

UNITED STATES COURT OF APPEALS FOR THE FEDERAL CIRCUIT

PETITION FOR REVIEW/NOTICE OF APPEAL

Notice is hereby given that the petitioner(s)/appellant(s) listed below hereby appeal(s) the below-noted case to the United States Court of Appeals for the Federal Circuit.

Originating Tribunal (*Name of Agency, Board, Commission, Office, Bureau, or Court whose decision is being appealed*): USPTO Patent Trial and Appeal Board

Case number being appealed: IPR2023-00033

Case title being appealed: Apple Inc. et al v. Zentian Limited

Date of final decision or order being appealed: 06/11/2024

Date decision or order was received: 06/11/2024

☒ I have attached a copy of the decision or order being appealed.

List all Petitioners/Appellants (List each party filing this appeal. Do not use “et al.” or other abbreviations. Attach continuation pages if necessary.)

Zentian Limited

Date: 08/09/2024

Signature: /s/ Kayvan B. Noroozi

Name: Kayvan B. Noroozi

Address: Noroozi PC

11601 Wilshire Blvd, Suite 2170

Los Angeles, CA 90025

Phone Number: (310) 975-7074

Email Address: kayvan@noroozipc.com

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC., AMAZON WEB SERVICES, INC., and
AMAZON.COM SERVICES LLC,
Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00033
IPR2023-01192
U.S. Patent No. 7,587,319

PATENT OWNER'S NOTICE OF APPEAL

Pursuant to 35 U.S.C. §§ 141(c), 142, and 319 and 37 C.F.R. §§ 90.2(a) and 90.3, Patent Owner hereby provides notice that it appeals to the United States Court of Appeals for the Federal Circuit from the Final Written Decision in Case No. IPR2023-00033¹ (Paper 34) entered on June 11, 2024, by the Patent Trial and Appeal Board (“the Board”), and from all orders, decisions, rulings, and opinions underlying the Final Written Decision. A copy of the Board’s Final Written Decision is attached hereto as Exhibit A.

For the limited purpose of providing the Director with the information requested in 37 C.F.R. § 90.2(a)(3)(ii), Patent Owner anticipates that the issues on appeal may include, but are not limited to:

- the Board’s holding that claims 46, 50, 54, 64, and 67 of the ’319 patent are unpatentable as obvious in view of U.S. Patent No. 6,526,380 B1 (“Thelen”), U.S. Patent No. 5,459,798 (“Bailey”), and U.S. Patent No. 5,428,803 (“Chen”);
- the Board’s holding that claims 47, 48, and 49 of the ’319 patent are unpatentable as obvious in view of Thelen, Bailey, Chen, and U.S. Patent No. 6,076,056 (“Huang”);
- the Board’s holding that claims 53, 58, and 59 of the ’319 patent are

¹ IPR2023-01192 was joined with this proceeding.

unpatentable as obvious in view of Thelen, Bailey, Chen, and U.S. Patent No. 6,374,219 B1 (“Jiang”);

- the Board’s holding that claim 51 of the ’319 patent is unpatentable as obvious in view of Thelen, Bailey, Chen, and U.S. Patent No. 4,831,550 (“Katz”);
- the Board’s conclusion that Petitioner’s evidence was sufficient to meet its burden of proof that the disclosures in Thelen, Bailey, and Chen render claims 46, 50, 54, 64, and 67 of the ’319 patent obvious;
- the Board’s conclusion that Petitioner’s evidence was sufficient to meet its burden of proof that the disclosures in Thelen, Bailey, Chen, and Huang render claims 47, 48, and 49 of the ’319 patent obvious;
- the Board’s conclusion that Petitioner’s evidence was sufficient to meet its burden of proof that the disclosures in Thelen, Bailey, Chen, and Jiang render claims 53, 58, and 59 of the ’319 patent obvious;
- the Board’s conclusion that Petitioner’s evidence was sufficient to meet its burden of proof that the disclosures in Thelen, Bailey, Chen, and Katz render claim 51 of the ’319 patent obvious;
- any finding or determination supporting or related to these issues;
- the Board’s findings that conflict with the evidence of record or are otherwise not supported by substantial evidence;

- the Board's legal errors in undertaking the aforementioned invalidity analysis; and
- all other issues decided adversely to Patent Owner in any orders, decisions, rulings, and opinions.

Pursuant to 37 C.F.R. § 90.3, this Notice of Appeal is timely, having been duly filed within sixty-three (63) days after the date of the Final Written Decision.

A copy of this Notice of Appeal is being filed simultaneously with the Patent Trial and Appeal Board, the Clerk's Office for the United States Court of Appeals for the Federal Circuit, and the Director of the United States Patent and Trademark Office.

Dated: August 9, 2024

Respectfully submitted,

By: /s/ Kayvan B. Noroozi
Kayvan B. Noroozi, *Pro Hac Vice*
NOROOZI PC
11601 Wilshire Blvd, Suite 2170
Los Angeles, CA 90025

Attorneys for the Patent Owner

CERTIFICATE OF FILING

I hereby certify that, in addition to being filed electronically through the Board's PTABE2E System, a true and correct copy of the foregoing Notice of Appeal was served and filed by FedEx Priority Overnight (Label No. 7778 8649 3473) on this 9th day of August 2024, with the Director of the United States Patent and Trademark Office, at the following address:

Director of the United States Patent and Trademark Office
c/o Office of the General Counsel
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

I further certify that a true and correct copy of the foregoing Notice of Appeal was filed with the Clerk's Office of the United States Court of Appeals for the Federal Circuit through the federal courts' Case Management and Electronic Case Files (CM/ECF) system on the 9th day of August, 2024, along with the requisite fee.

By: /s/ Kayvan B. Noroozi
Kayvan B. Noroozi, *Pro Hac Vice*
NOROOZI PC
11601 Wilshire Blvd, Suite 2170
Los Angeles, CA 90025

Counsel for the Patent Owner

CERTIFICATE OF SERVICE

The undersigned certifies, that on this 9th day of August 2024, the foregoing document is being served by electronic mail (e-mail) on counsel of record for Petitioner as follows:

Adam P. Seitz
Adam.Seitz@eriseip.com
PTAB@eriseip.com

Jennifer C. Bailey
Jennifer.Bailey@eriseip.com

Adam M. Sandwell
Adam.Sandwell@eriseip.com

Erise IP, P.A.
7015 College Blvd., Ste. 700
Overland Park, Kansas 66211

Christina Canino
Christina.Canino@eriseip.com

Erise IP, P.A.
5299 DTC Blvd., Suite 1340
Greenwood Village, CO 80111

By: /s/ Kayvan B. Noroozi
Kayvan B. Noroozi, *Pro Hac Vice*
Counsel for the Patent Owner

Exhibit A

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC., AMAZON WEB SERVICES, INC., and
AMAZON.COM SERVICES LLC,
Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00033¹
Patent 7,587,319 B2

Before KEVIN F. TURNER, JEFFREY S. SMITH, and
CHRISTOPHER L. OGDEN, *Administrative Patent Judges*.

Opinion for the Board filed by *Administrative Patent Judge* OGDEN

Opinion dissenting filed by *Administrative Patent Judge* SMITH

OGDEN, *Administrative Patent Judge*.

JUDGMENT
Final Written Decision
Determining All Challenged Claims Unpatentable
35 U.S.C. § 318(a)

¹ IPR2023-01192 has been joined with this proceeding.

I. INTRODUCTION

In response to a Petition (Paper 1, “Pet.”) filed by Petitioner Apple Inc. (“Apple”), the Board instituted an *inter partes* review of claims 46–51, 53, 54, 58, 59, 64, and 67 of U.S. Patent No. 7,587,319 B2 (Ex. 1001, “the ’319 patent”). Paper 10 (“Dec.”). Patent Owner Zentian Limited (“Zentian”) filed a Patent Owner Response (Paper 19; corrected version filed as Paper 31 (“PO Resp.”)), Apple filed a Reply to the Patent Owner Response (Paper 22 (“Pet. Reply”)), and Zentian filed a Sur-reply (Paper 27 (“PO Sur-reply”)).

We held an oral hearing on March 11, 2024, and the transcript is entered into the record (“Tr.”). Paper 33.

This is a final written decision under 35 U.S.C. § 318(a) as to whether the claims challenged in the *inter partes* review are unpatentable. For the reasons below, we conclude that Apple has shown that all the challenged claims are unpatentable.

II. BACKGROUND

A. RELATED PROCEEDINGS

As related matters, the parties identify *Zentian Ltd. v. Apple Inc.*, No. 6:22-cv-00122-ADA (W.D. Tex. filed Feb. 2, 2022) and *Zentian Ltd. v. Amazon.com*, No. 6:22-cv-00123-ADA (W.D. Tex. filed Feb. 2, 2022), as well as co-pending inter partes review proceedings IPR2023-00034 to -00037. Pet. 77; Paper 4, 1.

B. THE '319 PATENT (EX. 1001)

The '319 patent describes “a speech recognition circuit which uses parallel processors for processing the input speech data in parallel.” Ex. 1001, 1:4–6. The processors analyze models of words having common prefix components, and store lexical data in a lexical memory. *Id.* at code (57). The processors are arranged in clusters, each with its own “partial lexical memory” for storing part of the lexical data. *See id.* at 2:64–3:7.

Figure 2, reproduced below, is a schematic diagram of the speech-recognition circuit.

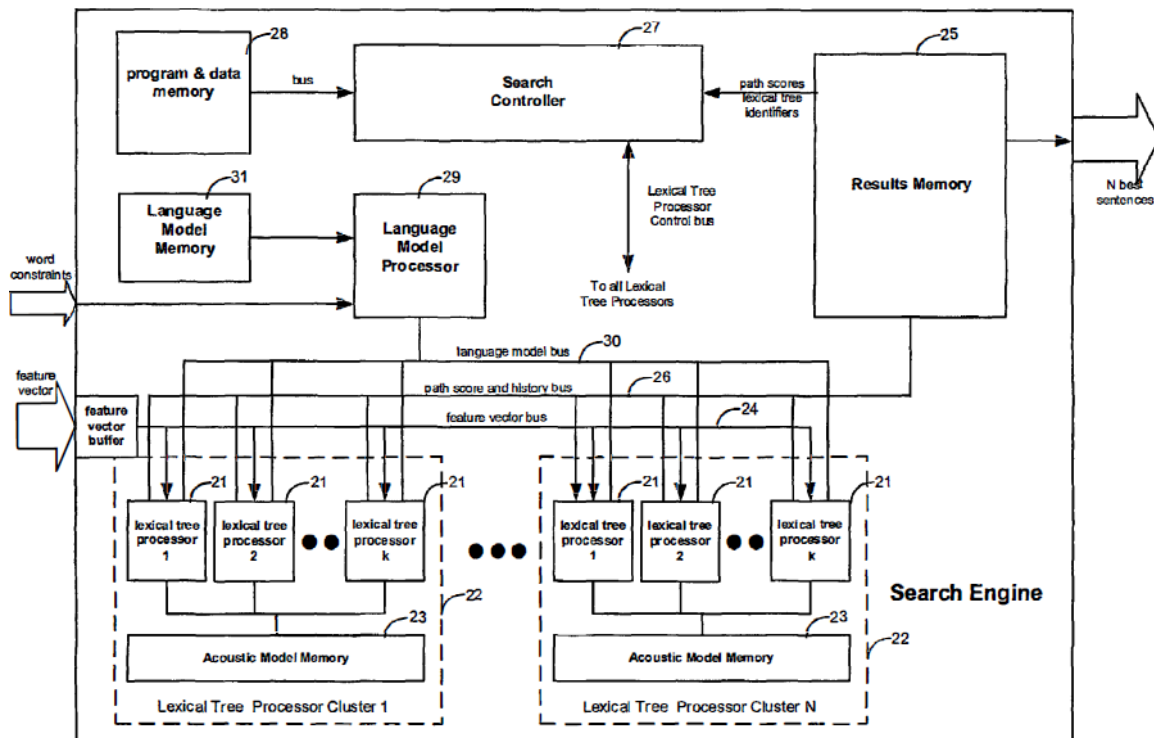


Fig 2

As shown above in Figure 2, there is a feature vector buffer (middle left) for receiving a set of speech data as feature vectors. *See* Ex. 1001, 4:55–57. Clusters 22 (of which there are N), each having k lexical tree processors 21 (and thus $N \times k$ total processors) read the feature vectors from feature vector

bus 26, process the feature vectors, and store the resulting lexical data in an acoustic model memory (23) that is part of each cluster. *Id.* at 4:65–5:13.

“[A]coustic model memories 23 store as a whole a complete set of lexical data, i.e. lexical tree data structures for use in the lexical tree processing by the lexical tree processors 21,” and “[e]ach acoustic model memory 23 contains part or a segment of the lexical tree data.” *Id.* at 5:6–10. For each cluster 22, “lexical tree processors 21 . . . access the same acoustic model memory 23.” *Id.* at 5:10–12.

C. CHALLENGED CLAIMS AND GROUNDS

Claim 46, the sole challenged independent claim, is as follows:

46. A speech recognition circuit comprising:

- [a] an input buffer receiving processed speech parameters;
- [b] a plurality of lexical memories containing in combination complete lexical data for word recognition, each lexical memory containing part of said complete lexical data;
- [c] a plurality of processors connected in parallel to said input buffer for processing the speech parameters in parallel,
- [d] said processors being arranged in groups of processors, each group of processors being connected to a lexical memory;
- [e] a control processor controlling each processor to process said speech parameters using partial lexical data read from a respective said lexical memory; and
- [f] a results memory storing the results of the processing of the speech parameters from said processors.

Ex. 1001, 15:55–16:2 (Apple’s reference numbers added).

Apple argues four grounds for *inter partes* review, as summarized in the following table:

Claim(s) Challenged	35 U.S.C. §	Reference(s)/Basis
46, 50, 54, 64, 67	103(a) ²	Thelen, ³ Bailey, ⁴ Chen ⁵
47–49	103(a)	Thelen, Bailey, Chen, Huang ⁶
53, 58, 59	103(a)	Thelen, Bailey, Chen, Jiang ⁷
51	103(a)	Thelen, Bailey, Chen, Katz ⁸

Pet. 6.

D. EXPERT TESTIMONY

As expert testimony, Apple submits a declaration by Christopher Schmandt. Ex. 1003. Zentian also submits an expert declaration by Dr. David Anderson. Ex. 2020.

Zentian does not challenge Mr. Schmandt’s qualification to provide expert testimony, or the admissibility of that testimony. *See* PO Sur-reply 13 (“[T]he issue here is not admissibility but rather credibility.”). Likewise, Apple does not challenge Dr. Anderson’s expert qualification or the admissibility of his testimony. *See* Pet. Reply. Neither party has submitted a motion to exclude any evidence.

² 35 U.S.C. § 103(a) (2006), *amended* by Leahy–Smith America Invents Act, Pub. L. No. 112-29 § 103, sec. (n)(1), 125 Stat. 284, 287, 293 (2011) (effective Mar. 16, 2013). This version of § 103 applies because the ’319 patent issued from an international application filed on February 4, 2003, which is before the effective date of the AIA amendments. *See* Ex. 1001, code (22).

³ Thelen et al., US 6,526,380 B1 (issued Feb. 25, 2003) (Ex. 1030).

⁴ Bailey et al., US 5,459,798 (issued Oct. 17, 1995) (Ex. 1031).

⁵ Chen et al., US 5,428,803 (issued June 27, 1995) (Ex. 1005).

⁶ Huang et al., US 6,076,056 (issued June 13, 2000) (Ex. 1034).

⁷ Jiang, US 6,374,219 B1 (issued Apr. 16, 2002) (Ex. 1004).

⁸ Katz, US 4,831,550 (issued May 16, 1989) (Ex. 1035).

Zentian does, however, argue that Mr. Schmandt's expert testimony is entitled to little credit because he, himself, has "*never* built the processor to memory architecture of *any* speech recognition system identified in his declaration." PO Resp. 10–12 (citing Ex. 2017, 32:3–9, 34:24–35:5, 106:10–23; Ex. 1003 ¶ 84); PO Sur-reply 13.

Apple responds that Mr. Schmandt's credibility does not depend on whether he has built or supervised building the claimed invention because he is a qualified expert, is at least a person of ordinary skill in the art, and has the necessary technical experience in the field to opine on the subject matter of his declaration. Pet. Reply 6–7. According to Apple, Mr. Schmandt has taken "courses covering multiprocessor architectures, parallelism, concurrency, and memory contention," has "co-authored a paper in 1993 on hardware architectures for speech recognition," and "authored a textbook including a chapter . . . discussing the various computer architectures for speech recognition." *Id.* at 7–8 (citing Ex. 1003 ¶¶ 109, 113; Ex. 1012, 9).

We agree, for the reasons Apple has argued, that Mr. Schmandt's education and experience are sufficient to provide credible testimony on the subject matter of his declaration, including computer architectures for speech recognition.

III. GROUNDS OF THE PETITION

For the reasons below, we determine that Apple has shown, by a preponderance of the evidence, that claims 46–51, 53, 54, 58, 59, 64, and 67 of the '319 patent are unpatentable as obvious.

A. LEVEL OF ORDINARY SKILL IN THE ART

The level of ordinary skill in the pertinent art at the time of the invention is a factor in how we construe patent claims. *See Phillips v. AWH Corp.*, 415 F.3d 1303, 1312–13 (Fed. Cir. 2005) (en banc). It is also one of the factors we consider when determining whether a patent claim is obvious over the prior art. *See Graham v. John Deere Co.*, 383 U.S. 1, 17–18 (1966).

To assess the level of ordinary skill, we construct a hypothetical “person of ordinary skill in the art,” from whose vantage point we assess obviousness and claim interpretation. *See In re Rouffet*, 149 F.3d 1350, 1357 (Fed. Cir. 1998). This legal construct “presumes that all prior art references in the field of the invention are available to this hypothetical skilled artisan.” *Id.* (citing *In re Carlson*, 983 F.2d 1032, 1038 (Fed. Cir. 1993)).

Relying on Mr. Schmandt’s testimony, Apple argues that a person of ordinary skill in the art would have had “a master’s degree in computer engineering, computer science, electrical engineering, or a related field, with . . . two years of experience in the field of speech recognition, or a bachelor’s degree in the same fields with . . . four years of experience in the field of speech recognition.” Pet. 5 (citing Ex. 1003 ¶¶ 24–26).

Zentian does not contest Apple’s proposed definition of the level of ordinary skill. PO Resp. 5; *see also* Pet. Reply 5–6 (citing Ex. 2017, 9:20–13:4). Zentian also acknowledges that such a person of ordinary skill would have had knowledge more generally “in the area of digital signal processing (DSP).” PO. Resp. 16 (citing Ex. 2020 ¶ 39). However, Zentian contends that “a specialist in speech recognition . . . would not also be expected to have specialized in parallel processing architectures and methods or high performance computing in addition to speech recognition.” *Id.* at 5–6, 16

(citing Ex. 2020 ¶¶ 17–19, 39); *see also* PO Sur-reply 3 (citing Ex. 2020 ¶ 39).

As evidence of this, Zentian argues that, although its expert Dr. Anderson exceeds the uncontested level of ordinary skill in the art as an expert in speech recognition, he, along with a team of engineers and a DSP expert from Texas Instruments, could not successfully implement a custom computer vision application within a particular multi-processor system supplied by Texas Instruments. PO Resp. 16–17 (citing Ex. 2020 ¶¶ 40–41); PO Sur-reply 5, 12–13 (citing Ex. 2020 ¶¶ 40–41). Zentian also argues that “Dr. Anderson has never taught a course that would have equipped a masters-level student in speech recognition to apply a parallel processing architecture . . . to known speech recognition techniques.” PO Resp. 6 (citing Ex. 2020 ¶ 19).

Apple counters that Dr. Anderson’s personal alleged lack of experience or skill in parallel processing is not relevant to the level of ordinary skill, as “[p]erhaps Dr. Anderson simply did not have the necessary skills to lead a parallel processing project in an unrelated field.” Pet. Reply 8. Apple also argues that, although a person of ordinary skill need not have “specialized” in parallel processing architectures or high-performance computing, there is evidence in the cited art that an ordinarily skilled artisan “would have been knowledgeable in both high performance computing and parallel processing.” *Id.* at 9. For example, Apple argues that Thelen and Bailey, which are in the field of speech recognition, are “directed to parallel processing architectures.” *Id.* at 10 (citing Ex. 1030, code (57); Ex. 1031, code (57)); Pet. 7 (arguing that Thelen and Bailey are speech-recognition references).

Because a person of ordinary skill in the art is a hypothetical person, we agree with Apple that Dr. Anderson’s inability to implement a particular computer vision application within the constraints of a particular Texas Instruments parallel-processing environment does not mean that more general knowledge and skill in parallel processing would have been beyond the level of ordinary skill. To the contrary, the evidence of record suggests that a person of ordinary skill in the art at the time of the claimed invention, who, for example, had a master’s degree and two years of experience in the field of speech recognition, would have also had knowledge and experience with parallel processing.

For example, the ’319 patent asserts that, at the time of the claimed invention, parallel processing had already been explored “to speed up the processing performed during . . . a search in a speech recognition system.” Ex. 1001, 1:30–43. The ’319 patent characterizes the claimed invention as an “improved circuit” that can perform such “parallel processing of speech parameters.” *Id.* at 1:44–46.

Also, we find persuasive Apple’s argument that Thelen and Bailey present problems and solutions within the speech-processing field involving parallel processing architectures. Ex. 1030, 2:38–39 (“[U]sing parallel speech recognition engines is known.”); Ex. 1031, 5:31–35 (describing “a specially optimized multiprocessing hardware unit capable of performing, in parallel, a multitude of steps required for pattern recognition procedures”); *Best Med. Int’l, Inc. v. Elekta Inc.*, 46 F.4th 1346, 1353 (Fed. Cir. 2022) (The “type of problems encountered in the art” and “prior art solutions to those problems” are “factors that may guide the fact finder in finding the appropriate level of skill in the art.”).

We have considered Dr. Anderson’s opinion that a “person of ordinary skill would not also be expected to have *specialized* in parallel processing architectures and methods or high performance computing *in addition to* speech recognition.” Ex. 2020 ¶ 18 (emphasis added). But we find this of little relevance, because the evidence of record indicates that the use of parallel-processing techniques were *an aspect of* the speech recognition field at the time of the claimed invention. Thus, a person of ordinary skill would have had at least the background necessary to apply parallel processing to a search in a speech recognition system as mentioned in the background section of the ’319 patent (Ex. 1001, 1:30–43), or to perform speech-recognition tasks in parallel using separate speech recognition engines or processors as in Thelen and Bailey (Ex. 1030, 2:30–33; Ex. 1031, 5:31–35).

Thus, we adopt Apple’s uncontested definition of the level of ordinary skill in the art as a master’s degree in computer engineering, computer science, electrical engineering, or a related field, with two years of experience in the field of speech recognition, or a bachelor’s degree in the same fields with four years of experience in the field of speech recognition. We also find that this level of ordinary skill would have included education and experience necessary to apply speech-recognition techniques in a parallel-processing environment.

B. CLAIM CONSTRUCTION

In an *inter partes* review, we construe a patent claim “using the same claim construction standard that would be used to construe the claim in a civil action under 35 U.S.C. 282(b).” 37 C.F.R. § 42.100(b) (2023). This generally includes “construing the claim in accordance with the ordinary and

customary meaning of such claim as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.” *Id.* The ordinary and customary meaning of a claim term “is its meaning to the ordinary artisan after reading the entire patent,” and “as of the effective filing date of the patent application.” *Phillips*, 415 F.3d at 1313, 1321. There are only two circumstances in which a construction departs from the ordinary and customary meaning: “1) when a patentee sets out a definition and acts as [their] own lexicographer, or 2) when the patentee disavows the full scope of a claim term either in the specification or during prosecution.” *Thorner v. Sony Comput. Entm’t Am. LLC*, 669 F.3d 1362, 1365 (Fed. Cir. 2012). Any such special meaning of a term “must be sufficiently clear in the specification that any departure from common usage would be so understood by a person of experience in the field of the invention.” *Multiform Desiccants Inc. v. Medzam Ltd.*, 133 F.3d 1473, 1477 (Fed. Cir. 1998).

To construe the claim terms, “we look principally to the intrinsic evidence of record, examining the claim language itself, the written description, and the prosecution history, if in evidence.” *DePuy Spine, Inc. v. Medtronic Sofamor Danek, Inc.*, 469 F.3d 1005, 1014 (Fed. Cir. 2006).

Apple states that it “applies the plain and ordinary meaning of all claim terms,” and does not propose any explicit constructions. *See* Pet. 7. Likewise, Zentian does not propose any explicit constructions. *See* PO Resp.

We do not find it necessary to explicitly construe any terms for this decision. To the extent that we need to interpret claim language, we address that language below in the context of the prior art. *See Nidec Motor Corp. v. Zhongshan Broad Ocean Motor Co.*, 868 F.3d 1013, 1017 (Fed. Cir. 2017) (“[W]e need only construe terms ‘that are in controversy, and only to the

extent necessary to resolve the controversy’” (quoting *Vivid Techs., Inc. v. Am. Sci & Eng’g, Inc.*, 200 F.3d 795, 803 (Fed. Cir. 1999))).

C. GROUNDS BASED ON THELEN, BAILEY AND CHEN (CLAIMS 46, 50, 54, 64, AND 67)

A claim is unpatentable under § 103 for obviousness “if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains.” 35 U.S.C. § 103; *see also KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 406 (2007).

We base our obviousness inquiry on factual considerations including (1) the scope and content of the prior art, (2) any differences between the claimed subject matter and the prior art, (3) the level of skill in the art, and (4) any objective indicia of obviousness or non-obviousness that may be in evidence. *See Graham*, 383 U.S. at 17–18.

Considering these factors,⁹ we determine that Apple has shown, by a preponderance of the evidence, that claims 46, 50, 54, 64, and 67 are unpatentable under § 103(a) as obvious over Thelen in view of Bailey and Chen.

1. *Overview of Thelen*

Thelen describes “[a] huge vocabulary speech recognition system for recognizing a sequence of spoken words.” Ex. 1030, code (57). Figure 3, reproduced below, is a block diagram of this system:

⁹ Neither party presents evidence of objective indicia of obvious or non-obviousness, so no such evidence factors into our decision.

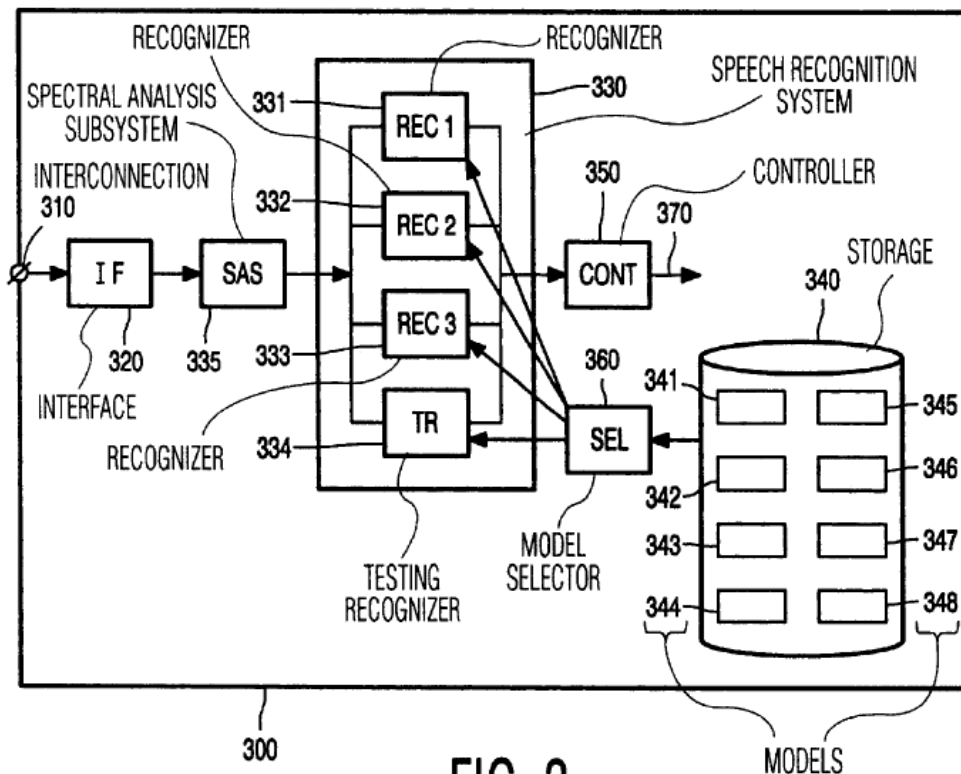


FIG. 3

The above figure depicts standalone system 300 with interconnection 310 and interface 320 for receiving a user's speech signal from, for example, a microphone or recording. Ex. 1030, 6:47–63. Spectral analysis subsystem (SAS) 335 processes the digital speech signal to produce a feature vector, which is fed to recognition system 330. *Id.* at 4:30–33, 6:63–66, 7:2–6.

Recognition system 330 “comprises a plurality of large vocabulary speech recognizers [331, 332, and 333], each being associated with a respective, different large vocabulary recognition model” that can be loaded from storage 340. Ex. 1030, 1:56–61, 6:66, 7:5–7, 7:34–35. “The recognizers may use the same algorithm, where the difference lies in the models being used, such as the vocabulary and the language model.” *Id.* at 7:7–9. “The recognizers operate in ‘parallel,’ in the sense that they independently recognize the same speech input at almost the same moment.

This can be realized by using separate resources for each of the recognizers, such as a separate processor.” *Id.* at 7:17–21.

Recognition models can be “hierarchically arranged from models with a more generic context to models with a more specific context.” Ex. 1030, 8:54–56. For example, there could be a general “sport” model and also more specific models for “football, baseball, athletics, car racing, etc.” *Id.* at 9:24–26. “[M]odel selector 360 is operative to enable recognition with a more specific model if recognition with a certain model obtains good recognition results,” so that the “more specific model (i.e. hierarchically lower) may be used as a replacement for the more generic model.” *Id.* at 9:8–12. Likewise, “[i]f recognition with a specific model at a certain moment gives low results, recognition is preferably continued with at least one model hierarchically above the specific model.” *Id.* at 9:29–32.

2. *Overview of Bailey*

Bailey describes a pattern-recognition system, useful for voice recognition, that “includes a specially optimized multiprocessing hardware unit capable of performing, in parallel, a multitude of steps required for pattern recognition procedures.” Ex. 1031, 5:27–34. Figures 3A and 3B, reproduced below, are block diagrams depicting two embodiments of this system:

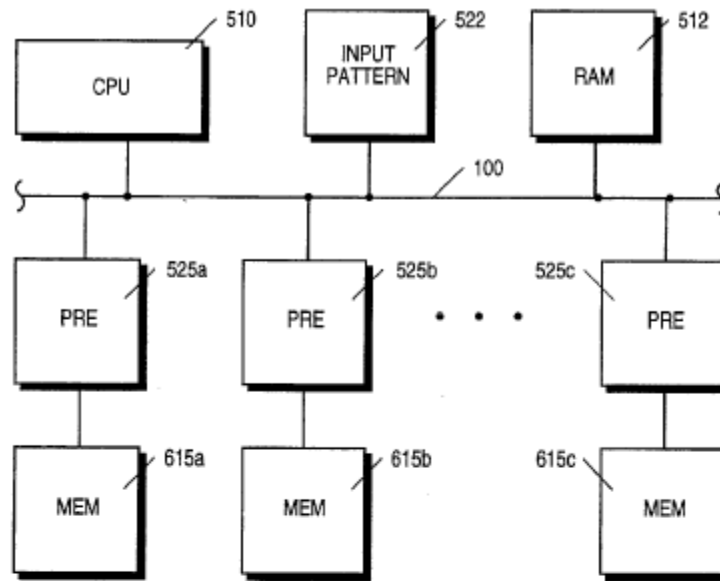


FIG. 3A

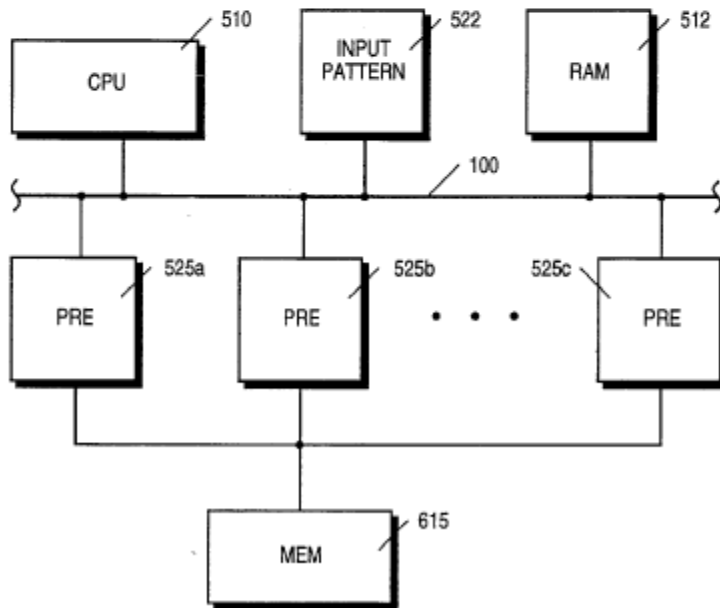


FIG. 3B

As shown above, the pattern-recognition system in each of Figures 3A and 3B includes CPU 510, memory 512, and input pattern block 522 coupled to host communication bus 100. Ex. 1031, 11:11–14. Input pattern block 522 obtains a voice sample from a microphone. *Id.* at 11:17–20. Also connected

to bus 100 are a “number of [pattern recognition] engines, 525a to 525c, each independently coupled to . . . bus 100.” *Id.* at 11:20–22. In Figure 3A, each of pattern recognition engines 525a–525c is also “interfaced with its own private memory 615a–615c.” *Id.* at 11:1–3. Alternatively, pattern recognition engines 525a–525c may share common memory 615 as shown in Figure 3B. *Id.* at 11:3–4.

3. *Overview of Chen*

Chen is directed to “parallel processing computer systems for performing multiple-instruction-multiple-data (MIMD) parallel processing.” Ex. 1005, 1:29–31. Chen organizes the MIMD multiprocessors into four or more physically separable clusters, each of which can uses a shared memory. *Id.* at 1:32–43.

Figure 4, reproduced below, is a block diagram of Chen’s system:

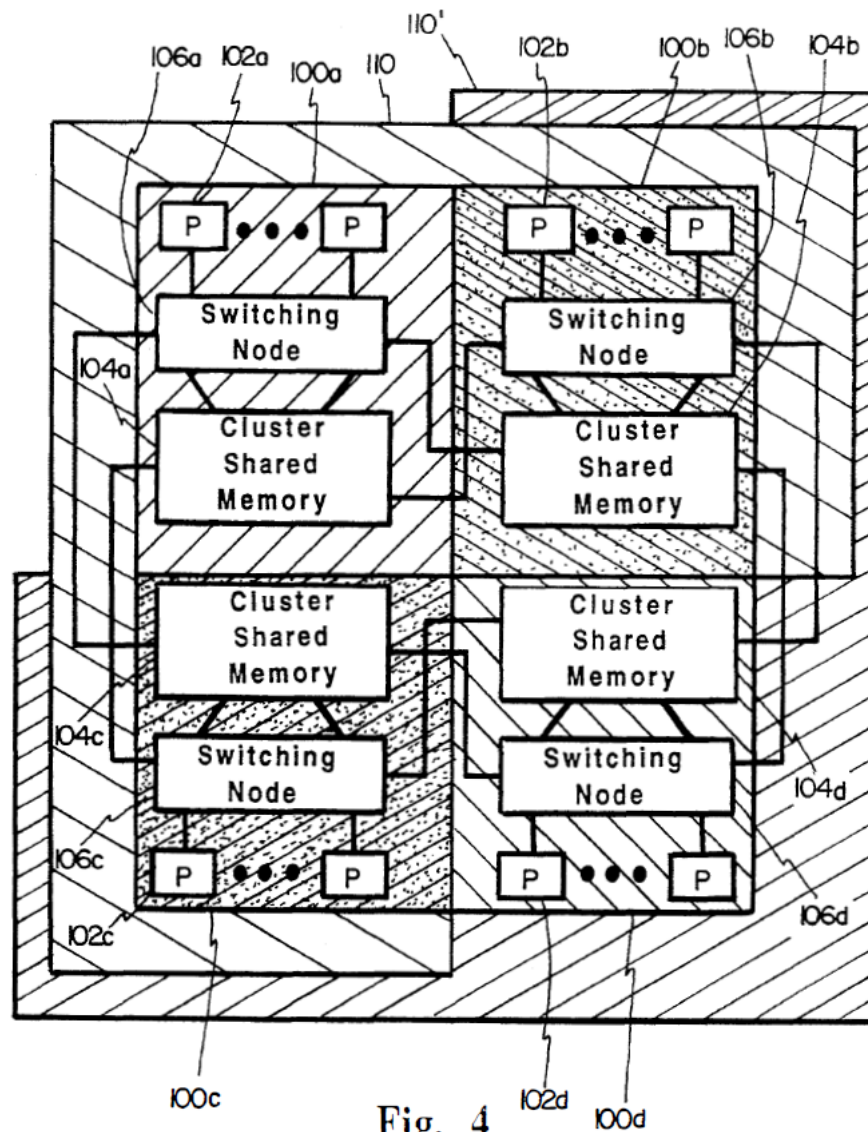


Fig. 4

Figure 4 depicts four clusters 100a–d, each with processors 102a–d and cluster shared memory 104a–d, where the clusters may be connected together in various shared or distributed memory configurations. Ex. 1005, 9:20–56. In particular, “[a] true shared memory model with symmetrical access to a common shared memory 104a is provided for all processors 102a in cluster 100a.” *Id.* at 9:46–48.

Chen teaches that this configuration can effectively use “many different types of processors,” such as commercially available “single chip

microprocessors” and memory, or alternatively, the system can be created by retrofitting “existing shared memory and extended shared memory multiprocessor systems.” Ex. 1005, 10:14–25.

4. *Independent Claim 46*

(a) Preamble and Limitations 46(a)–(c)

The preamble of claim 46 and limitation 46(a) recite “[a] speech recognition circuit comprising: an input buffer receiving processed speech parameters.” Ex. 1001, 15:55–56.

“To the extent the preamble is limiting,” Apple contends that Thelen “teaches or at least renders obvious a speech recognition circuit, as claimed” because it “teaches a speech recognition apparatus including a microphone, A/D converter, and processors.” Pet. 9–11 (citing Ex. 1030, code (57), 1:5–12, 2:30–37, 6:46–66, 7:20–26, Fig. 3; Ex. 1003 ¶¶ 68–69).

For limitation 46(a), Apple contends that Thelen’s recognition system 330 receives “observation vectors” from spectral analysis subsystem 335 which correspond to the recited “processed speech parameters.” *See* Pet. 12–16 (citing Ex. 1030, 4:29–43, 4:66–5:2, 6:53–7:7, 7:17–29, Figs. 1, 3; Ex. 1003 ¶¶ 70–75). Apple concedes that “*Thelen* does not expressly recite an ‘input buffer’” that received the processed speech parameters, but according to Apple, Bailey’s RAM 512 would constitute the recited buffer. Pet. 17. In particular, Apple points to Bailey’s teaching that an “unknown input *pattern* may be *temporarily stored* in memory array 512” which, Apple contends, would suggest to a person of ordinary skill in the art that RAM 512 is acting as an input buffer. Pet. 19 (quoting Ex. 1031, 11:16–17) (citing Ex. 1003 ¶ 72). According to Apple, there would have been motivation to

incorporate Bailey's input buffer into Thelen's system because such input "[b]uffers were well-known, inexpensive, and easy-to-implement solutions to improve system efficiency." Pet. 21 (citing Ex. 1003 ¶ 74).

Limitation 46(b) recites "a plurality of lexical memories containing in combination complete lexical data for word recognition, each lexical memory containing part of said complete lexical data." Ex. 1001, 15:57–59. According to Apple, Thelen describes dividing a huge vocabulary (corresponding to the "complete lexical data for word recognition") into separate recognition models, each of which contains only part of the complete lexical data. *See* Pet. 23–30 (citing Ex. 1030, code (57), 1:5–12, 1:53–2:33, 5:14–45, 5:49–52, 6:12–39, 7:17–23, 7:30–52, 8:1–5, 8:54–59, Figs. 1, 3; Ex. 1003 ¶¶ 76–78).

Apple argues that Thelen "does not expressly teach that each of its recognition models and associated lexicons (the mapped 'parts' of said complete lexical data) are stored in a plurality of lexical memories." Pet. 30–31. But according to Apple, Thelen "envision[s] [that] each recognizer (including its processing resources, such as processors and memory) may be associated with an exclusive recognition model in a 'fixed one-to-one relationship.'" Pet. 31 (citing Ex. 1030, 2:30–33, 7:42–49; Ex. 1003 ¶ 78). Apple also points to Bailey's Figure 3A embodiment, in which each parallel processor explicitly has its own memory. Pet. 30–32 (citing Ex. 1031, code (57), 1:16–20, 2:62–66, 10:18–11:44, 13:25–51, Fig. 3A; Ex. 1003 ¶ 79).

Limitation 46(c) recites "a plurality of processors connected in parallel to said input buffer for processing the speech parameters in parallel." Ex. 1001, 15:60–61. Apple identifies the recited processors with Thelen's recognizers 331–333 and argues that they are designed to process

the speech input in parallel, and that in the Thelen–Bailey combination they would be connected to the input buffer. *See* Pet. 33–38 (citing Ex. 1030, 2:30–37, 2:38–40, 7:2–7, 7:6–9, 7:17–23, 7:26–29, Figs. 2, 3; Ex. 1031, 3:60–4:1, 9:36–41, 11:9–29, 13:25–51, 14:15–29, Figs. 2, 3A; Ex. 1003 ¶¶ 81–82).

Zentian does not specifically contest that Thelen and Bailey teach the preamble and limitations 46(a)–(c), or that there would have been a motivation to combine Thelen and Bailey as per these limitations. *See* PO Resp. Because the evidence of record shows that Thelen teaches a speech recognition circuit (*see, e.g.*, Ex. 1030, Fig. 3 (system 300), 6:46–49 (stating that system 300 “preferably is implemented on a computer, such as a PC”)), we need not decide whether the preamble is limiting.

We also find Apple’s arguments persuasive that a person of ordinary skill in the art would have had reason to use an input buffer as per limitation 46(a), such as that exemplified in Bailey, to improve efficiency in processing a speech input, and that Thelen’s recognizers 331–333 would have been connected to that input buffer to process the input in parallel, as per limitation 46(c).

Finally, we find persuasive Apple’s arguments that Thelen and Bailey teach that the lexical memories contain, in combination, a complete set of lexical data as per limitation 46(b).

(b) Limitation 46(d)

Limitation 46(d) requires “said processors being arranged in groups of processors, each group of processors being connected to a lexical memory.” *Id.* at 15:62–64. We address Apple’s contentions for this limitation below, along with Zentian’s responsive arguments.

(1) *Apple's Proposed Combination and
Motivation to Combine*

For limitation 1(d), Apple relies on the modifying Thelen–Bailey in view of Chen.

Apple relies on Chen for teaching the arrangement of processors into groups, each group having its own shared memory. *See* Pet. 38–46.¹⁰ According to Apple, Chen teaches arranging the processing resources into clusters 100a–d, each comprising multiple processors 102a–d and shared memory 104a–d. Pet. 38–40 (citing Ex. 1005, 5:9–17, 6:17–21, 9:5–43, 10:14–35, Figs. 4, 5A, 6A–B; Ex. 1003 ¶ 83). In the proposed combination, “*Chen's* groups (clusters) of processors provid[e] the same functionality” as the single processors in Thelen–Bailey. Ex. 1003 ¶ 84.

Apple concedes that Chen “does not specifically teach a ‘lexical’ memory,” but proposes a modification of the Thelen–Bailey combination in which “each group of processors (embodied in recognizers) includes a memory, where the memory is a *lexical* memory as already taught by *Thelen–Bailey*.” Pet. 40. According to Apple, a person of ordinary skill in the art would have made this combination “for the purpose of associating the generic and specific contexts (e.g., all sports contexts) together in a group for efficient word recognition” which would “store all similar contexts on a lexical memory connected to the group of recognizers directed to those contexts.” Pet. 40–41 (citing Ex. 1030, 7:17–58, 8:1–3, 8:6–8, 8:54–9:32; Ex. 1031, Figs. 3A–B; Ex. 1003 ¶¶ 80, 84–85).

¹⁰ Apple also notes that Bailey’s Figure 3B embodiment suggests “connecting a group of recognition engines with a lexical memory.” Pet. 40; *see also* Pet. 32–33 (citing Ex. 1031, 11:1–4, Fig. 3B); Ex. 1003 ¶ 83.

Apple contends that Thelen already teaches dividing lexical data into a hierarchy of topics where “each recognizer is associated with one recognition model at a time,” and according to Apple, where multiple processors are assigned to similar subcontexts (e.g., multiple “sports” subcontexts), it would have been a logical extension that each such processor would have access to a shared memory “to facilitate more efficient control and processing.” Pet. 42–44 (citing Ex. 1005, code (57); Ex. 1030, 7:30–34, 8:54–56, 8:62–9:11, Fig. 4; Ex. 1003 ¶ 84).

In other words, “[i]t would have been a common[-]sense extension of *Thelen*’s grouped, hierarchical recognition models to similarly group the recognizers (per *Chen*), where each recognizer is associated with a recognition model (per *Thelen*), and each group of recognizers (and associated processors) is connected to a lexical memory, per *Bailey*.” Pet. 44 (citing Ex. 1003 ¶ 84). Apple argues that this grouping “would have yielded the predictable result of consolidating speech recognition for a particular context within a particular group of recognition processors” and “would have more efficiently allowed for sharing of lexical data common to each of the recognition models.” *Id.* (citing Ex. 1003 ¶ 84).

Apple also contends that grouping the processors with shared memory in this way “would have yielded the predictable result of reducing the time required to recognize a user’s speech, desirably speeding the system’s response.” Pet. 44 (citing Ex. 1003 ¶ 85). Apple argues that Thelen recognizes the importance, in real-time speech processing, ““that no significant delay occurs’ between a word being received in the system and recognizing the word,” and would have “simply substituted” *Chen*’s “more

efficient processing architecture” for the Thelen–Bailey recognizers. Pet. 44–45 (quoting Ex. 1030, 7:27) (citing Ex. 1030, 7:20–26; Ex. 1003 ¶ 85).

In its Response, Zentian first argues that Apple’s proposed motivation to modify Thelen to increase recognition speed “assumes that the clustered processor and memory arrangement of Chen would have necessarily improved the speed and power of Thelen’s speech recognition,” but that Apple has provided no analysis to why that would have been the case, or how much of a speed gain there would have been. *Id.* at 22–23; *see also* PO Sur-reply 18–19 (citing Ex. 2017, 137:21–138:24, 141:23–142:2, 144:22–145:3). And according to Zentian, the proposed modification would not actually have added processors to Thelen’s system. *Id.* at 23–24 (citing Ex. 2017, 130:23–131:4, 131:11–22, 136:3–6).

Zentian also argues that in another particular multiprocessing speech recognition system contemporary to the claimed invention, increasing the number of processors did not speed up the search stage of speech recognition, so knowing about this system, a person of ordinary skill in the art would not have expected that increasing the number of processors would lead to an improvement. PO Resp. 25–26 (citing Ex. 2021, 11; Ex. 2020 ¶¶ 52–55).

In addition, Zentian argues that using Chen’s hardware to run Thelen’s recognition system would have resulted in memory collisions, and an ordinarily skilled artisan would have expected this to slow things down, thus nullifying Apple’s proposed motivation that the combination would increase the processing speed. PO Resp. 28–30 (citing Ex. 1003 ¶¶ 83–84; Ex. 2017, 124:13–126:5, 155:7–17; Ex. 2020 ¶¶ 59–60).

And for Apple’s argument that the combination would have increased flexibility and scalability, Zentian argues that Apple has only shown an increase in flexibility and scalability for one particular prior-art system, and this does not provide any general motivation that would have been applicable to increasing flexibility and scalability in the Thelen–Bailey combination. PO Resp. 30–31 (citing Ex. 1003 ¶¶ 71–72; Ex. 2017, 105:8–106:4; Ex. 2020 ¶ 61); PO Sur-reply 23 (reiterating the argument). Zentian contends that Thelen was already flexible, scalable, and performing adequately, so there would have been no motivation to complicate that system by adding features of Bailey or Chen. PO Resp. 31 (citing Ex. 2017, 163:4–12; Ex. 2020 ¶ 62).

Apple counters that Zentian never explains why memory collision problems or slower processing “would [have been] present in the [Thelen–Bailey–Chen] combination but not the clustered-processor speech recognition system of Claim 46, nor does Zentian acknowledge that the claims do not recite real-time recognition.” Pet. Reply 14. Apple contends that a person of ordinary skill in the art would not have thoughtlessly added processors to the system, but would have exercised ordinary skill to address memory and speed issues that were well-known at the time of the claimed invention. *Id.*; *see also id.* at 18 (arguing that tradeoffs between advantages and disadvantages do not obviate a motivation to combine (citing Ex. 1003 ¶ 84; *Allied Erecting and Dismantling Co. v. Genesis Attachments, LLC*, 825 F.3d 1373, 1381 (Fed. Cir. 2016))).

Apple also argues that Chen includes substantial guidance about balancing the tradeoffs of parallel processing and memory access. Pet. Reply 15 (citing Ex. 1005, 5:47–53, 14:39–18:42). According to Apple, Chen

teaches that there can be a speedup when there are multiple processors in each cluster simultaneously performing word recognition on a specific part of the lexical data, and indicates that parallel processing can be implemented such that tradeoff is “balanced at the point at which the performance of inter-cluster memory access start to significantly degrade.” *Id.* at 15–16 (quoting Ex. 1005, 5:51–53) (citing Ex. 1005, 5:47–53, 14:39–18:42; Ex. 1003 ¶¶ 81, 84–85).

Apple argues that this would have indicated to one of skill in the art that one would not have to sacrifice performance within a *single* cluster, given that memory access within a cluster is faster than memory access between clusters. Pet. Reply 16–17 (citing Ex. 1005, 4:58–63, 5:33–41). According to Apple, Chen also teaches that its clustered architecture “*does not have to sacrifice performance* for the increased interconnection necessary to increase the size of the problem solution space provided by an extended shared memory model.” *Id.* at 16 (quoting Ex. 1005, 5:36–40) (citing Ex. 1005, 2:24–29, 4:17–28, 4:44–49, 5:33–41, 8:42–47, 8:58–63).

Further, Apple argues that its burden of persuasion “does not require building the combination and testing its efficacy, or that the combination be optimal or preferred.” Pet. Reply 22–23 (citing *In re Mouttet*, 686 F.3d 1322, 1332 (Fed. Cir. 2012); *Intel Corp. v. PACT XPP Schweiz AG*, 61 F.4th 1373, 1381 (Fed. Cir. 2023); *In re Nievelt*, 482 F.2d 965, 968 (CCPA 1973); *Elekta Ltd. v. ZAP Surgical Sys., Inc.*, 81 F.4th 1368, 1374 (Fed. Cir. 2023)). Apple points to Mr. Schmandt’s testimony that using Chen’s cluster arrangement would have resulted in cost savings because it could have used commercially-available processors, and would have added flexibility and

scalability to the Thelen–Bailey design. *Id.* at 23 (citing Ex. 1006, 19–20; Ex. 1003 ¶¶ 58, 83–84).

Apple also contends that its proposed combination is not contingent on adding more processors to Thelen’s design, but on a clustering arrangement that, among other things, increases processing power and cost savings due to reducing the duplication of lexical data, advantages that Apple contends are unrebutted in Zentian’s Response. Pet. Reply 24–25 (citing Ex. 2017, 121:3–136:6, 137:2–20, 138:2–24, 140:8–141:14; Ex. 1003 ¶¶ 83–84).

Finally, Apple argues that, contrary to Zentian’s characterization, Mr. Schmandt did not testify that it was *never* desirable to complicate an already-working speech recognition system with additional hardware. Pet. Reply 25 (citing PO Resp. 31; Ex. 2017, 163:4–12). According to Apple, Mr. Schmandt’s testimony was in the context of a small-vocabulary recognizer, unlike the large-vocabulary systems addressed by the Thelen–Bailey–Chen combination. *Id.* at 25–26 (citing Ex. 2017, 163:4–164:22; Ex. 1030, 7:30–35; Ex. 1003 ¶ 84).

In its Sur-reply, Zentian argues that it has no burden to show why memory collision problems would have been present in the Thelen–Bailey–Chen combination but not in the ’319 patent. PO Sur-reply 16. But according to Zentian, the ’319 patent does not require the assignment of multiple sub-context models to the same processing cluster, and thus would have avoided the alleged problems associated with Apple’s proposed combination. *Id.* at 16–17 (citing Ex. 1003 ¶ 83; Ex. 2017, 124:13–126:5; Ex. 2020 ¶¶ 43–48).

Zentian also points to the testimony of Dr. Anderson, and cross-examination admissions by Mr. Schmandt that allegedly show that using multiple processors within a single cluster would have resulted in a slowdown, not a speedup, because there would have been memory contention issues between processors in each cluster. *See* PO Sur-reply 19–22 (citing Ex. 2020 ¶¶ 44–46, 59–60; Ex. 2002 ¶¶ 53–55; Ex. 2017, 120:19–121:2, 130:23–131:4, 136:3–6, 155:7–17; Ex. 2021, 11; Ex. 2020 ¶¶ 51–53, 59–60, 62).

As to Apple’s allegation that there would have been cost savings from using commercially available components in the proposed combination, Zentian argues that Mr. Schmandt “did not undertake any qualitative or quantitative cost-benefit analysis in reaching his opinions regarding a motivation to combine.” PO Sur-reply 23 (citing Ex. 2017, 42:14–19, 63:8–15).

We find Apple’s arguments persuasive notwithstanding Zentian’s arguments to the contrary. We credit Mr. Schmandt’s testimony that consolidating the overlapping data within similar lexical models (such as “sports” models like football or baseball) and storing the consolidated data in a shared memory that multiple processors may access in parallel would have resulted in memory and processing efficiency. Ex. 1003 ¶¶ 83–84. This would have been an improvement over Thelen, where each processor has access to only one, distinct lexical model at a time and there is no shared memory in which multiple processors can access the same consolidated models data in parallel. *See* Ex. 1030, 7:30–34.¹¹

¹¹ The dissent characterizes Thelen’s common storage 340 as a shared memory. However, each of Thelen’s recognizers only has access (indirectly

The evidence of record suggests that the ordinarily skilled artisan would not have been dissuaded from the memory-efficiency benefits of modifying Thelen–Bailey by grouping processors and memory into clusters because memory collisions would have allegedly made recognition slower. Although memory collisions may have been an issue for the ordinarily skilled artisan to consider, Apple has persuasively shown that there are other motivating considerations, such as the known flexibility¹² and scalability of distributed processing systems in general, decreased cost because off-the-shelf components may be used, and the added efficiency of data sharing among related sub-context models. *See* Pet. 40–45.

Apple persuasively shows, and we find Mr. Schmandt’s supporting testimony credible, that a skilled artisan would have expected to be able to minimize the effect of memory collisions based on Chen’s teachings and the background knowledge in the art. *See* Pet. 14–19. For the same reasons, we do not find credible Dr. Anderson’s testimony (*see, e.g.*, Ex. 2020 ¶¶ 44–46)

via model selector 360) to the one particular model to which it has been assigned at any given time. *See* Ex. 1030, 7:30–34.

¹² The dissent suggests that Apple’s proposed combination is less flexible than Thelen because it removes Thelen’s ability to swap a processor’s lexical model for another model in the hierarchy. We disagree. Thelen describes two embodiments: one in which “each recognizer [is] associated with an exclusive recognition model in a fixed one-to-one relationship,” and another in which there are “more models than active recognizers.” Ex. 1030, 44–48; *see also id.* at 7:62–9:32. Apple’s combination relies on the former embodiment, where there would have been no need to swap out models because the processors are already accessing all models in parallel. *See* Ex. 1003 ¶¶ 77 (citing Ex. 1030, 7:42–49), 79 (explaining that the combination includes a “1-to-1 relationship” between processors and recognition models), 80 (each processor in Thelen–Bailey independently recognizes speech in parallel).

that a person of ordinary skill in the art would not have been able to increase the number of processors within a cluster without resulting in a slowdown with respect that cluster or the system as a whole.

Next, Zentian argues that implementing Thelen’s speech recognition techniques in Chen’s architecture would have “require[d] *significant* reprogramming effort” because, as is necessarily the case with distributed memory models, processes and data must be partitioned into separate tasks that can fit in each local memory but still be large enough to compensate for the overhead involved in passing messages within the distributed environment. PO Resp. 26–27 (quoting Ex. 1005, 2:8–9) (citing Ex. 1005, 2:6–17; Ex. 1003 ¶ 77; Ex. 1030, 7:30–34, 5:22–24, Fig. 3; Ex. 2020 ¶ 56). Zentian contends that Apple has not addressed this issue, and that as a result of this issue, a person of ordinary skill in the art would have had no motivation to implement Thelen’s recognition system using a cluster-based model. PO Resp. 27–28 (citing Ex. 2020 ¶¶ 57–58); *see also* PO Sur-reply 23–24 (reiterating the argument).

Apple responds that in building a speech recognizer, a person of ordinary skill in the art would have needed “to undertake reasonable programming efforts no matter the recognizer’s architecture,” so this would simply be *programming*, not *reprogramming*. Pet. Reply 20. And according to Apple, even if reprogramming were necessary, this “does not negate obviousness nor is finding the optimal approach required.” *Id.* (citing *Elekta*, 81 F.4th at 1374). Apple also disputes Zentian’s contention that the Thelen–Bailey–Chen combination would have required memory access between adjacent processing clusters, so any significant reprogramming effort to divide tasks between each cluster’s memory would not have been necessary.

Id. at 20–22 (citing Ex. 1005, 1:56–62, 1:66–2:1, 2:6–9, 2:1–29, 5:20–25; Ex. 1004, 7:40–43; Ex. 1003 ¶¶ 75–76, 83).

We find Apple’s arguments persuasive that programming the modified Thelen–Bailey–Chen system would have been within the ordinary skill in the art at the time of the claimed invention. *See infra* Section III.C.4(b)(2) (finding that a person of ordinary skill in the art would have been capable of addressing the complications of implementing Thelen–Bailey–Chen). Thus, any need for substantial programming or reprogramming would not have dissuaded a person of ordinary skill in the art from modifying Thelen–Bailey by grouping processors into clusters with shared memory.

Considering the evidence as a whole presented by both sides, we find Apple’s arguments persuasive that a person of ordinary skill in the art would have had reason to modify Thelen–Bailey so that its parallel recognizers are grouped in clusters with shared memory in view of Chen. Apple has persuasively shown, for example, that there would have been a motivation to facilitate more efficient control and processing because related recognition models (e.g., sub-context models within a context) may be assigned to a single cluster with multiple processors and processed together to allow for sharing of lexical data common to each of the recognition models.

Moreover, Thelen and Bailey present particular arrangements of memory and processing elements to achieve the known benefits of parallel processing, such as reduced cost, efficiency, flexibility, and scalability. Chen simply presents an alternative parallel-processing arrangement of memory and processing elements that address the same known problems. Apple has persuasively shown that modifying Thelen–Bailey by grouping processors into clusters with shared memory would have addressed known problems

using prior art elements according to their established functions. *See Intel v. PACT*, 61 F.4th at 1380.

(2) *Reasonable Expectation of Success*

Apple argues that there would have been a reasonable expectation of success in modifying Thelen–Bailey according to Chen’s teachings because this combination “would have only required arrangement of well-known memory components and relocation of lexical data already present in *Thelen*.” Pet. 45–46 (citing Ex. 1030, 2:30–37, 7:2–9, 7:26–29; Ex. 1031, 10:62–65, 11:4–8, 51:21–26; Ex. 1003 ¶ 85).

In response, Zentian first argues that “Chen . . . is not a reference in the field of speech recognition,” and “teaches nothing as to how to implement speech recognition techniques on its clustered parallel processing architecture, much less Thelen’s speech recognition techniques.” PO Resp. 8 (citing Ex. 2017, 120:19–121:2; Ex. 2020 ¶ 30). Zentian also contends that Thelen “does not teach how to run its speech recognition techniques on groups of clustered processors such as Chen’s.” *Id.* (citing Ex. 2017, 136:7–16; Ex. 2020 ¶ 31). According to Zentian, Apple and its expert Mr. Schmandt have not shown that, or how, a person of ordinary skill in the art would have been able to modify Thelen based on Chen’s teachings. *Id.* at 8–9 (citing Ex. 2017, 137:21–138:24, 141:23–142:2, 144:3–20, 144:22–145:3). Zentian also argues that Mr. Schmandt’s testimony is conclusory. *Id.* at 9–10 (citing Ex. 1003 ¶ 84; *TQ Delta, LLC v. Cisco Sys., Inc.*, 942 F.3d 1352, 1362–63 (Fed. Cir. 2019)).

Zentian also contends that modifying the Thelen–Bailey combination as Apple proposes would have been too complex for a person of ordinary skill in the art. PO Resp. 12 (citing Ex. 2020 ¶ 33). According to Zentian,

although Bailey teaches assigning each processor its own private memory (Fig. 3A), or alternatively assigning all processors to a shared memory (Fig. 3B), Bailey does not teach dividing processors into multiple groups, each with its own shared memory. *Id.* at 12–14 (citing Ex. 1031, 11:1–3, Figs. 3A, 3B; Ex. 1001, 3:50–54; Ex. 2017, 119:8–13; Ex. 2020 ¶¶ 33–37).

Zentian contends that grouping the processors, as Chen teaches, would have involved “avoiding memory conflicts, controlling task sharing in an efficient manner, addressing communication bandwidth and latency issues among the various hardware components, and developing a messaging strategy to coordinate information sharing between and within clusters,” as well as “developing hardware and software for centralized control of the individual processors, memories, and clusters” and making a real-time speech recognition system “in the sense that no significant delay occurs” during word recognition, all of which were beyond the level of ordinary skill in the art. PO Resp. 15–18 (quoting Ex. 1003 ¶ 84) (citing Ex. 2020 ¶¶ 38–42); *id.* at 21–22 (citing Ex. 1003 ¶ 84; Ex. 2021, 11; Ex. 2020 ¶¶ 46–47); *see also* PO Sur-reply 1–5 (citing Ex. 2021, 11; Ex. 2017, 32:3–9, 34:24–35:5, 106:10–23, 120:8–18, 120:19–121:2, 124:13–125:7, 136:7–16, 137:21–138:24, 141:23–142:2, 144:3–20, 144:22–145:3, 145:20–24, 154:8–155:17; Ex. 2020 ¶¶ 17–19, 27, 30–31).

In particular, Zentian argues that there would have been memory-access problems with Apple’s proposal to assign the models for each of Thelen’s “contexts” and corresponding “sub-contexts” to a single cluster, along with the subset of vocabulary associated with each context and sub-context. PO Resp. 18–20 (citing Ex. 1003 ¶ 83; Ex. 1030, 9:24–26; Ex. 2017, 124:13–126:5; Ex. 2020 ¶¶ 41, 43). But according to Zentian, this

would have required the processors in each cluster to access multiple sets of model data simultaneously—one model for each sub-context—and this would have led to memory collisions, with the result that the combination could not have been expected to operate in real time. *Id.* at 20–21 (citing Ex. 2020 ¶¶ 44–45); *see also* PO Sur-reply 1–2, 4 (reiterating the same argument).

Apple counters that the ’319 patent, itself, does not teach any of the details that Zentian contends were beyond the level of ordinary skill in the art, such as avoiding memory collisions, controlling task sharing, dealing with bandwidth limitations, or reducing latency. Pet. Reply 2. This, according to Apple, suggests that the person of ordinary skill, who is the intended audience of the ’319 patent, would have been capable of selecting between known methods to implement the elements of the claimed invention. *Id.* at 2–4 (citing Ex. 1001, 3:50–54; *Uber Techs., Inc. v. X One, Inc.*, 957 F.3d 1334, 1339 (Fed. Cir. 2020); *In re Epstein*, 32 F.3d 1559, 1568 (Fed. Cir. 1994)).

Apple also contends that the challenged claims do not require solutions to Zentian’s proposed complications. Pet. Reply 4–5 (citing Ex. 2017, 23:6–25, 26:2–27:3, 27:9–28:2). And according to Apple, Zentian has only alleged generic complications, without pointing to specific implementation problems in the proposed combination. *Id.* at 5. Apple contends that it is under no obligation “to predict and proactively explain every unrecited and generic implementation detail” that could arise when “performing known speech recognition techniques on clustered processors.” *Id.* (citing *Facebook, Inc. v. Windy City Innovations, LLC*, 973 F.3d 1321, 1343 (Fed. Cir. 2020)).

Finally, Apple argues that at the time of the claimed invention, a person of ordinary skill in the art would have known of various speech-recognition hardware architectures, and would have been concerned with the alleged complicating issues that Zentian discusses, as “[s]peech recognition researchers recognized issues and tradeoffs of speed, accuracy, power, scalability, monetary cost, ease of building, flexibility, expandability, miniaturization, etc.” Pet. Reply 10–11 (citing Ex. 1006, 6, 17–19; Ex. 2021, 17; Ex. 1007, 92, 111; Ex. 1012, 151–152); *id.* at 15–19 (arguing that Chen teaches that one may successfully address the tradeoff between parallel processing and memory access).

Zentian responds that Apple’s argument assumes, without showing, that omissions in the ’319 patent disclosure mean that the omitted material would have been within the level of ordinary skill at the time of the claimed invention, an assumption that amounts to an improper hindsight bias. PO Sur-reply 5–7 (citing *Al-Site Corp. v. VSI Int’l, Inc.*, 174 F.3d 1308, 1324 (Fed. Cir. 1999)). Zentian also distinguishes the ’319 patent from the Thelen–Bailey–Chen combination because “the ’319 Patent does not teach placing all of Thelen’s sub-context models for a given context into the *same* single cluster memory,” which would have resulted in “memory collisions that would [have] prevent[ed Apple]’s goal (and Thelen’s teaching) of real time speech recognition.” *Id.* at 7 (citing Ex. 1003 ¶ 83; Ex. 2017, 124:13–126:5; Ex. 2020 ¶¶ 43–45); *see also id.* at 11–14 (arguing that even knowledge of parallel architectures would have been insufficient to implement Apple’s specific proposed combination with real-time performance).

We disagree with Zentian that, to show there was a reasonable expectation of success, Apple must show how the proposed combination would have overcome various alleged complications such as memory collisions, task sharing issues, bandwidth limitations, or latency. “The reasonable expectation of success requirement refers to the likelihood of success in combining references to meet the limitations of the *claimed invention*.” *Intelligent Bio-Sys., Inc. v. Illumina Cambridge Ltd.*, 821 F.3d 1359, 1367 (Fed. Cir. 2016) (emphasis added). The claimed invention includes no recitations about the number of memory collisions, how task sharing should occur, how bandwidth is addressed, or any maximum latency, so even if the combination of Thelen, Bailey, and Chen would have reasonably been expected to result in a high-latency, inefficient system, this would still have been a “success” in terms of how the ’319 patent defines the claimed invention.

We also disagree with the argument in Zentian’s Sur-reply that Apple must show that the proposed combination would have achieved real-time speech recognition because, otherwise, “the reasonable expectation of success inquiry would collapse into the question of whether the claims are enabled.” PO Sur-reply 17–18 (citing Ex. 1003 ¶ 84; Ex. 2020 ¶ 42; Ex. 2017, 154:8–155:17; *Arctic Cat Inc. v. Bombardier Rec. Prods.*, 876 F.3d 1350, 1360–61 (Fed. Cir. 2017) (“[I]f an ordinarily skilled artisan would not believe that a particular combination would have a reasonable expectation of ‘anticipated success,’ the combination may not be obvious.”)). While the ability or inability of Apple’s proposed combination to have achieved real-time performance may be relevant to the motivation to combine, whether there was a reasonable expectation of success is based on the limitations as

claimed, and the challenged claims include no requirement for real-time performance. *See Intelligent Bio-Sys.*, 821 F.3d at 1367. *Arctic Cat* does not suggest otherwise; in that case, the question was whether an ordinarily skilled artisan would have anticipated success for the proposed prior-art combination in meeting the key limitations of the *claimed invention*. *See* 876 F.3d at 1363–64.

We also agree with Apple that the '319 patent's silence as to how to implement an architecture that deals with these alleged complications is strong evidence that solving these problems would have been within the level of ordinary skill in the art. *Epstein*, 32 F.3d at 1568 (“[T]he Board's observation that appellant did not provide the type of detail in his specification that he now argues is necessary in prior art references supports the Board's finding that one skilled in the art would have known how to implement the features of the references”). Zentian contends that *Epstein* is inapplicable because it addressed whether individual references were enabling rather than whether there was an expectation of success in combining the references. PO Sur-reply 9. But we do not consider that to be a meaningful distinction. In *Epstein*, as here, the question is what the patent applicant's omission suggests about what a skilled artisan would have already known and could have implemented without further disclosure.

As we discuss above, we find that the level of ordinary skill would have included education and experience necessary to apply speech-recognition techniques in a parallel-processing environment. *See supra* Section III.A. The '319 patent's omission of detail about addressing Zentian's alleged complications supports this understanding. Although we have considered Dr. Anderson's testimony that addressing such

complications would have been beyond the level of ordinary skill (*see, e.g.*, Ex. 2020 ¶¶ 38–48), this opinion is premised on an understanding of the level of ordinary skill in the art that we do not find credible, based on the evidence of record. *See id.* ¶ 18.

(3) *Enablement of the Prior Art*

In its Response, Zentian argued that the prior art, as a whole, must enable a skilled artisan to make and use the claimed invention. PO Resp. 7 (citing *Raytheon Techs. Corp. v. GE*, 993 F.3d 1374, 1380–81 (Fed. Cir. 2021)). But as Apple correctly notes, Zentian “d[id] not raise a non-enablement argument nor provide a *Wands* factors analysis.” Pet. Reply 12 (citing *In re Wands*, 858 F.2d 731, 737 (Fed. Cir. 1988)).

Nevertheless, according to Apple, “*Chen* already teaches an operable clustered processor architecture; the combination simply requires utilizing the architecture for speech recognition according to *Thelen*’s acoustic model hierarchy and employing *Bailey*’s parallel speech recognizers.” Pet. Reply 12–13; *see also id.* at 15 (arguing that Chen “is enabling for what it teaches, including a clustered computing system that balances the tradeoffs of parallel processing and memory access”). Apple argues that “Zentian’s alleged complications are specific to a clustered processor architecture and would occur regardless of the type of information being processed (e.g., for speech recognition or other uses). *Chen* further directly envisions accounting for memory latency.” *Id.* at 13.

In its Sur-reply, Zentian argues that Apple has not shown that Thelen, Bailey, and Chen enable the claimed invention, and that Chen, standing alone, does not enable using its architecture for speech recognition. PO Sur-reply 14–15 (citing Ex. 2017, 120:8–121:2, 136:7–16; Ex. 2020 ¶¶ 30–48; *In*

re Antor Media Corp., 689 F.3d 1282, 1287–88 (Fed. Cir. 2012); *Raytheon*, 993 F.3d at 1380–81).

Although Zentian mentioned enablement as a general legal requirement in its Response, it did not specifically raise the issue with respect to the facts of this case, and also did not properly or sufficiently raise the issue in its Sur-reply. Thus, we find the argument untimely. *See* Paper 11, 10 (Scheduling Order warning that “any arguments for patentability not raised in the response may be deemed waived.”); 37 C.F.R. § 42.23(b) (“A sur-reply may only respond to arguments raised in the corresponding reply”). In any event, we find Apple’s arguments persuasive that the prior art references, as a whole, enable the claimed invention.

(c) Limitation 46(e) and 46(f)

Limitation 46(e) recites “a control processor controlling each processor to process said speech parameters using partial lexical data read from a respective said lexical memory.” Ex. 1001, 15:65–67.

Apple associates the recited control processor with Thelen’s “model selector 360 in combination with controller 350,” and also Bailey’s CPU 510. Pet. 46–49 (citing Ex. 1030, 7:50–61, 7:63–8:1, 8:36–41, 9:8–10, 10:20–24, Fig. 3; Ex. 1031, 9:5–39, 11:4–8, 11:23–34, 11:59–64, 12:12–25, 13:5–23, Figs. 3A–B; Ex. 1003 ¶¶ 86–87). Apple contends that it would have been obvious to modify the Thelen–Bailey–Chen system to include Bailey’s CPU 510 for controlling the processors because it was well-known in the art to use a master processor to control a group of other processors. Pet. 50 (citing Ex. 1031, 12:2–11; Ex. 1003 ¶¶ 86–88).

Limitation 46(f) recites “a results memory storing the results of the processing of the speech parameters from said processors.” Ex. 1001, 16:1–2.

Apple identifies the recited results memory with Bailey’s RAM 512, and argues that a person of ordinary skill in the art would have been motivated to add such a results memory to Thelen because of a recognized need “to report outcomes of the comparisons.” Pet. 51–54 (citing Ex. 1030, 7:50–58; Ex. 1031, 11:23–34, 11:42–59, 12:2–11, 12:41–61, Fig. 4; Ex. 1003 ¶¶ 89–91).

Zentian does not specifically contest these arguments for limitations 46(e) and (f). *See* PO Resp. We find Apple’s arguments persuasive for the reasons outlined in the Petition, and determine that there would have been a reason to combine Thelen and Bailey in the manner Apple proposes for these limitations.

(d) Conclusion as to Claim 46

For all the above reasons, we find Apple’s arguments sufficiently persuasive that the combination of Thelen, Bailey, and Chen teaches all the limitations of claim 46 and that a person of ordinary skill in the art at the time of the claimed invention would have had reason to combine the references as Apple proposes, with a reasonable expectation of success. Thus, we conclude that claim 46 is unpatentable under § 103(a) as obvious over the combination of Thelen, Bailey, and Chen.

5. *Claim 50*

Claim 50 depends from claim 46 and further recites “wherein each processor compares the processed speech parameters with the lexical data in

a corresponding lexical memory to identify words as a word recognition event and sends information identifying the identified words to said results memory as the processing results.” Ex. 1001, 16:19–24.

Apple argues that in the Thelen–Bailey–Chen combination, each recognizer compares the received unknown input pattern (the “processed speech parameters”) with the recognition model (the “lexical data”) stored in a corresponding memory, and as Bailey teaches, sends the results to the system memory (the “results memory”). Pet. 54–55 (citing Ex. 1030, code (57), 7:50–52; Ex. 1031, 11:42–56, 12:41–61, 12:46–48, 15:41–45). Apple argues that a person of ordinary skill in the art would have recognized the interrupts that are generated after each recognizer completes its comparison as a signal to transfer results to the system memory, and would have been motivated to do so because there was a need in the modified system to indicate that the speech recognition is complete, and the use of interrupts were a well-known way of doing this. *Id.* at 55 (citing Ex. 1030, 7:50–58; Ex. 1031, 11:42–59, 12:41–61, 15:43–45; Ex. 1003 ¶¶ 92–93).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claim 50. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claim 50 (Pet. 54–55), we determine that Apple has shown by a preponderance of the evidence that claim 50 is unpatentable as obvious over Thelen in view of Bailey and Chen.

6. *Claim 54*

Claim 54 depends from claim 46 and further recites “wherein said processors process said speech parameters by comparing said speech

parameters with said lexical data or data derived from lexical data in respective said lexical memories.” Ex. 1001, 16:37–41.

Apple argues that Thelen teaches matching observation vectors from the input speech (the “speech parameters”) against a speech recognition unit which is part of the lexical data stored in each lexical memory. Pet. 56 (citing Ex. 1030, 5:2–52).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claim 54. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claim 54 (Pet. 56), we determine that Apple has shown by a preponderance of the evidence that claim 54 is unpatentable as obvious over Thelen in view of Bailey and Chen.

7. *Claim 64*

Claim 64 depends from claim 46 and further recites “wherein said input buffer is arranged to receive said speech parameters as feature vectors.” Ex. 1001, 17:12–14.

Apple argues that Thelen’s speech parameters are feature vectors, and that in the Thelen–Bailey–Chen combination, this would have been what the input buffer receives. Pet. 57 (citing Ex. 1030, 4:12–37; Ex. 1031, 9:1–39, Figs. 2, 3A).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claim 64. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claim 64 (Pet. 57), we determine that Apple has shown by a preponderance of the evidence that claim 64 is unpatentable as obvious over Thelen in view of Bailey and Chen.

8. *Claim 67*

Claim 67 depends from claim 46 and further recites “wherein said lexical memories store said lexical data as Hidden Markov Models, and each processor performs the Viterbi search algorithm using a respective part of said lexical data.” Ex. 1001, 17:22–25.

Apple argues that Thelen teaches using a Hidden Markov Model and that Bailey specifically teaches using the Viterbi search algorithm. Pet. 57–58 (citing Ex. 1030, 5:27–42; Ex. 1031, 7:51–8:36, Fig. 1B; Ex. 1003 ¶ 94). Apple contends that a person of ordinary skill would have applied the Viterby algorithm using *each* lexical tree processor “to maximize the benefit gained and maintain consistency between the lexical tree processors,” and each processor would only be acting on the lexical data to which is assigned. Pet. 58 (citing Ex. 1030, 5:29–31; Ex. 1003 ¶ 95).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claim 67. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claim 67 (Pet. 57–58), we determine that Apple has shown by a preponderance of the evidence that claim 67 is unpatentable as obvious over Thelen in view of Bailey and Chen.

D. GROUND BASED ON THELEN, BAILEY, CHEN, AND HUANG
(CLAIMS 47–49)

For its second ground, Apple argues that claims 47–49 are unpatentable under § 103(a) as obvious over Thelen in view of Bailey, Chen, and Huang. Pet. 6, 59–67.

Claim 47 depends from claim 46 and further recites “wherein each lexical memory stores lexical data as at least one lexical tree data structure, each lexical tree data structure comprises a model of words having common prefix components, and an initial component of each lexical tree data structure is unique.” Ex. 1001, 16:3–8.

Claim 48 depends from claim 47 and further recites “wherein each lexical memory stores said lexical tree data structure as an n phone model of words, where n is an integer, and said components comprise n phones.” Ex. 1001, 16:9–12.

Claim 49 depends from claim 48 and further recites “wherein each lexical memory stores said lexical tree data structure as a mono phone model of words, and said lexical tree processors use said mono phone models to generate context depend[e]nt phone models of words dynamically for use in processing the speech parameters.” Ex. 1001, 16:13–18.

Apple relies on its analysis for independent claim 46, which we discuss above, but argues that the Thelen–Bailey–Chen combination “does not . . . expressly teach structuring its lexical data in a lexical tree data structure.” Pet. 59. For this, Apple relies on Huang, which teaches representing lexical data for speech recognition in the form of “a tree structure wherein all of the words likely to be encountered by the system are represented in the tree structure.” Pet. 59–61 (quoting Ex. 1034, 2:35–37) (citing Ex. 1034, 2:31–58, 5:62–6:17, 7:21–52, 8:8–9, 10:16–33, 12:43–46, Figs. 1, 2; Ex. 1003 ¶ 96).

Apple argues that a person of ordinary skill in the art would have used Huang’s tree structure in Thelen–Bailey–Chen to reduce the magnitude of

computation required for speech recognition. Pet. 61–62 (citing Ex. 1031, 11:9–34, Fig. 3A; Ex. 1007, 68; Ex. 1003 ¶¶ 96–97).

In particular, for claims 48 and 49, Apple contends that Huang teaches building a lexical tree as a “n phone model” in which n is 1, which is an integer (i.e., a “mono phone model”), and generating context-dependent phone models dynamically. Pet. 62–63, 65–66 (citing Ex. 1034, 2:53–61, 6:62–8:20, 8:24–25, 11:64–12:46, Fig. 9; Ex. 1001, 6:40–7:35, 9:7–16, 11:18–20; Ex. 1003 ¶¶ 98–99, 101–102). According to Apple, a person of ordinary skill would have been motivated to do this “to minimize the amount of memory required for storing the trees” and “to minimize the computational load with the benefit of increased accuracy provided by accounting for the preceding phone.” Pet. 64, 66–67 (citing Ex. 1003 ¶¶ 99–100, 103–104).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claims 47–49. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claims 47–49 (Pet. 59–67), we determine that Apple has shown by a preponderance of the evidence that claims 47–49 are unpatentable as obvious over Thelen in view of Bailey, Chen, and Huang.

E. GROUND BASED ON THELEN, BAILEY, CHEN, AND JIANG (CLAIMS 53, 58, AND 59)

For its third ground, Apple argues that claims 53, 58, and 59 are unpatentable under § 103(a) as obvious over Thelen in view of Bailey, Chen, and Jiang. Pet. 6, 68–72.

Claim 53 depends from claim 51 (reproduced *infra* Section III.F) and further recites “wherein each processor determines and outputs scores for

words in the processing results at word recognition events.” Ex. 1001, 16:34–36.

Apple argues that Thelen and Bailey teach determining and outputting word scores from each recognizer, but Thelen–Bailey–Chen “does not expressly characterize this . . . process as a ‘word recognition event.’” Pet. 68 (citing Ex. 1030, 7:50–58; Ex. 1031, 12:12–15, 12:53–61). Apple thus relies on Jiang which, according to Apple, also teaches determining and outputting word scores, but does so at “word boundaries,” which Apple identifies as the recited “word recognition events.” Pet. 68–70 (citing Ex. 1004, 1:19–2:15, 2:47–67, 4:8–12, 6:39–42, 7:34–38, 7:48–8:4, 8:16–9:42, Fig. 4; Ex. 1003 ¶ 105). Apple argues that a person of ordinary skill in the art would have been motivated to output scores at word boundaries, as Jiang teaches, because this “would have provided the control processor with data necessary to evaluate word hypotheses at an appropriate step in the process (i.e., at the end of words).” Pet. 69–70 (citing Ex. 1003 ¶ 106).

Claim 58 depends from claim 46 and further recites “wherein said processors determine and output scores in the processing results during the processing of said speech parameters.” Ex. 1001, 16:56–59.

Claim 59 depends from claim 58 and further recites “wherein said processors prune the processing of the speech parameters in dependence upon the determined scores.” Ex. 1001, 16:60–62.

For these claims, Apple argues that Jiang teaches a pruning technique using the word scores, and that pruning in this way would have enabled computational savings and an improvement in efficiency. Pet. 71–72 (citing Ex. 1004, 2:25–46, 8:52–64, 10:51–11:24; Ex. 1003 ¶ 107).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claims 53, 58, or 59. For the reasons given above for claim 46 and the additional reasons articulated in the Petition for claims 53, 58, and 59 (Pet. 68–72), we determine that Apple has shown by a preponderance of the evidence that claims 53, 58, and 59 are unpatentable as obvious over Thelen in view of Bailey, Chen, and Jiang.

F. GROUND BASED ON THELEN, BAILEY, CHEN, AND KATZ (CLAIM 51)

For its fourth ground, Apple argues that claim 51 is unpatentable under § 103(a) as obvious over Thelen in view of Bailey, Chen, and Katz. Pet. 6, 73–75.

Claim 51 depends from claim 50 and further recites “including [a] language model processor providing a language model output for modifying the processing results at a word recognition event by . . . said processor.” Ex. 1001, 16:25–28.

Apple argues that Thelen teaches a sentence-level language model in addition to the word-level model discussed above in the context of claim 46, and thus modifies the processing results at the word-recognition level. Pet. 73 (citing Ex. 1030, 4:20–25, 5:54–6:10, 7:30–43; Ex. 1003 ¶¶ 108–109). According to Apple, Thelen “does not expressly state the language model output modifying the processing results is performed by a ‘language model processor,’” but Apple relies on Katz as teaching this limitation. Pet. 73–74 (citing Ex. 1035, 1:44–54, 2:34–44, 5:61–7:46, Figs. 2–3; Ex. 1003 ¶¶ 108–109). Apple argues that a person of ordinary skill in the art would have made this modification because Katz presents an operable structure for providing a language model output in this context, and that offloading the language

model processing to a separate processor would have conserved computational resources. Pet. 74–75 (citing Ex. 1030, 5:54–6:12; Ex. 1035, 1:44–54; Ex. 1003 ¶ 109).

Apart from its general arguments that we have discussed above, Zentian makes no particular arguments for claim 51. For the reasons given above for claims 46 and 50, and the additional reasons articulated in the Petition for claim 51 (Pet. 73–75), we determine that Apple has shown by a preponderance of the evidence that claim 50 is unpatentable as obvious over Thelen in view of Bailey, Chen, and Katz.

IV. CONCLUSION¹³

For the reasons above, Apple has shown by a preponderance of the evidence that claims 46–51, 53, 54, 58, 59, 64, and 67 of the ’319 patent are unpatentable under § 103(a).

¹³ Should Zentian wish to pursue amendment of claims in a reissue or reexamination proceeding after this decision, we draw attention to the April 2019 *Notice Regarding Options for Amendments by Patent Owner Through Reissue or Reexamination During a Pending AIA Trial Proceeding*. See 84 Fed. Reg. 16,654 (Apr. 22, 2019). If Zentian chooses to file a reissue application or a request for reexamination of the challenged patent, we remind Zentian of its continuing obligation to notify the Board of any such related matters in updated mandatory notices. See 37 C.F.R. § 42.8(a)(3), (b)(2).

In summary:

Claim(s)	35 U.S.C. §	Reference(s)/ Basis	Claims Shown Unpatentable	Claims Not Shown Unpatentable
46, 50, 54, 64, 67	103(a)	Thelen, Bailey, Chen	46, 50, 54, 64, 67	
47–49	103(a)	Thelen, Bailey, Chen, Huang	47–49	
53, 58, 59	103(a)	Thelen, Bailey, Chen, Jiang	53, 58, 59	
51	103(a)	Thelen, Bailey, Chen, Katz	51	
Overall Outcome			46–51, 53, 54, 58, 59, 64, 67	

V. ORDER

In consideration of the foregoing, it is

ORDERED that claims 46–51, 53, 54, 58, 59, 64, and 67 of the '319 patent are unpatentable;

FURTHER ORDERED that parties to this proceeding seeking judicial review of our decision must comply with the notice and service requirements of 37 C.F.R. § 90.2.

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC., AMAZON WEB SERVICES, INC., and
AMAZON.COM SERVICES LLC,
Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00033¹⁴
Patent 7,587,319 B2

Before KEVIN F. TURNER, JEFFREY S. SMITH, and
CHRISTOPHER L. OGDEN, *Administrative Patent Judges*.

SMITH, *Administrative Patent Judge*, dissenting.

¹⁴ IPR2023-01192 has been joined with this proceeding.

The majority, in this Final Written Decision (FWD), finds that Petitioner has shown that a person of ordinary skill in the art would have modified Thelen using the teachings of Bailey and Chen to produce the advantages of (1) the efficiency of sharing data among related sub-context models, (2) flexibility and scalability of distributed processing systems, and (3) decreased cost by using off-the-shelf components. FWD 27–31; *see* Pet. 40–45. I disagree. I would have found that Petitioner does not show any reason why a person of ordinary skill in the art would have modified Thelen to include the teachings of Bailey. I also would have found that Petitioner has not shown any reason to modify its combination of Thelen and Bailey to include the teachings of Chen. As discussed below, (1) Thelen already provides the efficiency of sharing data among related sub-context models by organizing its speech recognition models into a hierarchical data structure that shares data common to multiple speech recognition models, (2) Thelen is already scalable and flexible, and (3) Petitioner has not shown that Thelen’s components would have been more expensive than Chen’s components. Therefore, I respectfully dissent.

(1) The Efficiency of Sharing Data

To support the contention that the efficiency of sharing data among related sub-context recognition models provides motivation, Petitioner contends that grouping context-related recognizers together with a shared memory storing all context-related recognition models would have more efficiently allowed for sharing of lexical data common to each of the recognition models. Pet. 44 (citing Ex. 1003 ¶ 84). Petitioner contends that this grouping of recognizers into a cluster provides the benefit of reducing the duplication of lexical data, necessitating less memory to operate, further

reducing the overall financial cost of the system and increasing processing power. Pet. Reply 24 (citing, *inter alia*, Ex. 1003 ¶¶ 83, 84).

Petitioner's contentions are based on the premise that the combination of Thelen and Bailey would have required more memory to store Thelen's lexical models than the combination of Thelen, Bailey, and Chen. *See* Pet. 40–45. However, Petitioner does not provide any reason replace Thelen's single shared memory which efficiently stores models in a hierarchical data structure with Bailey's multiple private memories, which cannot. *See id.* at 33, 40–45. Petitioner also does not show that the combination of Thelen, Bailey and Chen would have used less memory than Thelen alone. In the following paragraphs, I briefly summarize the teachings of Thelen and Bailey and address Petitioner's failure to provide a reason for combining Thelen and Bailey. Then, I briefly summarize Chen and address Petitioner's failure to provide a reason for combining Thelen, Bailey, and Chen.

Figure 3 of Thelen is reproduced below.

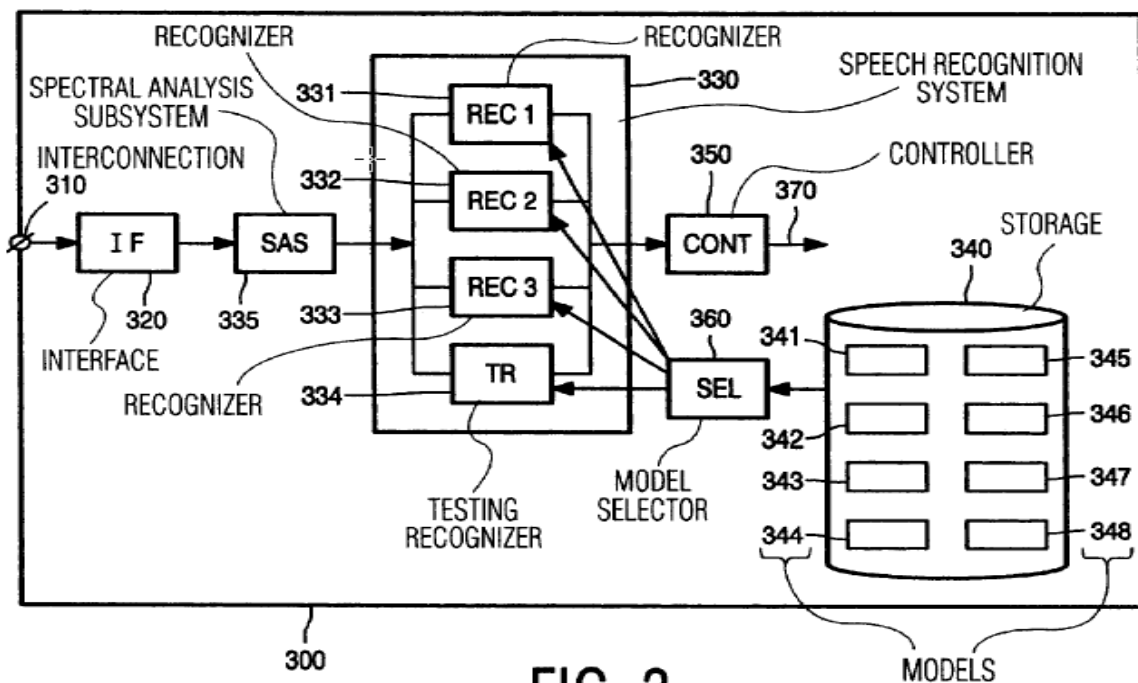


FIG. 3

Figure 3 above shows a block diagram of Thelen's speech recognition system. Ex. 1030, 4:15–16. As shown in Figure 3, storage 340 stores recognition models 341 to 348 which are used by recognizers 331 to 334. *Id.* at 7:5–7, 7:30–49, 8:40–42.

Figure 3A of Bailey is reproduced below.

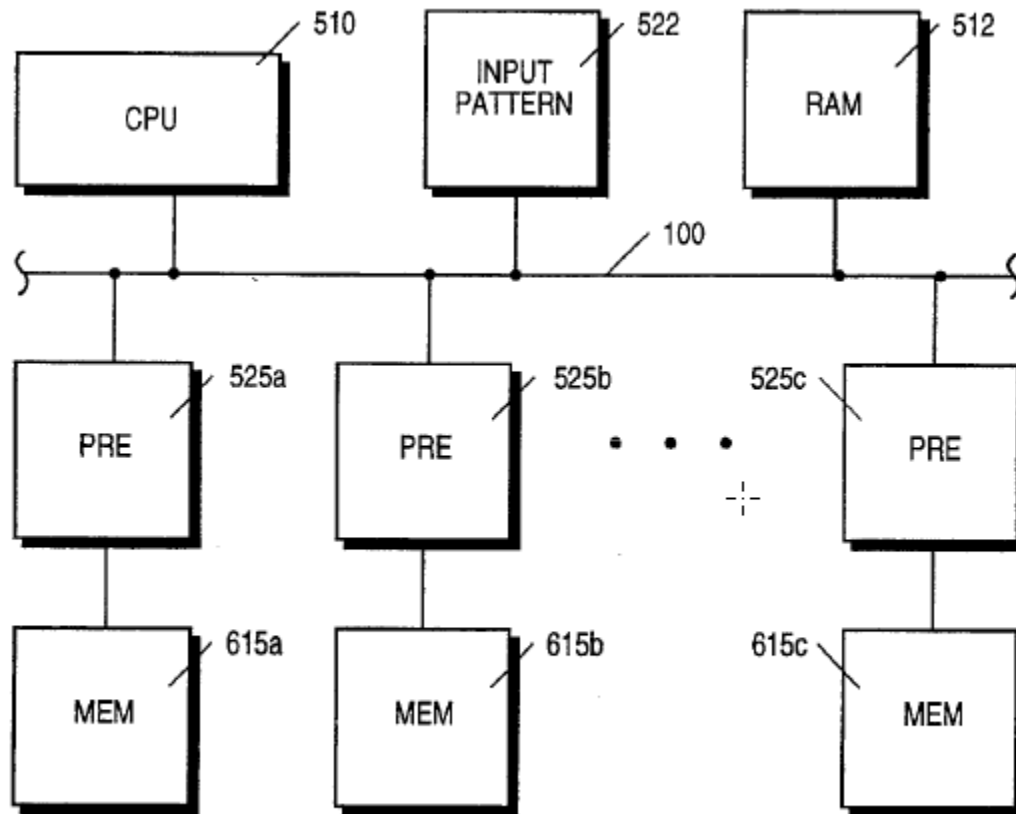


FIG. 3A

Figure 3A above shows a block diagram of a system for pattern recognition employing a multiprocessing apparatus with private memories. Ex. 1031, 4:44–47, Abstract. Each of processors 525a to 525c contains its own private memory 615a to 615c respectively, where each private memory stores a library of patterns. *Id.* at 11:9–23.

In combining Bailey with Thelen, Petitioner concedes that Thelen already discloses “a memory containing the complete lexical data” in describing that a lexicon and sub-word models are stored in storage 340. Pet. 30 (citing Ex. 1030, 7:30–34, 5:22–24, Fig. 3; Ex. 1003 ¶¶ 77–78). Petitioner further concedes that Thelen “does not expressly teach that each of its recognition models and associated lexicons . . . are stored in a plurality of lexical memories.” *Id.* at 30–31.

Petitioner replaces Thelen’s common storage 340 containing the complete lexical data with Bailey’s plurality of private memories, where each private memory is associated with one processor, and each of the private memories stores one of Thelen’s lexical models. Pet. 31–32, 40 (citing Ex. 1031, Fig. 3A). However, Petitioner has not explained why a person of ordinary skill in the art would have replaced Thelen’s single shared memory that “contains the complete lexical data” with Bailey’s private memories. The Petition states that a “motivation to combine Bailey’s plurality of lexical memories with Thelen is provided in Section V.A.5.b.” Pet. 33. Section V.A.5.b is entitled “Motivation to Combine Chen with Thelen-Bailey for Claims 46(b) and 46(d).” *Id.* at 40. This section discusses replacing Bailey’s individual private memories, where each private memory stores one of Thelen’s models, with Chen’s cluster memories. *Id.* at 40–46. This section does not state a reason for replacing Thelen’s single memory with Bailey’s multiple private memories. Petitioner has not provided a reason to replace Thelen’s single memory 340 with Bailey’s private memories 615a-c.

This is significant, because, as discussed in more detail below, Thelen’s hierarchy of recognition models shown in Figure 4, which is stored

in Thelen’s single memory shown in Figure 3, avoids duplicating data by sharing lexical data common to each of the recognition models. In contrast, Petitioner’s combination of Thelen and Bailey, by storing each recognition model in a separate memory, duplicates lexical data common to each of the recognition models. Petitioner does not identify any advantage in replacing Thelen’s single memory with Bailey’s multiple private memories that would have outweighed this disadvantage. *See* Pet. 40–45.

The Federal Circuit has held that “[O]bviousness concerns whether a skilled artisan not only *could have made* but *would have been motivated to make* the combinations or modifications of prior art to arrive at the claimed invention.” *Belden Inc. v. Berk-Tek LLC*, 805 F.3d 1064, 1073 (Fed. Cir. 2015). The mere fact that a skilled artisan could have replaced Thelen’s single memory with Bailey’s private memories does not provide a reason that a skilled artisan would have modified Thelen to replace its single memory that efficiently stores recognition models in a hierarchical data structure with Bailey’s less efficient private memories which must duplicate data common to the recognition models. *Virtek Vision International ULC v. Assembly Guidance Systems, Inc.*, 97 F.4th 882, 887 (Fed. Cir. 2024) (“The mere fact that these possible arrangements existed in the prior art does not provide a reason that a skilled artisan would have substituted [a component of one reference] with [a component] disclosed in [another reference].”). Here, this case involves nothing other than an assertion that because private memories were disclosed in a prior art reference and were therefore “known,” that satisfies the motivation to replace Thelen’s more efficient single memory with Bailey’s less efficient multiple private memories.

Therefore, I would find all challenged claims patentable over Petitioner's combination for this reason alone.

Turning now to the motivation to modify the combination of Thelen and Bailey to include Chen, Petitioner contends that a person of ordinary skill in the art "would have been motivated to . . . modify Thelen to include grouping recognizers together for the purpose of associating the generic and specific contexts (e.g., all sports contexts) together in a group for efficient word recognition." Pet. 40.

Figure 4 of Chen is reproduced below.

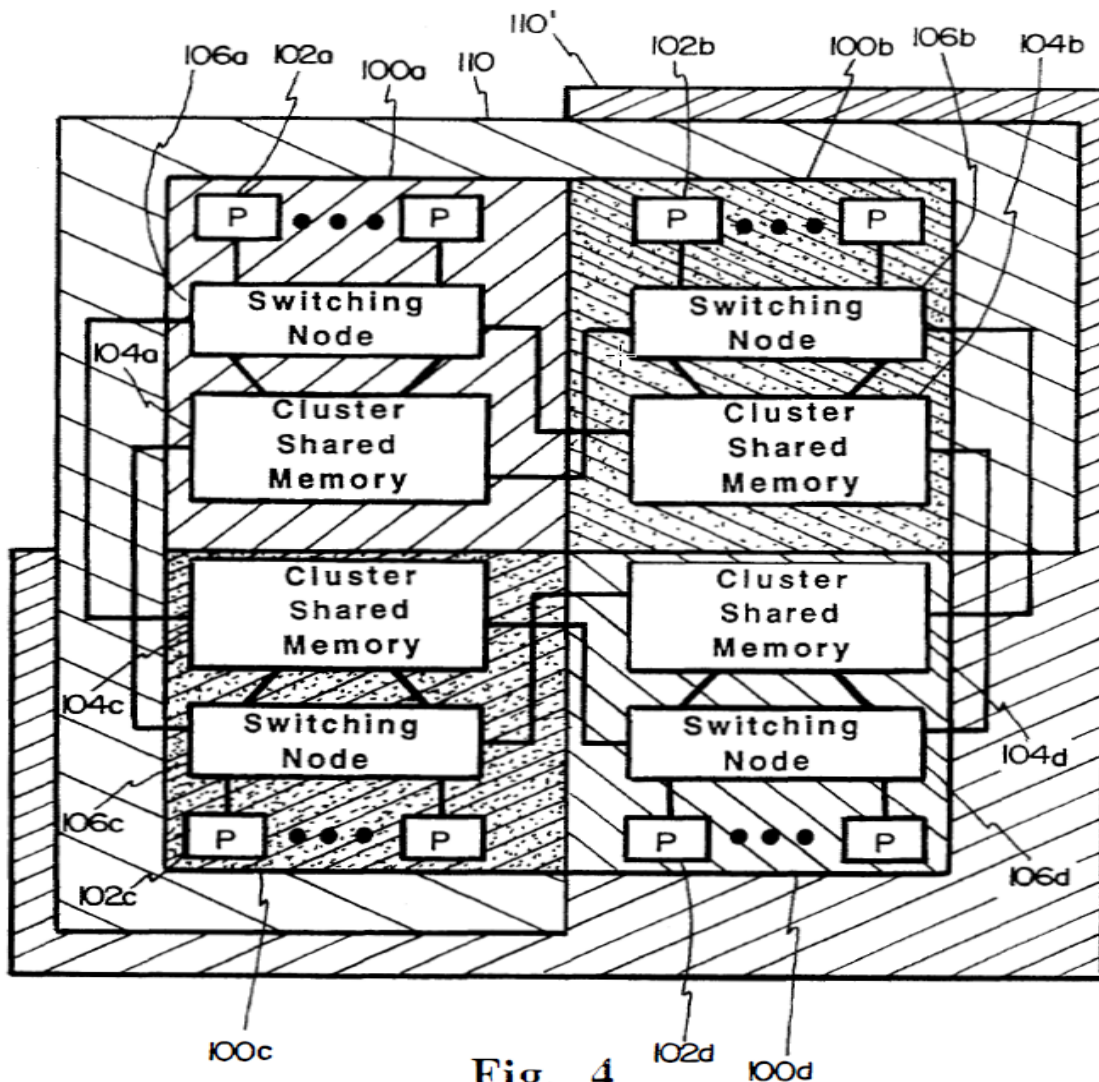


Fig. 4

Figure 4 above shows a parallel processing system that includes multiple clusters of processors 100a-100d. Ex. 1005, 9:10–13. Each cluster, such as cluster 100a, includes at least two processors 102a connected to cluster shared memory 104a. *Id.* at 9:14–17.

According to Petitioner’s combination of Thelen, Bailey, and Chen, all sports contexts, such as all of Thelen’s sports models, would be stored in one of Chen’s cluster memories, such as cluster memory 104a of cluster 100a. *See* Pet. 41, 43–45. In Petitioner’s combination, each of processors 102a of cluster 100a would recognize speech using an associated one of the sports models stored in memory 104a. *See id.* Petitioner contends that replacing the processors and private memories in the combination of Thelen and Bailey with Chen’s clusters of processors, where the processors of each cluster are associated with a cluster memory that stores all generic and specific models for one topic, such as sports, “would have more efficiently allowed for sharing of lexical data common to each of the recognition models.” *Id.* at 44; *see id.* at 45.

Although the combination of Thelen, Bailey, and Chen may have been more efficient than Petitioner’s inefficient combination of Thelen and Bailey, Petitioner does not explain why the combination of Thelen, Bailey, and Chen would have been more efficient than Thelen alone. I agree with Patent Owner and Dr. Anderson that Thelen discloses a single shared memory structure, because all of Thelen’s processors obtain lexical data from Thelen’s single common memory 340. PO Resp. 26–27 (citing Ex. 2020 ¶¶ 53–54; Ex. 1003 ¶ 77; Ex. 1030, 7:30–34, 5:22–24, Fig. 3). Dr. Anderson persuasively testifies that “Thelen’s existing architecture relies on a shared memory structure in which all components ultimately obtain their lexical

data from a single common memory.” Ex. 2020 ¶ 56 (citing Ex. 1003 ¶ 77 (Mr. Schmandt testifying that Thelen’s storage 340 stores the “complete lexical data.”); Ex. 1030 7:30–34, 5:22–24, Fig. 3). As discussed above, Petitioner concedes that Thelen already discloses “a memory containing the complete lexical data” and “does not expressly teach that each of its recognition models and associated lexicons . . . are stored in a plurality of lexical memories.”¹⁵ Pet. 30–31.

¹⁵ The majority, in its footnote 10, contends that Thelen’s common storage 340 is not a shared memory because each recognizer has access to storage 340 indirectly via model selector 360. However, the presence of model selector 360 does not change the fact that memory 340 contains the complete lexical data as conceded by Petitioner (Pet. 30), and that every recognition processor of Thelen obtains its lexical data from memory 340 as shown in Figure 3. *See* Ex. 1030, 7:30–49. Further, the majority’s statement that Thelen’s storage 340 is not a shared memory is inconsistent with Chen’s disclosure that “[i]n a shared memory model, data is stored in a common shared memory that is equally accessible to all processors in the parallel processing system.” Ex. 1005, 1:59–62. The majority’s statement is also inconsistent with the testimonies of both experts and the contentions of both parties. *See* Ex. 1003 ¶ 77; Ex. 2020 ¶ 56; Pet. 30–31; PO Resp. 26–27.

Further, in Thelen’s embodiment where the number of models equals the number of recognizers, model selector 360 would not be used, because each recognizer is associated with a model stored in memory 340. Ex. 1030, 7:43–47.

Figure 4 of Thelen is reproduced below.

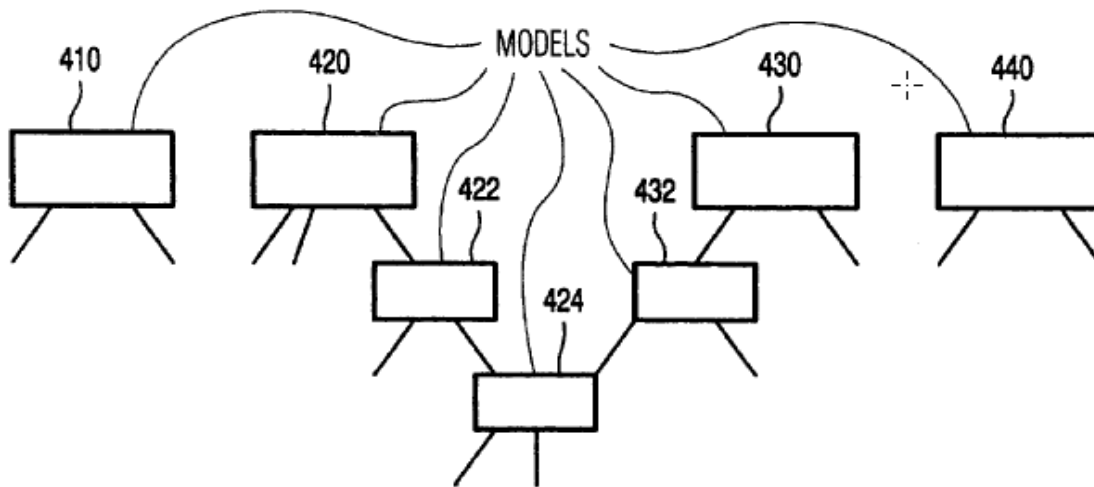


FIG. 4

Figure 4 above shows a hierarchy of recognition models. Ex. 1030, 4:37.

As shown in Figure 4 above, Thelen organizes all recognition models into a single hierarchical data structure and stores the single data structure in Thelen's single memory 340 shared by Thelen's recognizers. Ex. 1030, Figs. 3 and 4, 7:30–35, 8:54–64. This data structure shares lexical data common to each of the recognition models. *Id.* For example, health farms submodel 424 is hierarchically below parent model 422 that relates to hospitals/medical centers, and also below parent model 432 that relates to travel. Ex. 1030, Fig. 4, 8:65–9:8.

Because Thelen uses a single hierarchy of recognition models, submodel 424 does not need to be duplicated in its memory 340 in order to descend from both parent models 422, 432. In contrast, Petitioner's combination needs to duplicate model 424 in Chen's cluster memory that stores all travel models, and also needs to duplicate model 424 in Chen's cluster memory that stores all health models, because Petitioner's combination divides Thelen's single hierarchy of recognition models among

Chen's multiple cluster memories 104a-104d. Thus, as shown in Thelen's Figure 4, "Thelen already teaches recognition models associated with a particular context and ranging from generic to specific contexts for a particular topic, such as sports," and "grouping the recognition models according to the hierarchical association" as contended by Petitioner (Pet. 42), and does so by duplicating less data than would have been duplicated by storing Thelen's lexical models among the multiple cluster memories of Chen.

As a result, Thelen already provides the benefit of "consolidating speech recognition for a particular context," such as sports, "within a particular group of recognition processors" so that "[t]he group of processors would then share the same lexical memory" which "more efficiently allow[s] for sharing of lexical data common to each of the recognition models." *See* Pet. 44. For example, in Thelen's embodiment where the number of recognition models corresponds to the number of recognition processors (Ex. 1030, 7:43–47), each generic and specific sports model stored in Thelen's shared lexical memory would always be assigned to a corresponding processor. This embodiment of Thelen would achieve the same result as Petitioner's combination of Thelen, Bailey, and Chen, by causing Thelen's processors to perform speech recognition by accessing all of the generic and specific models for the sports contexts, as well as all of the generic and specific models for all other contexts, from Thelen's single hierarchy of recognition models stored in Thelen's single common memory 340. This embodiment of Thelen would have the added advantages of duplicating less data as well as using a single "shared memory model having a globally shared physical address space," which is, "[f]rom the

programmer's perspective, the most attractive parallel processing system.”
Ex. 1005, 1:66–2:1.

As another example, in Thelen's embodiment where the number of processors is less than the number of models, Thelen teaches that if one processor is obtaining good results using a generic model, such as a sports model, then a model selector causes another processor using another model giving worse results, such as health, to stop using the health model and instead use more a specific model within the sports context, so that the system performs recognition using the more specific sports model in addition to the generic sports model. Ex. 1030, 9:6–24. If several more specific sports models exist, such as football, baseball, athletics, and car racing, then the selector can cause all of these specific sports models to also be tested by Thelen's processors. *Id.* at 9:24–26; *see id.* at 3:4–9, 3:27–30, 3:47–52, 7:63–8:1, 8:36–53. In this embodiment, Thelen's model selector consolidates speech recognition to the particular context best suited for recognition, such as sports, within Thelen's group of recognition processors.

Thus, in any embodiment, Thelen already provides the same, if not better, advantages of sharing data than those provided by Petitioner's combination. The alleged advantages of adding Chen to the combination of Thelen and Bailey already exist in Thelen alone, and Petitioner has not shown otherwise.

Rather, Petitioner's combination would have performed less efficiently than Thelen alone. Petitioner admits that its combination would have required duplicating a core vocabulary of lexical data in each cluster memory of Chen. Pet. Reply 22. Thelen's single memory, shared by all recognizers, would not need to duplicate the core vocabulary among

multiple memories as required by Chen. Thus, storing Thelen's single hierarchical data structure of lexical models in Thelen's single shared memory results in sharing common lexical data more efficiently than diving and storing the hierarchical data structure among Chen's multiple cluster memories as required by Petitioner's combination. *Id.*; see PO Sur-Reply 26–27. Petitioner has not provided persuasive evidence or argument to the contrary.

In fact, Petitioner cites Thelen to support the contention that a single memory storing a single hierarchy of recognition models would store overlapping data in fewer locations. See Pet. 27–29 (citing, *inter alia*, Ex. 1030, 7:40–42, 2:3–8), 42–44 (citing, *inter alia*, Ex. 1030, Fig. 4, 8:54–56, 8:62–9:11). For example, according to Petitioner, “Thelen already teaches recognition models associated with a particular context and ranging from generic to specific contexts for a particular topic, such as sports.” Pet. 42 (citing Ex. 1030, Fig. 4). Thus, Petitioner relies on Thelen to teach the benefit of storing generic and specific models for a topic in a single memory.

The majority's statement that Thelen discloses that “there is no shared memory in which multiple processors can access the same consolidated model data in parallel” (FWD 27) is inconsistent with Thelen's description of consolidating data by organizing recognition models into the hierarchy of recognition models as shown in Figure 4, and storing the hierarchy of recognition models in memory 340 as discussed above. The majority's statement is also inconsistent with Thelen's description that each recognizer accesses data for a particular model from memory 340, such that Thelen's group of processors can process all generic and specific models for a given topic such as sports in parallel according to any embodiment of Thelen as

discussed above. *See* Ex. 1030, 9:17–30. In addition, the majority’s statement that Thelen does not consolidate the data of the recognition models is inconsistent with the contentions of both parties and the testimony of both experts. *See* Pet. 27–29, 42–43 (citing Ex. 1030, Fig. 4); PO Resp. 26–27; Ex. 1003 ¶ 83 (citing Ex. 1030, Fig. 4); Ex. 2020 ¶ 56. Further, the disclosures of Thelen and Chen, the testimonies of both experts, and the contentions of both parties show that Thelen’s memory 340 is a shared memory as discussed above.

The majority’s statement that Thelen does not disclose a shared memory (FWD 27) is based on the premise that “each processor [of Thelen] has access to only one distinct lexical model at a time.” However, the majority does not explain why the processors in Petitioner’s combination would have processed the models any differently than Thelen. In particular, neither Petitioner nor Mr. Schmandt contends that a given processor in Petitioner’s combination would have had access to more than one model at a time. Rather, according to the Petition, the “collective groups of recognizers” of Chen (Pet. 43), “where each recognizer is associated with a recognition model (per Thelen) . . . would have yielded the predictable result of consolidating speech recognition for a particular context within a particular group of processors” (Pet. 44), which Thelen alone already describes. Contrary to the majority’s statement, Petitioner’s combination relies on Thelen’s disclosure of associating each recognizer with a model.

The alleged benefit of using Chen’s cluster memories in place of Bailey’s private memories, which is “consolidating speech recognition for a particular context within a particular group of recognition processors” to “more efficiently allow[] for sharing of lexical data common to each of the

recognition models” (Pet. 44), already exists in Thelen alone, because Thelen organizes all recognition models into a single hierarchical data structure stored in Thelen’s single shared memory 340 in order to store overlapping lexical data in fewer unique locations. Ex. 1030, Figs. 3 and 4, 7:30–35, 8:54–64; PO Sur-Reply 26–27; Pet. Reply 21–22; *see* Pet. 27–29, 42–43. A person of ordinary skill in the art would have recognized that Thelen alone already has a single memory storing a single hierarchy of recognition models, providing the benefit of efficiently sharing lexical data common to each of the recognition models. Petitioner has not shown a reason to replace Thelen’s single memory that stores all lexical models in a single hierarchical data structure with Chen’s multiple cluster memories, where each cluster memory stores only a portion of Thelen’s lexical models, thus requiring duplication of more data than Thelen alone. *See id.* Contrary to Petitioner’s contentions, the evidence of record shows that a person of ordinary skill in the art would have stored Thelen’s hierarchy of recognition models in Thelen’s single shared memory rather than in Chen’s multiple cluster memories to achieve the benefit of duplicating less data than Petitioner’s combination, as well as the benefit of “us[ing] a shared memory model having a globally shared physical address space,” which is, “[f]rom the programmer’s perspective, the most attractive parallel processing system.” Ex. 1005, 1:66–2:1.

In conclusion, Petitioner has not shown any reason to modify Thelen to include the teachings of Bailey. Petitioner also has not shown that the efficiency of sharing data, which is found in Thelen alone, would have been a reason to modify Thelen to include the teachings of Bailey and Chen. Petitioner has the burden of showing that the combination of Thelen, Bailey,

and Chen would have shared data more efficiently than Thelen alone, and Petitioner has not met this burden. *Dynamic Drinkware, LLC v. Nat'l Graphics, Inc.*, 800 F.3d 1375, 1378–79 (Fed. Cir. 2015).

(2) Flexibility and Scalability

To support the contention that scalability and flexibility provide motivation, Petitioner contends that “the Petition and supporting declaration provide ample support that a [person of ordinary skill in the art] would have been motivated by the flexibility and scalability provided by Hon to utilize Chen’s processing clusters in Thelen’s speech recognition system, regardless of respective speeds and/or costs of the systems.” Pet. Reply 23 (citing Pet. 45; Ex. 1003 ¶¶ 58, 84; Ex. 1006, 19–20).

In paragraph 58 of his Declaration, Mr. Schmandt testifies that Hon’s clustered architecture is scalable because it can be expanded by adding more processors. In paragraph 84, Mr. Schmandt testifies that “[o]ther motivation for the modification would have come from [Hon] . . . praising parallel processing configurations for their flexibility and scalability in speech recognition systems.” (citing Ex. 1006, 19–20). Dr. Anderson, relying on the deposition testimony of Mr. Schmandt, testifies that Hon’s system was flexible because it had a programmable pipeline. Ex. 2020 ¶ 61 (citing Ex. 2017, 105:8–106:4). Dr. Anderson further testifies that Hon’s discussion of flexibility is specific to Hon’s system compared to another specific system, rather than to parallel processing generally. *Id.*

Dr. Anderson testifies that Mr. Schmandt “has made no showing that Thelen itself, without modification, was not already both flexible and scalable.” Ex. 2020 ¶ 61; *see* PO Resp. 30–31. I agree with Dr. Anderson. Mr. Schmandt does not testify, in either paragraph 58 or 84, that Thelen

without modification is not already flexible and scalable. Petitioner does not provide persuasive evidence or argument to rebut Dr. Anderson's testimony. I agree with Dr. Anderson, that Mr. Schmandt has not shown that Thelen's parallel processing speech recognition system, without modification, was not already flexible and scalable.

Further, I would have found that Thelen alone is already scalable, because Thelen discloses that the speech recognition system comprises M recognition models and N active recognizers. Ex. 1030, 7:63–64. Thus, the number of Thelen's recognizers, N, can be increased or decreased, which provides scalability. *Id.*; *see id.* at 7:43–49; Ex. 2020 ¶ 61.

In addition, I would have found that Thelen is already flexible, because when Thelen assigns a processor to a model, such as health, that gives low recognition results, then Thelen may terminate using this model in favor of a model that gives better results, such as sports in general or football specifically. Ex. 1030, 9:17–24; *see id.* at 9:13–31. If several more specific sports models exist, such as football, baseball, athletics, and car racing, then all of these models may be assigned to corresponding processors. Ex. 1030, 9:24–26.

In contrast, in the combination relied on by Petitioner, one group of processors would always be associated with the “sports” context, and another group of processors would always be associated with the “health” context. *See* Pet. 42–43. Thus, when the speech to be recognized is about sports, the processors in the health group would continue to perform recognition using the health models even though the health models give low results and the sports models give better results. *See* Tr. 41:1–43:10.

To the extent Petitioner relies on Thelen’s embodiment where the number of recognition models corresponds to the number of recognition processors, Petitioner has not shown any difference in flexibility between Thelen and Petitioner’s combination, “where each processor is associated with a recognition model (per Thelen).” Pet. 44. I agree with Dr. Anderson, that Mr. Schmandt has not shown that Thelen, without modification, was not already flexible and scalable. Ex. 2020 ¶ 61.

Therefore, I would have found that scalability and flexibility, which is already present in Thelen alone, would not have motivated a person of ordinary skill in the art to modify Thelen. Petitioner has the burden of showing that the combination of Thelen, Bailey, and Chen would have been more flexible and scalable than Thelen alone, and Petitioner has not met this burden. *Dynamic Drinkware, LLC*, 800 F.3d at 1378–79).

(3) Decreased Cost

To support the contention that decreased cost from using off-the-shelf components provides motivation, Petitioner contends that a person of ordinary skill in the art would have utilized Chen’s clusters of commercially-available processors to reduce the cost of computing resources. Pet. Reply 23 (citing Ex. 1003 ¶ 83). Mr. Schmandt testifies that each of Chen’s processors may be formed using commercially available processors. Ex. 1003 ¶ 82 (citing Ex. 1005, 9:5–19). Mr. Schmandt does not testify, however, that Thelen requires processors that would have been more expensive than those of Chen. In addition, Mr. Schmandt does not testify that Thelen’s system would not use commercially available

processors. In fact, Thelen discusses using convention components throughout its disclosure. *See* Ex. 1030, 6:55–56, 7:23–24, 9:36–38.

I agree with Patent Owner that Mr. Schmandt has not undertaken any cost-benefit analysis in reaching his opinions regarding a motivation to combine. PO Sur-Reply 23 (citing Ex. 2017, 42:14–19, 63:8–15). I agree with Patent Owner that Petitioner has not stated what the combination relied on by Petitioner would be cheaper than, or what costs would have been reduced. *Id.* at 25. In particular, Petitioner does not state that the combination relied on by Petitioner, in using conventional components, would have been cheaper than Thelen’s speech recognition system, or that Thelen’s components would have been more expensive than Chen’s.

Given that (a) Petitioner does not state what the combination relied on by Petitioner would have been cheaper than, (b) Mr. Schmandt does not testify that the components of Thelen would have been more expensive than Chen’s conventional components, and (c) Thelen already teaches using conventional components, I would have found that Petitioner has not provided sufficient evidence or argument showing that Chen’s clusters of processors would have been cheaper than Thelen’s speech recognition system. Petitioner has the burden of showing that the combination of Thelen, Bailey, and Chen would have been cheaper than Thelen alone, and Petitioner has not met this burden. *Dynamic Drinkware, LLC*, 800 F.3d at 1378–79).

(4) Improved Speed and Power

I also agree with Patent Owner that Mr. Schmandt has not persuasively shown that the clustered processor and memory arrangement of the combination of Thelen, Bailey, and Chen would have necessarily

improved the speed and power of Thelen's speech recognition system. Patent Owner persuasively contends that Mr. Schmandt testified at his deposition that (a) the combination of Thelen and Bailey with Chen would not increase the number of processors in Thelen, (b) he did not know whether the combination relied on by Petitioner would have increased the number of processors in Thelen, and (c) his Declaration does not explain how the combination relied on by Petitioner would have increased processing power. PO Resp. 23–24 (quoting Ex. 2017, 130:23–131:4), 25 (citing Ex. 2017, 136:3–6). In addition, Petitioner concedes that the “combination's motivation to combine is not contingent on increasing processing power via adding more processors beyond what is taught in Thelen.” Pet. Reply 24. I would find that Petitioner has not explained how Petitioner's combination would have increased the speed and power of Thelen's speech recognition system.

Further, when Petitioner's combination does not “add[] more processors beyond than what Thelen alone teaches” (Pet. Reply 24), Thelen's system would have at least as much, if not more, speed and power than Petitioner's combination. For example, when Chen has four clusters and each cluster has two processors for a total number of eight processors, then, for a given topic such as sports, the two processors in the sports cluster would perform speech recognition using the sports model and submodels stored in the associated cluster memory, while the six processors in the other clusters would be using models other than the sports model and thus would provide low, if any, results. *See* Tr. 41:1–43:10.

In contrast, in Thelen's embodiment where the number of processors is less than the number of models, Thelen would assign many, if not all, of

its eight processors to the sports model and submodels, thus providing greater processing speed and power to the speech recognition than the combination relied on by Petitioner. *See* Ex. 1030, 7:62–8:5, 9:13–32. To the extent Petitioner relies on Thelen’s embodiment where the number of recognition models corresponds to the number of recognition processors, I agree with Dr. Anderson, that Mr. Schmandt has not shown that the clustered processor and memory arrangement of Chen would have improved the speed and power of Thelen without modification, given that both Thelen and Petitioner’s combination would process the same number of models using the same number of processors. *See* Ex. 2020 ¶ 51; *see id.* at ¶¶ 52–58, 62; Pet. Reply 24 (Petitioner contending that Petitioner’s combination would not add more processors “beyond what is taught in Thelen.”).

Therefore, I agree with Patent Owner, that Petitioner and Mr. Schmandt have not persuasively shown that modifying Thelen to use the clustered architecture of Chen would have increased processing speed and power over Thelen. To the contrary, the evidence shows that Thelen would have had more processing speed and power than the combination relied on by Petitioner. Petitioner has the burden of showing that the combination of Thelen, Bailey, and Chen would have performed with more power and speed than Thelen alone, and Petitioner has not met this burden. *Dynamic Drinkware, LLC*, 800 F.3d at 1378–79.

(5) Not the Most Desirable Combination

Petitioner contends that the combination relied on by Petitioner does not have to be the most optimal, desirable, or preferred. Pet. Reply 22–23 (citing *Elekta Ltd. v. ZAP Surgical Systems, Inc.*, 81 F.4th 1368, 1374 (Fed. Cir. 2023)). However, Petitioner has not shown any advantage in modifying

Thelen. The mere fact that Petitioner's possible arrangements existed in the prior art does not provide a reason that a skilled artisan would have replaced Thelen's single memory with Bailey's multiple private memories, then replaced Bailey's multiple private memories with Chen's multiple cluster memories, in order to produce a system that requires more memory than Thelen alone, has less flexibility than Thelen alone, and has less processing power and speed than Thelen alone. *See Virtek Vision*, 97 F.4th at 887; Ex. 2020 ¶ 62. The Petition does not show, by a preponderance of the evidence, that a person of ordinary skill in the art would have modified Thelen as contended in the Petition.

(4) Conclusion Regarding a Reason to Combine

I agree with Patent Owner that Petitioner has not shown that Petitioner's combination of Thelen, Bailey, and Chen provides greater efficiency of sharing data among related sub-context models, because Thelen already has this efficiency, and because Petitioner's combination results in sharing less data than Thelen alone. I agree with Patent Owner that Petitioner has not shown that Petitioner's combination provides the advantage of flexibility and scalability of distributed processing systems over Thelen alone, because Thelen is already scalable and flexible, and because Petitioner has not shown how its combination would have been flexible. I agree with Patent Owner that Petitioner has not shown that Petitioner's combination provides the advantage of decreased cost by using off-the-shelf components, because Thelen already discloses using conventional components, and because Petitioner has not shown that Thelen's components would have been more expensive than Chen's components. I agree with Patent Owner, that Petitioner has not shown that

Petitioner's combination would not have had more processing speed and power than Thelen.

Petitioner has the burden of establishing by a preponderance of the evidence that a person of ordinary skill would have had a reason to modify Thelen. *See Dynamic Drinkware, LLC*, 800 F.3d at 1378–79. I would have found that Petitioner has not met this burden. I agree with Dr. Anderson, that “the evidence demonstrates that Thelen would perform *worse*, not better, after the proposed combination, and that Thelen has no need or deficiency that the [person of ordinary skill in the art] would look to address through a combination or modification.” Ex. 2020 ¶ 62. Therefore, because Petitioner has not shown an advantage in modifying Thelen, I would conclude that Petitioner has not shown any of the challenged claims unpatentable.

(7) Disadvantages of the Combination relied on by Petitioner

I also agree with Patent Owner, that a person of ordinary skill in the art would have recognized multiple disadvantages in using Chen's clusters to perform Thelen's speech recognition process. For example, I agree with Patent Owner and Dr. Anderson, that Thelen discloses a single shared memory structure. PO Resp. 26–27 (citing Ex. 2020 ¶¶ 53–54; Ex. 1003 ¶ 77; Ex. 1030, 7:30–34, 5:22–24, Fig. 3); *see* Ex. 2020 ¶ 56 (“Thelen's existing architecture relies on a shared memory structure in which all components ultimately obtain their lexical data from a single common memory.”). I agree with Patent Owner, that replacing Thelen's single shared memory with Chen's multiple private memories would have required a significant reprogramming effort. PO Resp. 26–28.

For example, Chen discloses that a parallel processing system that uses a shared memory model is the most attractive from a programmer's perspective because most programs are written for a shared memory model, as opposed to a distributed or private memory model. Ex. 1005, 1:66–2:6. Chen discloses that effectively porting an existing software program from a shared memory model to a distributed memory model can require significant reprogramming effort. *Id.* at 2:6–9. Chen discloses that a distributed memory model requires processes and data to be partitioned into separate tasks that are small enough to fit into the private local memory of a processor and large enough to recover the overhead required to establish a parallel task in a message passing environment. *Id.* at 2:9–17.

I agree with Dr. Anderson, that a person of ordinary skill in the art “would have been aware of Chen’s warning that porting Thelen’s speech recognition technique to Chen’s distributed memory architecture would require ‘significant reprogramming effort,’ including the complex task of finding the optimal approach to partitioning the processes and data of Thelen’s speech recognition techniques to balance the competing problems of memory size and overhead.” Ex. 2020 ¶ 57. I also agree with Dr. Anderson, that Mr. Schmandt’s Declaration does not address these issues, which are explicitly identified by Chen itself. *Id.*

Although Petitioner in Reply contends that using Chen’s architecture to perform Thelen’s speech recognition techniques “does not require any significant reprogramming effort” (Pet. Reply 22), Petitioner’s contention is attorney argument unsupported by evidence. Petitioner’s contention that “the clustered processor architecture [of Chen] provides increased benefits over a single-processor architecture” (*id.* at 20) is irrelevant, because Thelen

(a) already uses a multiple-processor architecture and (b) has the additional advantage of using a single “shared memory model” which is “the most attractive parallel processing system” as taught by Chen. Ex. 1005, 1:66–68; *see id.* at 2:24–29 (“Thus, it would be desirable to have a parallel processing system” in which “processors all have equal and symmetric access to any and all memory locations in a common shared memory.”). Petitioner has not explained why a person of ordinary skill would have replaced Thelen’s single shared memory, which is “the most attractive parallel processing system” with Chen’s less attractive parallel processing system that uses multiple private cluster memories distributed over at least four clusters.

I agree with Patent Owner and Dr. Anderson, that, contrary to Mr. Schmandt’s contention, contemporary real-world evidence provided by Mathew I (Ex. 2021) shows that a person of ordinary skill in the art could not have reasonably expected the combination of Thelen, Bailey, and Chen to speed up Thelen’s speech recognition process, and that Mr. Schmandt has not explained why Chen would speed up Thelen. PO Resp. 25; Ex. 2020 ¶¶ 50–52. Mathew I discloses that a two-processor parallel processing design was faster than the original sequential version, but a five-processor design was slower than the two-processor version due to the high synchronization overhead. Ex. 2021, 11.

Additional evidence supports Patent Owner’s contention. For example, Hon discloses that “[t]he long communication latency also make[s] Hon’s parallel processing system] unsuitable for algorithms that require frequent synchronization, like . . . natural language processing.” Ex. 1006, 25. Chen discloses that in an extended shared memory model, to account for latency of inter-cluster memory access, the number of inter-cluster memory

accesses should be less than about 10–12% of the number of intra-cluster memory accesses, and preferably should be about 1–5% of the total number of memory accesses for any given parallel processing task. Ex. 1005, 8:63–65, 8:68–9:4; *see id.* at 8:42–9:4, Figs. 3A, 3B. Petitioner has not shown that inter-cluster memory accesses in the combination relied on by Petitioner would have been less than 12% of intra-cluster memory accesses.

Petitioner contends that Mathew’s extremely limited discussion of its dissimilar architecture would not serve to dissuade a person of ordinary skill from the combination relied on by Petitioner. Pet. Reply 18. I disagree. The prior art recognized the difficulty in using a distributed memory parallel processing system in place of a shared memory parallel processing system. Mathew I, together with Hon, Chen, and Dr. Andersons’s testimony, persuasively shows that a person of ordinary skill in the art would have preferred Thelen’s shared memory system over the combination’s distributed memory system.

I agree with Patent Owner and Mr. Schmandt that “Thelen in unmodified form expresses no doubt or concern about achieving real time recognition.” PO Resp. 28 (citing Ex. 2017, 155:7–17). Petitioner has not shown any advantage to modifying Thelen, which already achieves real time recognition, already shares data common to each lexical model, is already scalable and flexible, and already uses conventional components. Further, Patent Owner has shown multiple disadvantages to modifying Thelen. I agree with Patent Owner and Dr. Anderson, that a person of ordinary skill in the art would have recognized that using Chen’s clusters to perform Thelen’s speech recognition techniques would have yielded several disadvantages and would not have yielded any advantage. Ex. 2020 ¶ 62; *see id.* at ¶¶ 51–58.

IPR2023-00033
Patent 7,587,319 B2

Therefore, I would have concluded that all challenged claims are patentable over Petitioner's combination.

IPR2023-00033
Patent 7,587,319 B2

For PETITIONER:

Jennifer Bailey
Adam Seitz
Adam Sandwell
ERISE IP, P.A.
jennifer.bailey@eriseip.com
adam.seitz@eriseip.com
adam.sandwell@eriseip.com

J. Hadden
Dargaye Churnet
Saina Shamilov
FENWICK & WEST LLP
dhadden-ptab@fenwick.com
dchurnet@fenwick.com
sshamilov-ptab@fenwick.com

For PATENT OWNER:

Peter Knops
Kayvan Noroozi
NOROOZI PC
peter@noroozipc.com
kayvan@noroozipc.com

Katherine Rhoades
BARTLIT BECK LLP
katherine.rhoades@bartlitbeck.com