (i.e., the “HMMs which represent phonemes”) (*Jiang*, 7:35-37) is “used

in searching lexicon 70 to **determine a word**” representative of the utterance. *Jiang*, 7:51; *Dec.*, 209-213.

Jiang’s discussion of the prefix tree further supports that the score for the most likely phoneme is used in determining the word. *Jiang* teaches that feature extraction module 66 divides the uttered word into constituent phonemes, e.g., AO-R-IX-N-JH for the word ORANGE, where the AO phoneme is more likely to correspond to the uttered speech than the AE and T phonemes considered in *Jiang*’s prefix tree at nodes 82-86. *Jiang*, 8:25-43. Per Mr. Schmandt, identifying the word ORANGE comprised of the phonemes AO-R-IX-N-JH is using one or more calculated distances. *Dec.*, 209-213. The distances are the scores assigned to each node representing the likelihood of the phoneme, i.e., the score for node 82 indicating the phoneme corresponding to the codeword is highly likely to be AO. *Dec.*, 209-213.

Mr. Schmandt explains that in the modified *Jiang*, where the score calculation of the codeword to the HMM states (discussed at 7:30-43) is performed by a distinct second programmable device (*see* Section V.D.4), the third programmable device obtains the calculated score from the second programmable device to perform the word search traversing the lexical tree. *Dec.*, 209-213. As the tree is traversed, because the phoneme AO was scored as highly likely (from the scoring performed by comparison of the codeword to the HMMs), then the prefix tree uses the

AO-identified phoneme to eventually perform *word identification* for the word ORANGE. *Dec.*, 209-213 (*citing Jiang*, 8:5-28).

Yet further, *Jiang* notes that because there are a finite number of phonemes, there is a relatively “small number of initial branches of the tree,” with the initial branches corresponding to the approximately 40-50 known phonemes. *Jiang*, 3:1-21. The score determined for the initial branch is, or is based on, the score used to determine the initial phoneme of the codeword (i.e., the feature vector). *Dec.*, 209-213 (*citing Jiang*, 8:5-43). Thus, at the least, the initial branches of the prefix tree used for word identification uses the calculated distances, i.e., one or more probability scores for the beginning one or more phonemes of the word, e.g., the AO, AE, and T shown in Fig. 3, to perform *word identification*. *Id.*

Finally, because the probability score was calculated by the modified second programmable device, i.e., the tree search engine 74 modified to include a second programmable device as discussed in Section V.D.5, the prefix tree is performing word identification using one or more distances “calculated by the second programmable device,” as claimed. *Dec.*, 170-182, 209-213.

H. Claim 1(g)

Petitioners incorporate by reference the mapping for Claim 1(f), which maps identification of words using distances (i.e., the probability scores). Additionally, the word identification uses distances “calculated from a first feature vector,” as

claimed. As discussed in Section V.B.3, the digital samples of the user's spoken utterance are sampled into frames. *Jiang*, 1:58-67. The frames are encoded into feature vectors, and the feature vectors are, in some embodiments, further encoded into codewords using vector quantization. *Jiang*, 6:62–7:10. Per Mr. Schmandt, in the instance where the identification of spoken words is for an utterance corresponding to a first feature vector (i.e., the utterance is converted to a digital sample, which is then encoded as a first feature vector), the word uses one or more distances (i.e., the probability score for a phoneme) from the first feature vector. *Dec.*, 214. The word identification thus uses one or more distances calculated from a first feature vector because the word identification uses a phoneme determined from a first feature vector. *Id.*

I. Claim 1(h)

Petitioners incorporate by reference Sections V.G-H, which map using the calculated distances (i.e., the probability scores) to identify words within a lexical tree (i.e., *Jiang*'s prefix tree). To summarize the mapping, *Jiang* teaches calculating distances, namely a score indicating the most likely phoneme based on the codeword (feature vector) from the HMM. *Jiang*, 7:30-43, 8:15-42, 7:49-54 (“[t]he information received by tree search engine 74 based on its accessing of the phonetic speech unit models in memory 72 [the probability or likelihood score] is used in searching lexicon 70 to determine a word which most likely represents the codewords received

by feature extraction module 66”). This score is then used during a search stage to determine a most likely phoneme represented by the codeword based on the HMMs. The sequence of most likely phonemes identifies a word. *Jiang*, 8:25-51; *Dec.*, 215-218.

Per Mr. Schmandt, a POSITA would have understood *Jiang*’s determination of a word by traversing prefix tree is a “search stage,” because *Jiang* expressly states such in labeling the engine 74 as a “tree **search**” engine. *Dec.*, 215-218. That is, by traversing prefix tree, a search stage is performed for word identification. *Jiang*, 7:67–8:4; *Dec.*, 215-218.

Mr. Schmandt also opines a POSITA would have understood prefix tree is a “lexical tree,” as claimed. *Jiang* expressly teaches “[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.” *Jiang*, 2:47-52. Because the words of lexicon 70 are “presented to tree search engine 74 in the form of a prefix tree” comprising branches and nodes, as is conventional, *Jiang*’s prefix tree is a “lexical tree.” *Dec.*, 215-218 (*citing Ravishankar*, EX1012).

Finally, *Jiang*’s prefix tree (i.e., the claimed “lexical tree”) comprises a model of words, as claimed. *Jiang* expressly teaches all the words likely to be encountered are represented in the tree structure. *Jiang*, 2:49-52. Additionally, the words of the lexicon 70 (which contains information representative of all words of the system’s

vocabulary) are presented in the form of the prefix tree. *Jiang*, 7:65–8:4. Because *Jiang*’s prefix tree comprises a lexicon 70 of all words in the system’s vocabulary, *Jiang* teaches a “lexical tree comprising a model of words,” as claimed. *Dec.*, 215-218.

VI. GROUND 2: CLAIMS 2-3 ARE OBVIOUS OVER *JIANG*, *SMYTH*, AND *NGUYEN*

A. Claim 2

Jiang-Smyth teaches or renders obvious Claim 1. *Jiang-Smyth-Nguyen* teaches Claim 2.

Nguyen teaches a speech recognition system where at step 12, “***a similarity measure is computed*** between each frame of input speech and each reference pattern. The process defines ***a local measure of closeness*** between acoustic feature vectors[.]” *Nguyen*, 2:30-38. After a similarity measure is computed, a decoding step 16 occurs, wherein the “probability that a particular phoneme was spoken is provided by the acoustic models as part of the pattern matching process. A sequence of words can then be constructed by concatenating the phonemes observed during the pattern matching process.” *Nguyen*, 3:11-24. Figure 1 depicts *Nguyen*’s pattern classification (i.e., distance calculation) step 12 and decoding step 16.

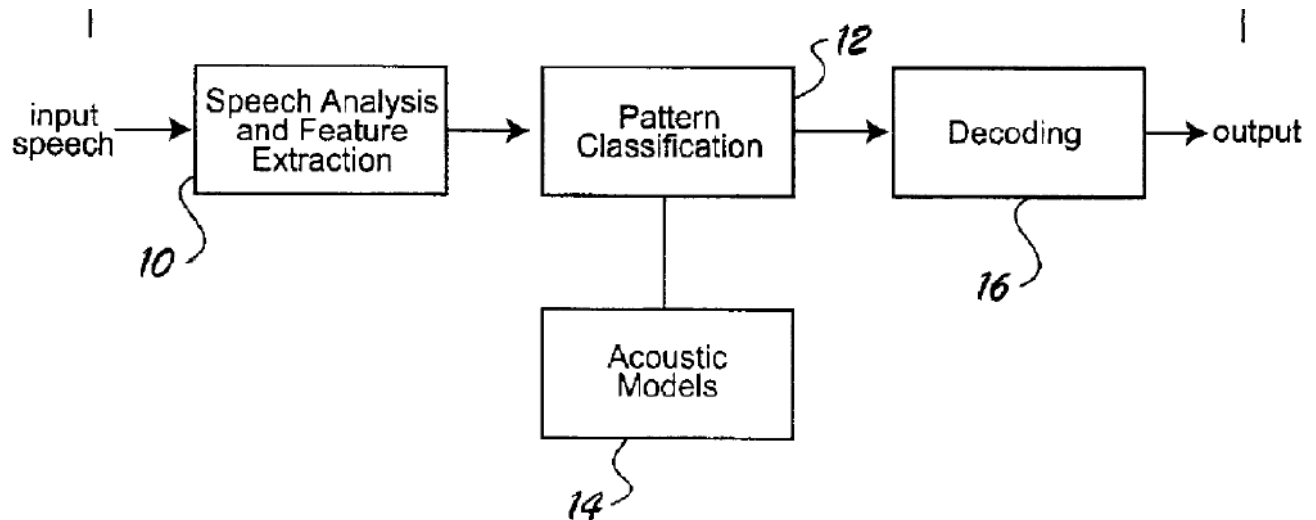


FIG. 1

Nguyen, Fig. 1 (excerpt).

Nguyen further teaches an embodiment where a similarity measure is computed at step 36 for *each* acoustic feature vector in a *first group* of vectors. *Nguyen*, 3:51-52, 3:61-4:7. After the similarity measure is computed, a search process is performed at step 44. *Nguyen*, 4:8-16. *Nguyen* teaches that “[c]ontemporaneous with the search process, **a subsequent group of acoustic feature vectors** may be retrieved into the memory workspace at step 48. **A similarity measure is computed for each acoustic feature vector in this subsequent group** as described above.” *Nguyen*, 4:17-22. Therefore, a POSITA would have understood that *Nguyen* teaches performing a similarity measure (i.e., distance calculation) while contemporaneously performing the search step. *Dec.*, 219-227.

A POSITA would have found it obvious and been motivated to modify the *Jiang-Smyth* speech recognition system to operate using the contemporaneous processing taught in *Nguyen*. *Dec.*, 219-227. *Nguyen* teaches performing the similarity computation (i.e., the distance calculation) for a group of acoustic feature vectors reduces the number of I/O operations required to load and unload each acoustic model. *Nguyen*, 3:57-60; *Dec.*, 219-227. A POSITA also would have been motivated to modify the *Jiang-Smyth* system's multiple DSPs to use *Nguyen's* contemporaneous distance calculation step and searching step to advantageously increase system performance. *Dec.*, 219-227. *Smyth* teaches one DSP for performing distance calculations (i.e., classifier processor 341 implemented as a DSP) and another DSP for performing speech recognition (i.e., parsing processor 351 implemented as a DSP). Modifying *Jiang-Smyth* according to *Nguyen's* teachings would have advantageously resulted in the *Jiang-Smyth* system's separate DSPs processing contemporaneously, thus improving system performance. *Dec.*, 219-227. Indeed, a POSITA would have understood "contemporaneous" processing was well known in the prior art. *Ravishankar* provides one example wherein the CMU Sphinx-II speech recognizer realized "a speed up of 3.9" by "exploit[ing] the large degree of concurrency within its algorithmic steps[.]" *Ravishankar*, 105; *Dec.*, 225.

There would have been a reasonable expectation of success in the proposed modification. For example, a POSITA would have understood and found obvious to

modify *Jiang-Smyth* such that feature vectors would have been calculated and batched then sent to classifier processor 341 for performing distance calculations. *Dec.* 219-227. Batch distance calculation results would have been transmitted to parsing processor 351 for speech recognition processing, resulting in parsing processor 351 performing speech recognition processing while classifier processor 341 performs distance calculations on a new batch of calculated feature vectors. *Dec.* 219-227.

Because *Jiang-Smyth* renders obvious utilizing separate DSPs, such a modification would have resulted in improved efficiency in *Jiang-Smyth's* speech recognition system by causing separate processing components to work in parallel (i.e., contemporaneously). *Dec.* 219-227. A POSITA would have understood that *Smyth* already teaches utilizing multiple DSPs, such that modifying *Jiang-Smyth* to utilize *Nguyen's* contemporaneous processing functionality would have had a reasonable expectation of success and would have required only minor modifications. *Dec.* 219-227. Further, because such contemporaneous processing was already well known in the prior art, as taught by *Nguyen*, such a combination would have had a reasonable expectation of success. *Dec.* 219-227.

B. Claim 3

The *Jiang-Smyth-Nguyen* combination teaches or renders obvious Claims 1-2. *See* Sections V-VI. *Smyth* teaches Claim 3.

The claimed “processing circuit” is described in the ’377 Patent only with respect to the speech processing circuit. ’377 *Patent*, 24:5-10. The ’377 Patent does not ascribe any distinction, either structural or functional, to a “processing circuit” as compared to a “programmable device,” and therefore, a POSITA would have understood a processing circuit is met or rendered obvious by a programmable device, such as a DSP. *Dec.*, 228-230; ’377 *Patent*, 5:52-55, 24:27-33.

Smyth teaches a speech recognition system with three physically distinct processing circuits, namely three separate DSPs for performing speech recognition. *Smyth*, 5:49-53, 6:5-8, 6:43-51; *Dec.*, 150-156, 228-230; Section V.D.4.a. Claim 1, as mapped, identifies three separate processors (either CPU or DSP), namely: (1) *Jiang’s* feature extraction module implemented as an array processor (Section V.B.3) or modified to include *Smyth’s* DSP (Section V.B.4); and (2) *Jiang’s* tree search engine modified to be two separate, distinct processors as taught by *Smyth* (Section V.D.5). Claim 1 thus maps that the first, second, and third programmable devices are “physically distinct processing circuits,” as recited in Claim 3. *Dec.*, 230.

VII. GROUND 3: CLAIM 4 IS OBVIOUS OVER *JIANG, SMYTH, NGUYEN, AND BOIKE*

The *Jiang-Smyth-Nguyen* combination teaches or renders obvious Claims 1-3. *Jiang-Smyth-Nguyen-Boike* teaches Claim 4.

Boike teaches a “system-on-a-chip (SoC) integrated circuit 10 containing multiple digital signal processors (DSPs).” *Boike* (EX1006), 2:52-54. *Boike* depicts such DSP subsystems 18-21 as part of DSP/SoC 10 (i.e., integrated circuit 10). “FIG. 1 is a block diagram depicting a typical configuration of a DSP/SoC multiple digital signal processor integrated circuit.” *Boike*, 2:42-44. Annotated Fig. 1 below illustrates a single integrated circuit 10 (annotated blue) and DSP subsystems 18-21 (annotated green).

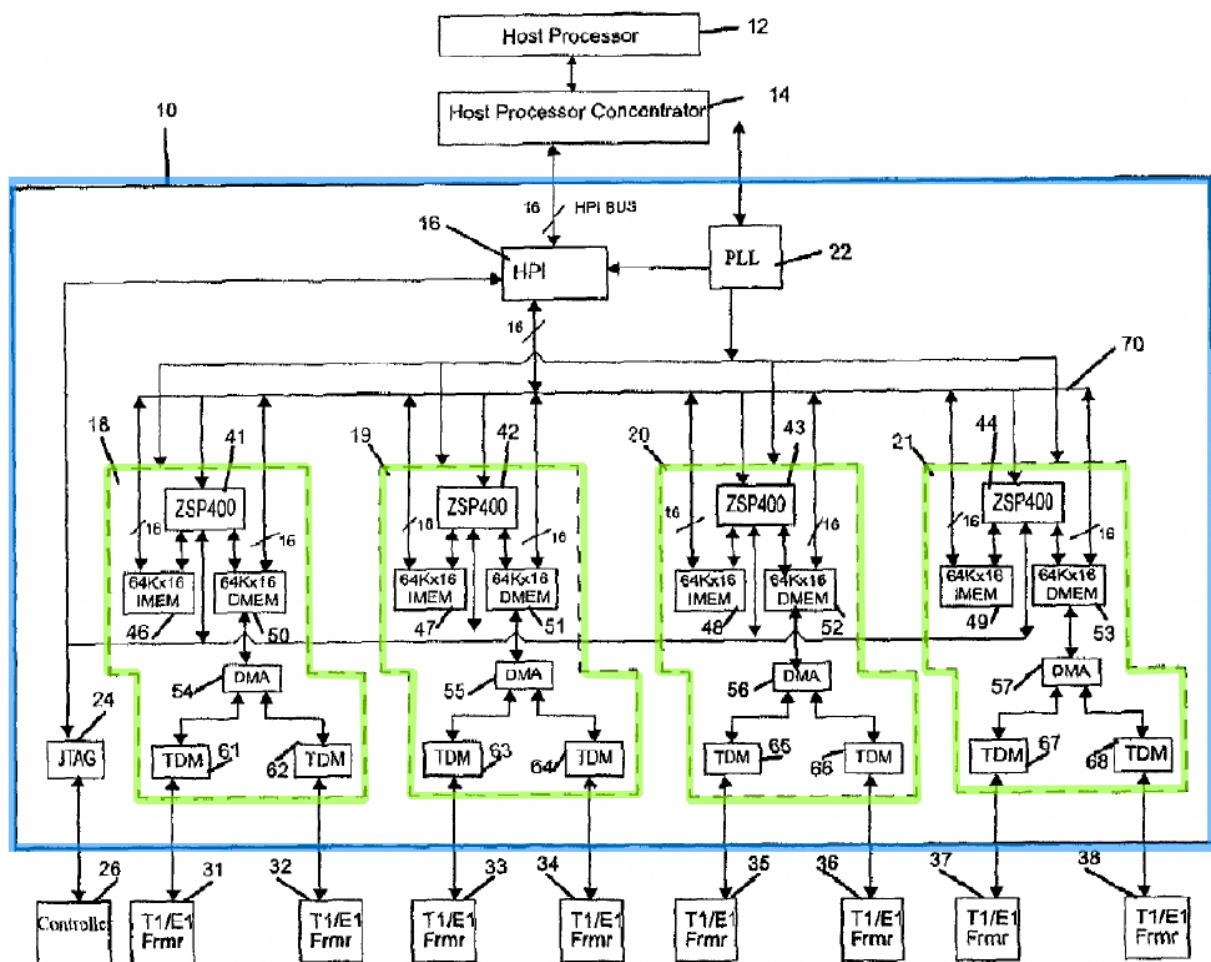


FIGURE 1

Boike, Fig. 1 (annotated). *Boike* teaches numerous advantages of an architecture placing multiple DSPs on a single IC, including reduced space and costs and conserved power. *Boike*, 1:19-42, 2:3-9, 2:21-35; *Dec.*, 231-234.

A POSITA would have found it obvious and been motivated to modify *Jiang-Smyth-Nguyen* to include the three distinct processing circuits on a single IC, as taught by *Boike*. *Dec.*, 231-234. *Boike* alone expressly motivates the combination, informing a POSITA that there is reduced space and costs and conserved power in forming multiple DSPs on a single IC. *Dec.*, 231-234 (*citing Boike*, 1:19-42, *Ozcelik*, 1:47-51).

VIII. GROUND 4: CLAIM 5 IS OBVIOUS OVER *JIANG, SMYTH, NGUYEN*, AND *BAUMGARTNER*

The *Jiang-Smyth-Nguyen* combination teaches or renders obvious Claims 1-2. *Jiang-Smyth-Nguyen-Baumgartner* teaches or renders obvious Claim 5.

Jiang teaches a speech recognition system 60 comprising computer 20 including CPU 21. *Jiang*, Figs. 1-2, 5:4-9. CPU 21 may include multiple processors: “CPU 21 (which may include one or more processors).” *Id.* Because *Jiang*’s CPU may include multiple processors, *Jiang*’s CPU is a “multi-processor,” as claimed. *Dec.*, 235-245.

The Claim 1(a) mapping mapped the first programmable device as *Jiang*’s array processor (Section V.B.3). The Claim 1(e) mapping mapped the third programmable device as *Jiang*’s tree search engine (which may be a dedicated

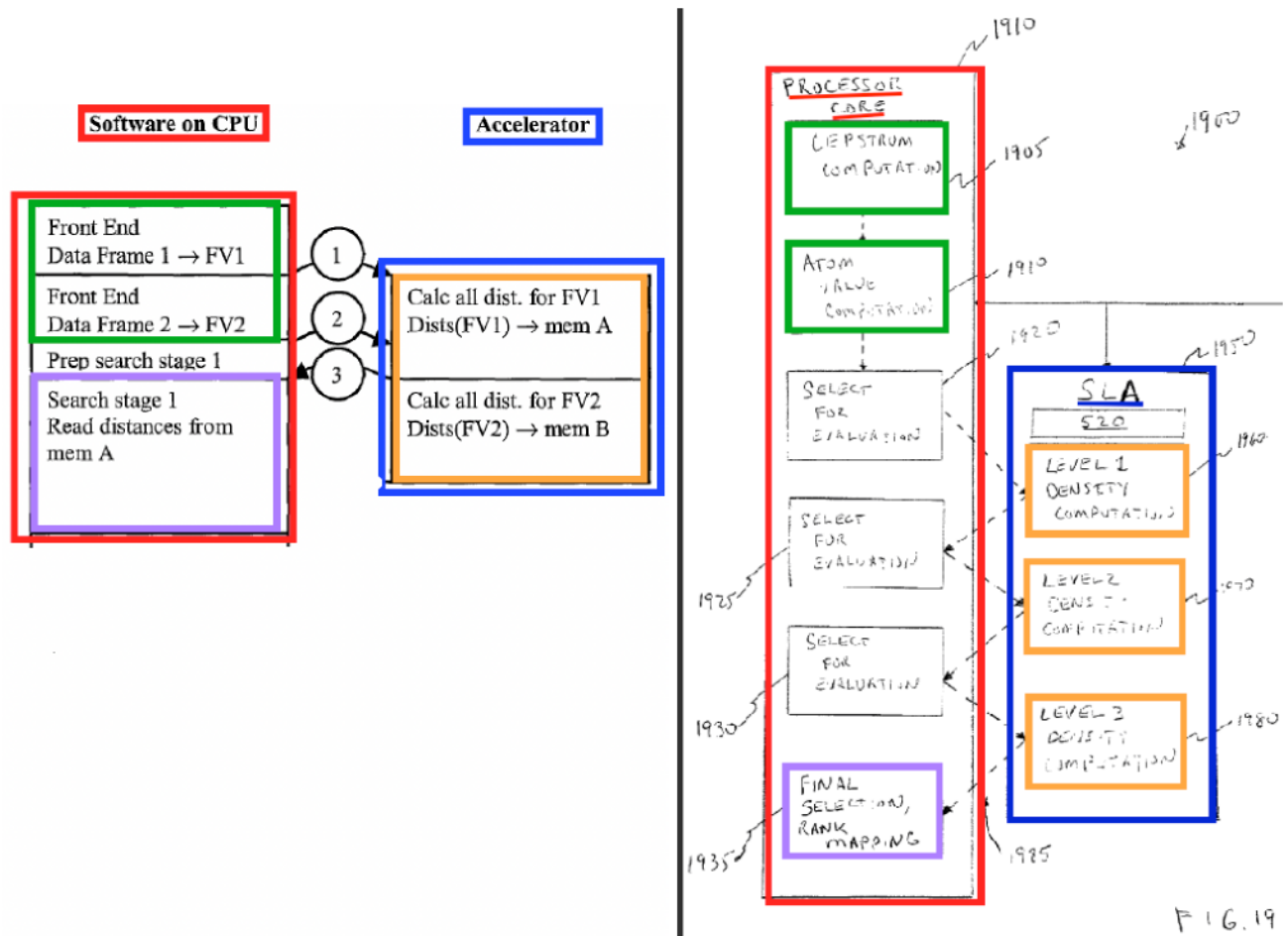
processor) modified to perform only the word identification (Section V.F.1.a) or *Smyth*'s DSP 351 (Section V.F.2.a).

As discussed below, *Baumgartner* teaches the functionalities of feature vector calculation and word identification being performed by one physical processor, namely processor core 1910, and the functionality of distance calculation being performed by a speech label accelerator including a circuit for distance calculation.

In the proposed modification of *Jiang* for Claim 5, it would have been obvious to perform the functionalities of the *Jiang* array processor or *Smyth* DSP (feature vector calculation) and *Jiang* tree search engine (word identification) using a multiprocessor (as otherwise taught by *Jiang*), such as *Baumgartner*'s processor core 1910. It further would have been obvious to perform computationally-intensive distance calculation via a separate processor, such as *Baumgartner*'s SLA including a circuit.

Baumgartner teaches it is desirable for distance calculation to be performed on a physically distinct device separate from feature vector calculation and word identification. *Baumgartner* teaches feature vector calculation and word identification performed on a first processor core 1910. *Baumgartner*, [0031]-[0033], [0156]-[0158], Fig. 19. In *Baumgartner*, distance calculation is performed on a separate, physically distinct processor. *Id.*

Baumgartner's Fig. 19 (annotated below) illustrates a **cepstrum computation** (i.e., feature vector calculation) 1905 and atom value computation 1910 and a **final selection rank mapping** (i.e., word identification) 1935. Each of the **cepstrum computation** 1905, and **atom value computation** 1910, and **rank mapping** 1935 is implemented on **processor core** 1910. **Density computations** (i.e., distance calculations) 1960-1980 are implemented on **Speech Label Accelerator (SLA)** 520.



'377 Patent, Fig. 21 (annotated, excerpted, left); *Baumgartner*, Fig. 19 (annotated, excerpted, right). *Dec.*, 240-243.

Regarding the claimed “physically distinct processing *circuit*,” Petitioners incorporate Section VI.B, Claim 3. *Baumgartner* also teaches the distance calculation is performed by a circuit, which is the “heart” of the SLA. *Baumgartner*, [0043]. Therefore, *Baumgartner* teaches a physically distinct processing circuit, as claimed. Additionally, when *Jiang* is modified to include the *Smyth* DSP for performing distance calculation (Section V.D.4.a), the *Smyth* DSP is a circuit. *See* Section VI.B.

Because *Jiang* already teaches CPU 21 includes multiple processors, it would have been obvious to modify *Jiang*, to the extent any modification is required, to configure the multiple processors of CPU as including (1) *Jiang*’s processor or *Smyth*’s DSP for performing feature vector calculation; and (2) *Jiang*’s tree search engine in the form of a dedicated processor and modified to only perform word identification or *Smyth*’s DSP 351. *Dec.*, 235-245. Such a modification is motivated by *Baumgartner*’s processor core 1910 performing similar functionality and further because *Jiang* already recognizes multiprocessor computer configurations can be employed and that the CPU comprises multiple processors. *Id.*

A POSITA would have also found it obvious and been motivated to modify *Jiang-Smyth-Nguyen* with *Baumgartner* as proposed to place the computationally-intensive processing load of distance calculation on a physically distinct circuit or programmable device. *Baumgartner* explains “[s]peech recognition...is currently a

computationally intensive task,” that “it would be beneficial to reduce the load on the processors of the system when the system performs speech recognition,” and “a need exists for techniques to reduce the computational load of a processor during speech recognition.” *Baumgartner*, [0003]-[0006]. A POSITA would have recognized *Baumgartner*’s “invention provides Speech Label Accelerators (SLAs),...speeds processing during the labeling stage of a speech recognizer[,] and frees one or more processors from performing labeling.” *Dec.*, 245 (*quoting Baumgartner*, [0007]), 245 (explaining *Baumgartner*’s “labeling stage” performed by SLA is similar to distance calculation). A POSITA would have further understood that *Baumgartner*’s hardware architecture—utilizing an accelerator for expensive calculations and implementing the functionality of feature vector calculation and word identification collectively by a single processor—provides an efficient solution to the computational load issues identified by the prior art. *Dec.*, 132, 175, 225, 245 (*citing Ravishankar*, 33, 105); *Baumgartner*, [0200]–[0205].

Because (1) *Baumgartner* suggests performing processing of the functionality of feature vector calculation and word identification collectively on processor core 1910, (2) *Smyth* teaches performing the three stages of speech recognition on separate DSPs; and (3) *Jiang* teaches CPU includes multiple processors, including array processor for feature vector calculation and tree search engine for word

identification, the collective teachings of *Jiang-Smyth-Nguyen-Baumgartner* render obvious Claim 5.

IX. GROUND 5: CLAIM 6 IS OBVIOUS OVER *JIANG, SMYTH, NGUYEN, BAUMGARTNER, AND BOIKE*

The *Jiang-Smyth-Nguyen-Baumgartner* combination teaches Claims 1-2 and 5. *Jiang-Smyth-Nguyen-Baumgartner-Boike* teaches Claim 6.

As discussed for Claim 4, *Boike* teaches multiple processors, such as DSPs, formed on a single IC, as claimed. *See* Section VI.A; *Boike*, 2:42-54, FIG. 1. A POSITA would have found it obvious and been motivated to modify *Jiang-Smyth-Nguyen-Baumgartner* to include a single IC comprising the first DSP (i.e., the claimed “first programmable device”) and the third DSP (i.e., the claimed “third programmable device”), as taught by *Boike*. *Boike* expressly motivates the combination, informing a POSITA that there is reduced space and costs and conserved power in forming multiple DSPs on a single IC. *Dec.*, 246-248 (*citing Boike*, 1:19-42, *Ozcelik*, 1:47-51).

Additionally, *Baumgartner* expressly teaches a CPU core and SLA 500 “reside on the **same physical die.**” *Baumgartner*, [0180], [0174] (“SLA 500 is built on the same silicon substrate as CPU core 2005.”); *Dec.*, 246-248 (discussing a physical die and silicon substrate both refer to an integrated circuit and a POSITA would have been motivated to combine teachings from the embodiments of Figs. 19 and 20).

X. THE BOARD’S DISCRETION UNDER 35 U.S.C. §314(a) AND §325(d)

§314(a). None of the factors articulated in *General Plastic Industrial Co., Ltd. v. Canon Kabushiki Kaisha*, IPR2016-01357, Paper 19 at 16 (PTAB Sept. 6, 2017) favor discretionary denial. Regarding Factors 1–5, Petitioners have filed no prior petition regarding the ’377 Patent. Regarding Factors 6–7, joinder will neither consume additional Board resources nor affect the timing of a final written decision, because Petitioners agree to assume an understudy role behind Apple (with the same case schedule and arguments) and not actively participate in proceedings, unless Apple’s participation terminates.

The considerations in *Apple Inc. v. Fintiv, Inc.*, IPR2020-00019 (PTAB, Mar. 20, 2021) also disfavor discretionary denial. Trial in the co-pending litigation, scheduled for April 22, 2024 (EX1072), will begin approximately at the same time a Final Written Decision is due, approximately May 1, 2024. (Factor 2). The court only recently issued its Claim Construction Order on June 16, 2023, and dispositive motions are not due until January 26, 2024. Therefore, the parties and the court are at the early stages of the invalidity analysis. (Factor 3). The Petition also presents a strong showing of unpatentability; indeed, the Board has already found Apple’s Petition, that presented grounds of unpatentability identical to those presented here, has shown that there is a reasonable likelihood it would prevail with respect to at

least one of the claims challenged in the Petition. (IPR2023-00035, Paper 10; Factor 6).

§325(d). Petitioners seek to join already instituted IPR2023-00035, based on the same grounds as applied to the same claims. Petitioners agree to assume an understudy role and will only actively participate in proceedings if Apple's participation terminates.

XI. CONCLUSION

For the forgoing reasons, Petitioners respectfully request *inter partes* review of the Challenged Claims of the '377 Patent.

Respectfully submitted,

Dated: July 12, 2023

/J. David Hadden/
J. David Hadden
Reg. No. 40,629

Attorneys for Petitioners
Amazon Web Services, Inc. and
Amazon.com Services LLC

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

A. Real Party-In-Interest

Amazon.com Services LLC and Amazon Web Services, Inc. are the real parties-in-interest. 37 C.F.R. § 42.8(b)(1).

B. Related Matters

The '377 Patent is currently the subject of *Apple Inc. v. Zentian Ltd.*, Inter Partes Review No. IPR2023-00035 (P.T.A.B.) (instituted Jun. 12, 2023). This Petition is substantively identical to the petition submitted in IPR2023-00035 and Petitioners have concurrently filed a motion to join that proceeding.

The '377 Patent is also asserted in the following pending patent infringement lawsuits filed by Zentian Ltd:

- Zentian Ltd v. Apple Inc., No. 6:22-cv-00122 (W.D. Tex. Feb. 2, 2022);
- Zentian Ltd v. Amazon.com, Inc., No. 6:22-cv-00123 (W.D. Tex. Feb. 2, 2022).

C. Lead and Back-Up Counsel

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

LEAD COUNSEL	BACK-UP COUNSEL
J. David Hadden (Reg. No. 40,629) dhadden@fenwick.com Fenwick & West LLP 801 California Street	Saina Shamilov (Reg. No. 48,266) sshamilov@fenwick.com Fenwick & West LLP 801 California Street

Petition for *Inter Partes* Review of
U.S. Patent No. 10,062,377

Mountain View, CA 94041 Telephone: (650) 988-8500 Facsimile: (650) 938-5200	Mountain View, CA 94041 Telephone: (650) 988-8500 Facsimile: (650) 938-5200 Dargaye Churnet (Reg. No. 71,288) dchurnet@fenwick.com Fenwick & West LLP 555 California Street, 12th Floor San Francisco, CA 94104 Telephone: (415) 875-2300 Facsimile: (415) 281-1350
---	--

D. Service Information

Service of any documents may be made to the postal mailing address above.

Petitioners consent to electronic service at Amazon-Zentian-IPR@fenwick.com. A Power of Attorney is being filed concurrently with the designation of counsel in accordance with 37 C.F.R. § 42.10(b).

CLAIM LISTING APPENDIX

U.S. PATENT NO. 710,062,377 FOR CLAIMS 1-6

Claim Designation	Claim Language
Claim 1(Pre)	A speech recognition circuit, comprising:
Claim 1(a)	a first programmable device programmed to calculate a feature vector from a digital audio stream,
Claim 1(b)	wherein the feature vector comprises a plurality of extracted and/or derived quantities from said digital audio stream during a defined audio time frame;
Claim 1(c)	a second programmable device programmed to calculate distances indicating the similarity between a feature vector and a plurality of acoustic states of an acoustic model
Claim 1(d)	wherein said feature vector is received by the second programmable device after it is calculated by the first programmable device; and
Claim 1(e)	a third programmable device programmed to identify spoken words in said digital audio stream using Hidden Markov Models and/or Neural Networks
Claim 1(f)	wherein said word identification uses one or more distances that were calculated by the second programmable device,
Claim 1(g)	wherein said identification of spoken words uses one or more distances calculated from a first feature vector; and
Claim 1(h)	a search stage for using the calculated distances to identify words within a lexical tree, the lexical tree comprising a model of words.
Claim 2	The speech recognition system of claim 1 wherein the time period taken for said identification of spoken words overlaps the time period taken for said second programmable device to calculate one or more distances from a second feature vector.

Claim Designation	Claim Language
Claim 3	The speech recognition system of claim 2 wherein the first programmable device, the second programmable device, and third programmable device are physically distinct processing circuits.
Claim 4	The speech recognition system of claim 3 wherein said physically distinct processing circuits are all formed on a single integrated circuit.
Claim 5	The speech recognition system of claim 2 wherein the first programmable device and the third programmable device comprise one physical uni-processor or multi-processor, and the second programmable device is a physically distinct processing circuit.
Claim 6	The speech recognition system of claim 5 wherein said one physical uni-processor or multi-processor and said second programmable device are all formed on a single integrated circuit.

CERTIFICATION OF WORD COUNT

The undersigned certifies pursuant to 37 C.F.R. §42.24 that the foregoing Petition for *Inter Partes* Review, excluding any table of contents, mandatory notices under 37 C.F.R. §42.8, certificates of service or word count, or appendix of exhibits, contains 13,985 words according to the word-processing program used to prepare this document (Microsoft Word).

Dated: July 12, 2023

/J. David Hadden/

J. David Hadden

Reg. No. 40,629

Attorneys for Petitioners

Amazon Web Services, Inc. and

Amazon.com Services LLC

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

I hereby certify, pursuant to 37 C.F.R. §§ 42.6(e) and 42.105, that a complete copy of this Petition for *Inter Partes* Review including all exhibits are being served via Federal Express on July 12, 2023, upon Patent Owner by serving the correspondence address of record with the USPTO as follows:

Blank Rome LLP
1825 Eye Street, NW
Washington DC 20006-5403

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

John M. Hughes
Katherine L.I Hacker
Bartlit Beck LLP
1801 Wewatta Street, Suite 1200
Denver, CO 80202

Katherine E. Rhoades
Nevin M. Gewertz
Rebecca T. Horwitz
Bartlit Beck LLP
54 West Hubbard Street, Suite 300
Chicago, IL 60654

John Francis Luman , III
Daniels & Tredennick PLLC
6363 Woodway, Suite 700
Houston, TX 77057

/J. David Hadden/
J. David Hadden
Reg. No. 40,629

Attorneys for Petitioners
Amazon Web Services, Inc. and
Amazon.com Services LLC