UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC.,
Petitioner,

v.

Zentian Limited
Patent Owner.

Case IPR2023-00033
Patent No. 7,587,319

PATENT OWNER'S RESPONSE

TABLE OF CONTENTS

I. Introduction 1

II. Background of the '319 Patent 2

III. Person of ordinary skill in the art 5

IV. The Petition fails to demonstrate that an ordinary artisan would have had a reasonable likelihood of success in combining Thelen, Bailey, and Chen 6

V. The Petition fails to show a motivation to combine Thelen, Bailey, and Chen in the manner proposed, and the POSA would not have been motivated to do so 22

VII. Conclusion 32

EXHIBIT LIST

Exhibit No.	Description
2017	Deposition of Christopher Schmandt dated September 6, 2023
2018	*Intentionally left blank*
2019	*Intentionally left blank*
2020	Declaration of Christpher Anderson, Ph.D
2021	Binu K. Mathew et al., A Gaussian Probability Accelerator for SPHINX 3, ("Mathew I")

I. Introduction

The Petition proposes that an ordinary speech recognition artisan would have modified Thelen, Bailey, and Chen in multiple respects to arrive precisely at the highly specific and novel invention of the challenged claims.

But an ordinary artisan in the field of speech recognition would not have had a reasonable expectation of success with respect to the combination of Chen with Thelen and Bailey. None of the references enables the Petition's proposed combination, and as Dr. Anderson explains, the obstacles implicated by the combination would have been well beyond the ordinary speech recognition artisan's capability. Indeed, Apple's expert Mr. Schmandt admitted he has never supervised anyone involved in mapping a speech recognition system to a clustered processor architecture like Chen's, nor has Mr. Schmandt ***ever built or designed*** the processor to memory architectures of any of the speech recognition systems disclosed in the background of his declaration. Mr. Schmandt's own (in)experience thus belies his conclusory assurances that the combination would be merely a "simple substitution." Indeed, Mr. Schmandt admitted he did not undertake a particularized analysis of what the modification would require the POSA to do before he reached the conclusion that the POSA could have done it. By contrast, Dr. Anderson explains that even a team of engineers working under his supervision

for *six months*, with extensive assistance from a digital signal processing application expert from Texas Instruments, could not successfully implement a far simpler application on a Texas Instruments multi-processor system.

Moreover, the specific combination proposed by Mr. Schmandt would be overrun by memory contention problems, which would cause the system to fall far behind Mr. Schmandt's requirement of a real time speech recognition system. Indeed, the system of Mr. Schmandt's combination would be ***slower*** than Thelen's unmodified system, which could already perform in real time without issues. At the same time, the modification would have required significant reprogramming effort, among other obstacles. The ordinary speech recognition artisan would thus have had no motivation to pursue the Petition's proposed combination.

Accordingly, the Board should not find any challenged claim unpatentable.

II. Background of the '319 Patent

U.S. Patent 7,587,319, titled "Speech recognition using parallel processors," is directed to an improved speech recognition circuit that "uses parallel processors for processing the input speech data in parallel." Ex. 1001, 1:4-6. The '319 patent teaches multiple processors "arranged in groups or clusters," with each group or cluster of processors connected to one of several "partial lexical memories" that "contains part of the lexical data." Ex. 1001, 2:64-3:3. "Each lexical tree processor

is operative to process the speech parameters using a partial lexical memory and the controller controls each lexical tree processor to process a lexical tree corresponding to partial lexical data in a corresponding partial lexical memory." Ex. 1001, 3:3-7. The '319 patent further teaches that the invention "provides a circuit in which speech recognition processing is performed in parallel by groups of processors operating in parallel in which each group accesses a common memory of lexical data." Ex. 1001, 3:44-47.

Figure 2 of the patent, annotated below, further illustrates that architecture by showing two groups of lexical tree processors, with each group containing multiple processors 1-*k*, and each group of processors connected to a dedicated "acoustic model memory," such that there are least two acoustic model memories for at least two groups of processors. Ex. 1001, Fig. 2 (annotated).

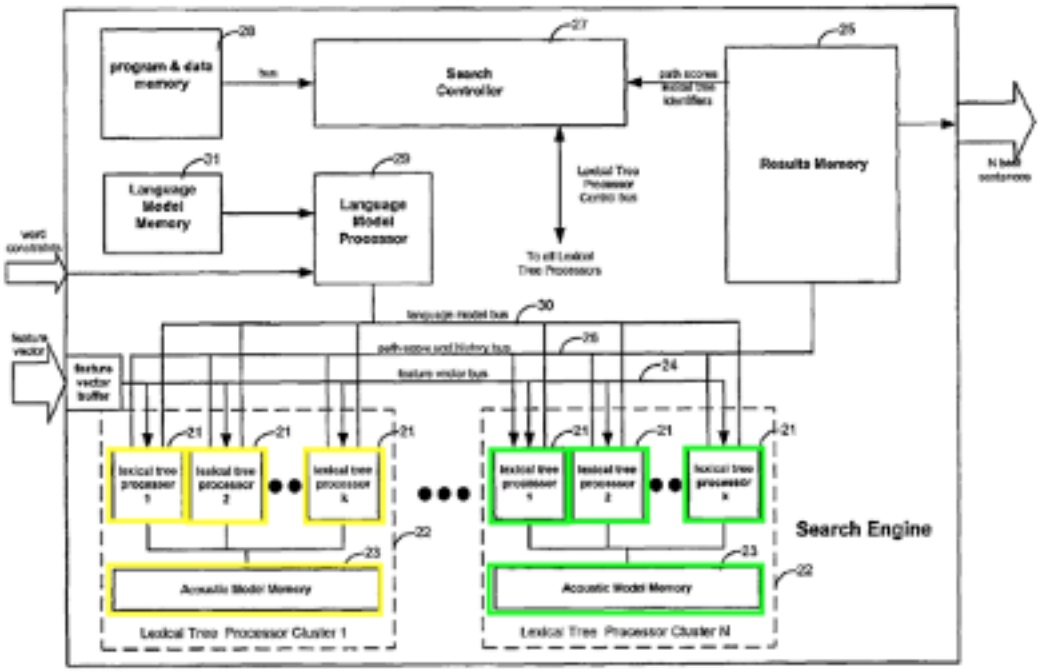


Fig 2

The '319 patent further distinguishes its architecture from two prior known alternative designs, which the patent describes as less advantageous. The '319 patent teaches that: "By providing a plurality of processors in a group with a common memory, flexibility in the processing is provided without being bandwidth limited by the interface to the memory that would occur if only a single memory were used for all processors. The arrangement is more flexible than the parallel processing arrangement in which each processor only has access to its own local memory and requires fewer memory interfaces (i.e. chip pins)." Ex. 1001, 3:50-58. The patent thus distinguishes its design from (1) a one-memory-to-all-processors design, which

it describes as bandwidth (access) limited in the processor to memory interface; and (2) a one-memory-to-one-processor design, which would require more memory interfaces and affords less flexibility in controlling the extent of parallel processing of speech parameters.

III. Person of ordinary skill in the art

This Response addresses the Petition's deficiencies by assuming the Petition's proffered definition of the person of ordinary skill in the art.

As Dr. Anderson explains, however, the field of electrical engineering encompasses many specialties, including fiber optics, analog circuits, digital circuits, very large-scale integration (VLSI), systems and controls, digital signal processing, bio-medical sensors and systems, electrical energy, electromagnetics, nanotechnology, telecommunications, computer systems and software, and others. Ex. 2020 ¶ 17.

Thus, if a student desires to study speech recognition, they would specialize in the area of digital signal processing (DSP) and then pursue studies specific to the advanced subspecialty of DSP in their graduate studies. On the other hand, a student who desires to learn computer architecture would take a different set of core courses in one of the computer engineering specialties. Ex. 2020 ¶ 17.

Mr. Schmandt's proposed person of ordinary skill in the art is a specialist in speech recognition. Any such 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. Indeed, Dr. Anderson has never taught a course that would have equipped a masters-level student in speech recognition to apply a parallel processing architecture like Chen's to known speech recognition techniques. Ex. 2020 ¶ 18. While a person with a master's degree in one of the fields identified in Mr. Schmandt's POSA definition could be a person of ordinary skill in speech recognition *or* a person of ordinary skill in high performance computing and parallel processing, that person would not be both. Ex. 2020 ¶ 19.

IV. The Petition fails to demonstrate that an ordinary artisan would have had a reasonable likelihood of success in combining Thelen, Bailey, and Chen

The Petition's theory as to all challenged claims relies on at least a combination of Thelen, Bailey, and Chen. In particular, the Petition proposes to "group" Thelen's recognizers into clusters of processors, with each cluster of processors connected to a dedicated memory, such that each dedicated cluster memory is a lexical memory containing part of the lexical data, and all of the cluster memories combined contain complete lexical data. Pet. at 40.

"To render a claim obvious, the prior art, taken as a whole, ***must enable*** a skilled artisan to make and use ***the claimed invention***." *Raytheon Techs. Corp. v. GE*, 993 F.3d 1374, 1380 (Fed. Cir. 2021) (emphasis added). "[E]ven though a non-enabling reference can play a role in an obviousness analysis, the evidence of record must still establish that a skilled artisan could have made the claimed invention." *Id.* at 1381.

Here, the Petition's combination relies on a specific clustered processor and memory embodiment from Chen to meet limitations 46(d) and 46(b).

8 In your declaration at
9 paragraph 82, you rely on Chen for the
10 requirement of groups of processors, each
11 group being connected to a lexical memory;
12 right?
13 A. Yes, that is correct.
14 Q. You don't contend that either
15 Bailey or Thelen already teach groups of
16 processors with each group being connected
17 to a lexical memory; fair?
18 A. Correct.

Ex. 2017 at 120:8-18.

Chen, however, is not a reference in the field of speech recognition, and as Mr. Schmandt admitted, Chen teaches nothing as to how to implement speech recognition techniques on its clustered parallel processing architecture, much less Thelen's speech recognition techniques. Ex. 2017 at 120:19-121:2; Ex. 2020 ¶ 30.

Thelen likewise does not teach how to run its speech recognition techniques on groups of clustered processors such as Chen's. Ex. 2017 at 136:7:16 (agreeing that Thelen does not discuss applying its techniques "in terms of an architecture of clustered processors"); Ex. 2020 ¶ 31.

Accordingly, the evidence of record beyond Thelen, Bailey, and Chen must establish that the *ordinary* artisan could have implemented Thelen's speech recognition techniques within a clustered processor environment where each cluster of processors is connected to a memory, each cluster memory is a lexical memory containing part of the lexical data, and all of the cluster memories combined contain complete lexical data.

Mr. Schmandt and the Petition, however, do not provide such evidence. As Mr. Schmandt admitted under cross-examination, his declaration does not explain how a person of *ordinary* skill in the field of speech recognition would manage the parallel operation of Thelen's different recognition models working at the same time within Chen's clusters. Ex. 2017 at 137:21-138:24. Instead, Mr. Schmandt

attempted to dismiss that fundamental question as "trivial." Ex. 2017 at 138:2. Moreover, Mr. Schmandt admitted that his declaration does not identify any challenges the POSA would face in making the proposed combination. Ex. 2017 at 141:23-142:2. Mr. Schmandt further admitted that his declaration does not identify the tasks the POSA would need to undertake to enable running Thelen's speech recognition techniques on Chen's clusters of parallel processors and distributed cluster memories. Ex. 2017 at 144:3-20. Crucially, Mr. Schmandt admitted that he did not even analyze with particularity "the tasks that would be required" for the POSA to enable the Petition's proposed combination; rather, Mr. Schmandt reached his conclusion of a reasonable expectation of success without any such particularized analysis. Ex. 2017 at 144:22-145:3.

Ultimately, the only testimony Mr. Schmandt's declaration provides regarding a reasonable expectation of success with respect to the combination of Chen with Thelen and Bailey comes in a ***single sentence***. Ex. 1003 ¶ 84. According to Mr. Schmandt, a POSA would have had a reasonable expectation of success in combining Chen with Thelen and Bailey because both references teach using multiple processors, so "using *Chen's* processor architecture in the combined system constituted a simple substitution of *Bailey's* single processors providing

each pattern recognition engine for *Chen*'s groups (clusters) of processors providing the same functionality." Ex. 1003 ¶ 84.

An expert's "*ipse dixit* declaration," however, cannot prove obviousness. *TQ Delta, LLC v. Cisco Sys.*, 942 F.3d 1352, 1362 (Fed. Cir. 2019). Mr. Schmandt's "conclusory statements and unspecific expert testimony are [ ] thus inadequate to support the Board's finding" of obviousness. *Id.* at 1363.

In addition to being devoid of evidentiary value, Mr. Schmandt's conclusory assurances as to the "simplicity" of the proposed substitution are not credible given his admission that he has ***never*** supervised anyone involved in the process of mapping a speech recognition model to a clustered processor and memory architecture like Chen's.

```
20        Q.     Have you ever supervised someone
21   who had to map a speech recognition model to
22   a clustered processor and memory
23   architecture like Chen's?
24        A.     No.
```

Ex. 2017 at 145:20-24.

Moreover, although Mr. Schmandt's declaration claims he has "built" a number of speech recognition systems, *see* Ex. 1003 ¶¶ 4, 6, Mr. Schmandt

admitted under cross-examination that he has ***never*** built the processor to memory architecture of ***any*** speech recognition system identified in his declaration.

24 To summarize, with respect to
25 paragraph 4 and the systems identified in
2 your paragraph 4, you did not build the
3 processor to memory architecture for any of
4 those systems; right?
5 A. That's correct.

Ex. 2017 at 34:24-35:5.

3 Q. In paragraph 6, you say that you
4 built a number of speech recognition
5 client/server distributed systems. Did you
6 build the processor to memory architecture
7 for any of the systems you're referring to
8 in paragraph 6?
9 A. No.

Ex. 2017 at 32:3-9, *see also* 106:10-23.

Mr. Schmandt thus lacks an adequate basis from his personal knowledge and experience to opine that the substituting "*Bailey's* single processors providing each pattern recognition engine for *Chen*'s groups (clusters) of processors providing the

same functionality" so as to run *Thelen*'s speech recognition techniques would have been a "simple substitution" for an *ordinary* artisan in the field of speech recognition. *See* Ex. 1003 ¶ 84.

By contrast, as Dr. David Anderson explains, implementing the Petition's proposed combination of Chen with Thelen and Bailey would have been highly complEx. and far outside the skill level of the ordinary artisan in the field of speech recognition as defined by the Petition. Ex. 2020 ¶¶ 17-19, 27.

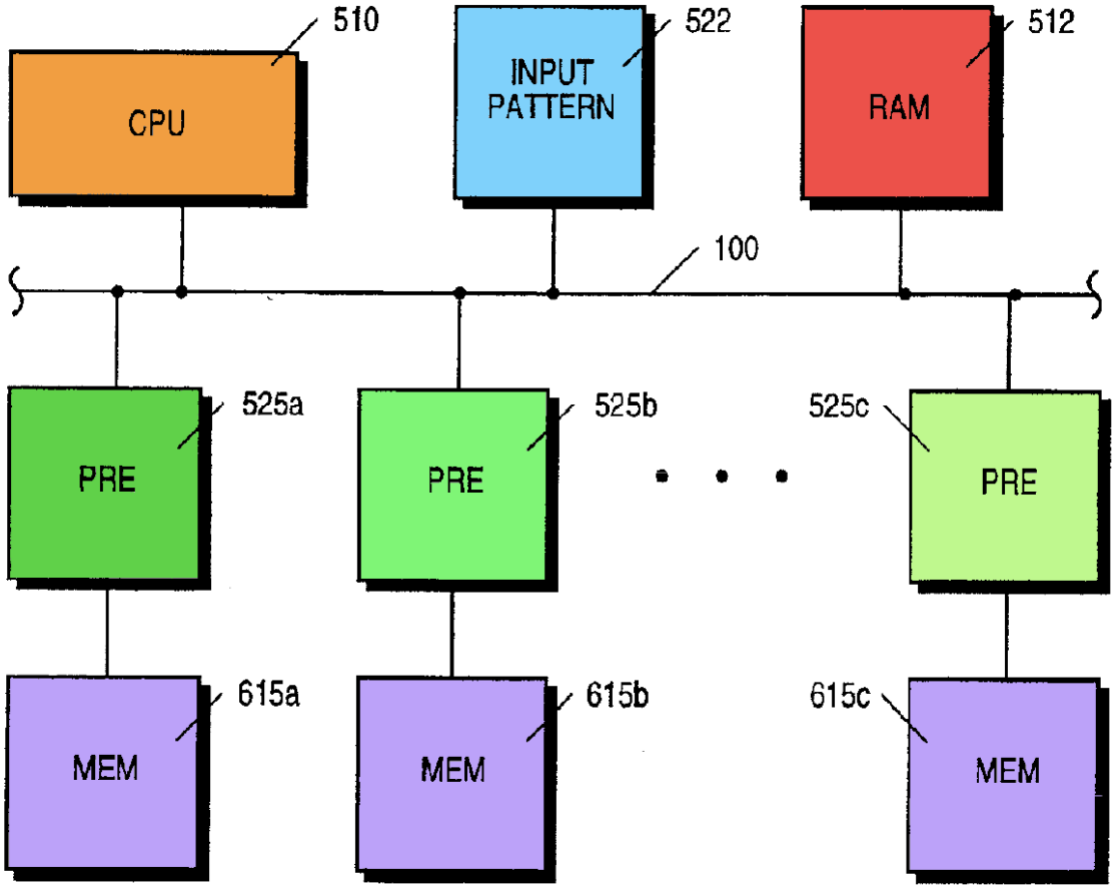
As a starting point, it is undisputed that Bailey does not teach ***groups*** of processors with each group being connected to a lexical memory. Ex. 2020 ¶ 28.

> 8 Q. With respect to claim 46(d),
> 9 which requires groups of processors with
> 10 each group connected to a lexical memory,
> 11 you don't contend that Bailey teaches that;
> 12 correct?
> 13 A. That's correct.

Ex. 2017 at 119:8-13.

Rather, Bailey teaches two processor-to-memory architectures that were expressly described as the prior art on the face of the '319 Patent. Ex. 2020 ¶ 34. In the embodiment of Bailey's Fig. 3A (below), each individual processor (labeled

Pre) is connected to its own dedicated memory (labeled mem) such that there is one processor to one memory. Ex. 2020 ¶ 35.



Ex. 1031, Fig. 3 (Petition's annotated version). As Bailey states, "each PR engine 525a-525c [is] interfaced with its own ***private memory*** 615a-615c." Ex. 1031, 11:1-3 (emphasis added); Ex. 2020 ¶ 35. The embodiment of Bailey's Fig. 3A is thus the one-memory-to-***one***-processor design distinguished by the '319 patent. Ex. 2020 ¶ 35. Accordingly, while Bailey's Fig. 3A teaches a "plurality" of memories,

it certainly does not teach the memory architecture of the '319 patent, but rather teaches the prior art design that the '319 patent innovated beyond. Ex. 2020 ¶ 35.

The design of Bailey's Figure 3B, by contrast, is identical to the other design distinguished by the '319 patent. Bailey's Figure 3B illustrates a design in which ***all*** of the processors PRE are connected to a ***single*** memory MEM 615. Ex. 2020 ¶ 36; Ex. 1031, Fig. 3B, 11:1-4.

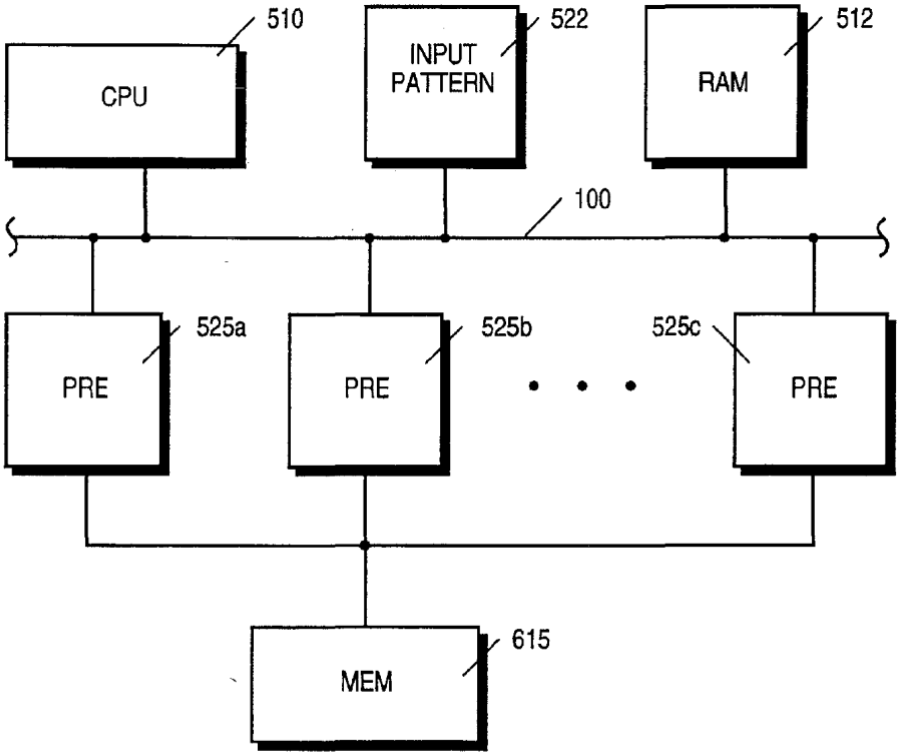


FIG. 3B

Thus, while Bailey's Fig. 3B teaches connecting a plurality of processors to ***one*** memory, it does ***not*** teach multiple *groups* of processors, each group

containing multiple processors, and each group of processors connected to a lexical memory that contains a portion of the overall lexical data, with the lexical memories combined containing all of the lexical data. Ex. 2020 ¶ 37. The '319 patent describes that with such a model, the interface to the memory would be bandwidth limited (i.e., the memory would have difficulty providing the required data to all of the processors at the required rate). Ex. 1001 at 3:50-54.

As Dr. Anderson explains, moving Thelen's speech recognition techniques to Chen's clustered, parallel processors and memories, and distributing Thelen's lexical data across Chen's distributed memories as required by limitations 46(d) and 46(b), would have been a highly complex undertaking beyond the skill of an ordinary speech recognition artisan. Ex. 2020 ¶ 38. Such a modification would have required 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. Ex.2020 ¶ 38. Moreover, whereas Chen teaches decentralized control of its processors, memories, and clusters, the Petition's proposed modification would require developing hardware and software for centralized control of the individual processors, memories, and clusters. *Id.*

Such tasks are not within the skill level of the ordinary artisan in the field of speech recognition, even under the Petition's definition. Ex. 2020 ¶ 39. As Dr. Anderson explains, the *ordinary* artisan in the field of speech recognition would have specialized in the area of digital signal processing (DSP) within electrical engineering or computer science, and would have further pursued studies specific to speech recognition, an advanced subspecialty of DSP, in their graduate studies. Ex. 2020 ¶ 39. The POSA in the field of speech recognition would not have also been a specialist in parallel processing architectures or high performance computing. Ex. 2020 ¶ 39. Rather, the ordinary speech recognition artisan would have not been capable of addressing the complex challenges of combining Chen with Thelen and Bailey in the manner the Petition has proposed any more than a dermatologist could be expected to also perform open heart surgery. Ex. 2020 ¶ 39.

Dr. Anderson's own experience in supervising a team that attempted to implement an application within a parallel processing environment is illustrative. Ex. 2020 ¶ 40. In addition to support from Dr. Anderson, this team of engineers worked in close collaboration with a DSP application expert at Texas Instruments for six months in order to implement a custom computer vision application on a multi-processor system supplied by Texas Instruments (Texas Instrument's multi-processor daVinci platform). Even with training, expert oversight, and extensive

efforts on behalf of the team, the inter-processor communication details prevented the successful implementation of their system. Rather, the final product was only able to do a simple tracking of a black ball on a white background—something that could have more easily been performed on a single processor. While this system was not a speech recognition system, it was a signal processing system that was simpler than a speech recognition system in terms of complexity (only a few hundreds of lines of code). The failure of this design example merely demonstrates that transitioning a signal processing system to any particular multi-processing architecture is not necessarily trivial or practical, and that even those beyond the level of ordinary skill in the field of speech recognition could not necessarily expect success with respect to far simpler combinations than the one the Petition has proposed. Ex. 2020 ¶ 41; *Kloster Speedsteel AB v. Crucible, Inc.*, 793 F.2d 1565, 1574 (Fed. Cir. 1986) ("If the invention here would not have been obvious to one of extraordinary skill, it follows that in this case it would not be obvious to one with lesser skills.").

The problems with the Petition's combination, and the lack of a reasonable expectation of success, would have been even more acute in view of the fact that the Petition's combination must be a ***real-time*** speech recognition system "in the

sense that no significant delay occurs between" receiving the word in the system and word recognition. Ex. 1003 ¶ 84; Ex. 2020 ¶ 42.

Data access to the various acoustic models was known to be one of the limiting factors in the performance speed of a speech recognition system. Ex. 2020 ¶ 43. The Petition's theory proposes to dedicate one of Chen's clusters of processors and cluster memories to each of Thelen's speech "contexts" and related sub-contexts, e.g., dedicating the "sports" contexts and all subcontexts such as "football, baseball, etc." to one of Chen's clusters. Ex. 1003 ¶ 83. Thelen teaches that sub-contexts such as football, baseball, and so forth, may each have their own models, and where those models exist, "then all of those models may be tested." Ex. 1030 at 9:24-26; Ex. 2020 ¶ 41. According to Mr. Schmandt himself, his combination would require placing all of the sub-context models for the same context (e.g., the models for football, baseball etc. within the sports context) into the ***same*** single cluster memory. Ex. 1003 ¶ 83 (stating "[a] POSITA would have recognized that associated contexts, such as the sports contexts (baseball, football, etc.) would have shared a larger amount of lexical data than contexts which are not associated" and proposing to store "context-related recognition models" for "associated contexts" "in the same lexical memory."); Ex. 2017 at 124:13-125:7

("In using Chen's multiple processors in a cluster, each of those contexts that are related . . . could share some parts of the memory"); Ex. 2020 ¶ 43.

As Mr. Schmandt further explained at his deposition, his combination would require each of the processors in that cluster to request "the particular subset of vocabulary . . . that's germane to its context" from the shared cluster memory.

8 Q. So are you saying that you
9 understand Thelen to mean that even within
10 the context of sports as a whole, as a topic
11 category, there are additional individual
12 memories, in your opinion, that store
13 language models just related to football and
14 then another memory that stores lexical
15 models for baseball and so forth?
16 A. Yes.
17 Q. That's how you understand
18 Thelen?
19 A. Yes.
20 Q. And so what you're really just
21 proposing is to take the collection of all
22 sports-related language models, for
23 instance, and keep them in one memory;
24 right?
25 A. They're in one memory, and each

2 of the processors connected to that memory
3 is looking at the particular subset of
4 vocabulary in that memory that's germane to
5 its context; yes.

Ex. 2017 at 125:8-126:5.

That approach, however, would require multiple processors in one of Chen's clusters to attempt to access the single shared clustered memory ***simultaneously*** in order for each processor to obtain the necessary recognition model data for the relevant sub-context assigned to that processor. Ex. 2020 ¶ 44. For instance, one processor in Chen's "sports" cluster would make a memory call to the cluster memory for portions of the "football" recognition model, while another processor in the same cluster would make a memory call at the same time to the same cluster memory for other model portions, e.g., portions of the "baseball" recognition model. Ex. 2020 ¶ 44. Each processor could only request as much data as it could hold in its cache and compute with at a given time, which is less than an entire model, so each processor would have to continuously make more memory calls to the cluster memory as it moves along the frame of data. Ex. 2020 ¶ 44. Since Mr. Schmandt's combination puts all of the work for a given context and associated sub-contexts into ***one*** cluster and ***one*** cluster memory, the multiple processors in

the cluster would be continuously making memory calls to the cluster memory to address a common speech recognition task using multiple different associated models, causing a problem known in the art as "memory collisions." Ex. 2020 ¶ 45. Each memory collision causes one or more processors to pause in its computation as the processor waits for its turn to retrieve data from memory. Ex. 2020 ¶ 45. As a result of the numerous memory collisions that would ensue from Mr. Schmandt's proposal, the combined system of Chen, Thelen, and Bailey could not be expected to operate in real time or near real time. Ex. 2020 ¶ 45.

Indeed, it was known in the art of speech recognition at the time that adding processors to the speech recognition tasks in a parallel processing architecture could in fact ***slow down*** the system's performance. *Mathew I*, for instance, compares a two-processor system (one for distance calculation and one for search) to a five-processor system (one processor for the calculation stage, and four for search), and reports that "the resulting 5 processor version ***was slower*** than the 2 processor version due to the high synchronization overhead." *Id.*; Ex. 2020 ¶ 46.

Thus, even though Bailey and Chen both describe architectures that use multiple processors, implementing Thelen's speech recognition teachings using Chen's clustered processor and distributed private memory architecture would

have been far from a "simple substitution," as Mr. Schmandt's declaration simplistically presumes. *Compare* Ex. 1003 ¶ 84 *with* Ex. 2020 ¶ 47.

The evidence of record thus does not demonstrate that an ordinary artisan in the field of speech recognition could have had a reasonable expectation of success with respect to the combination of Chen with Thelen and Bailey. 2020 ¶ 48.

V. The Petition fails to show a motivation to combine Thelen, Bailey, and Chen in the manner proposed, and the POSA would not have been motivated to do so

The Petition's motivation to combine with respect to the combination of Chen with Thelen and Bailey rests on the purported expected benefit of "increase[ing] the processing power of the speed recognition circuit while speeding the responsiveness of the system and . . . keeping the financial cost of the overall system low." Ex. 1003 ¶ 84. According to Mr. Schmandt's declaration, a POSA would have been motivated to combine with Chen based on the alleged benefit of increased "recognition speed." *Id.* Mr. Schmandt's declaration additionally alleges that parallel processing systems generally offer the advantages of "flexibility and scalability," as purportedly taught in *Hon*, which would have further motivated the combination. *Id.* The evidence, however, does not support any such benefits, or the existence of a motivation to combine.

Mr. Schmandt's declaration assumes that the clustered processor and memory arrangement of Chen would have necessarily improved the speed and power of Thelen's speech recognition. Ex. 1003 ¶ 84. But Mr. Schmandt's declaration does not undertake any quantitative or other particularized analysis to demonstrate *why* Chen would allegedly speed up Thelen, or *by how much*. Rather, Mr. Schmandt seems to simply ***assume*** the conclusions that (1) the combination with Chen would add more processing power to Thelen's system and (2) that more processors from Chen would have sped up Thelen's speech recognition.

The evidence of record, however, refutes both of Mr. Schmandt's declaration assumptions.

First, Mr. Schmandt's deposition testimony refutes the assumption that combining Chen with Thelen would add processing power beyond what Thelen already had. According to Mr. Schmandt's initial testimony, the combination with Chen would not increase the number of processors in Thelen at all.

```
23                So according to your theory,
24   Thelen in unmodified form would already have
25   the same number of processors as Thelen in
```

2 modified form would have after being
3 modified in view of Chen?
4 A. Yes.

Ex. 2017 at 130:23-131:4. But as Mr. Schmandt admitted, his declaration does not explain how a POSA could expect increased processing power—or be motivated to combine Chen with Thelen and Bailey based on such an expectation—if the combination with Chen would not increase Thelen's processors.

11 Q. You don't explain how the
12 processing power would be increased given
13 that Thelen already has the same number of
14 processors as would exist after the
15 modification in view of Chen; right?
16 A. Yes. For the -- yes, you're
17 right.
18 Q. And you see in paragraph 84 you
19 refer to speeding the responsiveness of the
20 system? Do you see that?
21 (Witness reviewing document.)
22 A. Yes.

Ex. 131 at 11-22. And while Mr. Schmandt then sought to equivocate about whether the combination with Chen would increase the number of processors in Thelen, he ultimately admitted that he simply ***did not know*** whether it would.

3 Q. So maybe your theory requires
4 adding processes to Thelen and maybe it
5 doesn't; fair?
6 A. Correct.

Ex. 2017 at 136:3-6.

Accordingly, the first assumption in Mr. Schmandt's "speed improvement" motivation—that the combination with Chen would increase the processing power of Thelen + Bailey—is baseless in view of his deposition testimony.

Second, contemporaneous real-world evidence demonstrates that the POSA could ***not*** have reasonably expected the combination of Chen with Thelen and Bailey to speed up Thelen's speed recognition process. Ex. 2020 ¶ 50. *Mathew I* provides a detailed report on computer architecture modifications intended to improve the speed of a speech recognizer, and in particular through the addition more processors. Ex. 2021 at 11. *Matthew I* explains that it developed "a parallel version" of the Sphinx speech recognition system in which the distance calculation (GAU, or Gaussian probability estimation) was run on a separate processor from

the search stage (HMM processing) with parallelization. *Id.* This two-processor parallel processing design was only 1.67x faster than the "original sequential version," not 1.97x faster as the authors had predicted based on Amdahl's law. *Id.*; Ex. 2020 ¶ 51. The authors then further modified the search stage (HMM phase) "to use 4 processors instead of 1," thus moving from a two processor system (one for distance calculation and one for search) to a five processor system (one processor for the calculation stage, and four for search). *Id.*; Ex. 2020 ¶ 51. As *Mathew I* reports, "the resulting 5 processor version ***was slower*** than the 2 processor version due to the ***high synchronization overhead***." *Id.*; Ex. 2020 ¶ 52.

As *Mathew I* demonstrates, the ordinary artisan could not have assumed that Chen's complEx. clustered parallel processing architecture with numerous processors in each cluster and multiple distributed memories would have improved the speed of Thelen's speech recognition system (whether combined with Bailey or not) in view of the high synchronization overhead, memory collisions, and other complications that such a system would have implicated. Ex. 2020 ¶ 53.

As Dr. Anderson further explains, the POSA would not have been motivated to implement Thelen's speech recognition techniques in Chen's clustered processing architecture. Ex. 2020 ¶ 54. 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. 1003 ¶ 77 (Mr. Schmandt stating that Thelen's storage 340 stores the "complete lexical data," and that Thelen's recognition models are "loaded from a storage 340."); Ex. 1030 (Thelen) at 7:30-34, 5:22-24, Fig. 3; Ex. 2020 ¶ 54. The relevant embodiment the Petition selects from Chen, however, relies on multiple private memories, with one memory per cluster, in which access to those private memories is extended also to directly adjacent clusters. Ex. 2020 ¶ 54. As Chen itself states, "[t]o effectively port an existing software program from a shared memory model to a private memory model can require ***significant*** reprogramming effort." Ex. 1005 at 2:6-9 (emphasis added); Ex. 2020 ¶ 55. That is because "a distributed memory model necessarily requires that the processes and data which comprise a parallel programming problem be partitioned into separate tasks that are each small enough to fit into the private local memory of a processor and at the same time are large enough to recover the ***overhead*** required to establish a parallel task in a message passing environment." Ex. 1005 at 2:9-17 (emphasis added); Ex. 2020 ¶ 54. The POSA would have been aware of Chen's warning that porting Thelen's speech recognition techniques 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 ¶ 55. Mr. Schmandt's declaration does not address those issues, much less provide evidence that the POSA would have been motivated to combine Chen with Thelen and Bailey in view of those significant impediments. Ex. 2020 ¶ 56.

Mr. Schmandt's particular proposal as to how the POSA would have allegedly utilized Chen's hardware to perform Thelen's speech recognition only further demonstrates that a POSA would not have been motivated to make the combination. Thelen is a real time speech recognition system, and Mr. Schmandt proposes that the combination would have also been required to perform in real time. Ex. 1003 ¶ 84. Moreover, as Mr. Schmandt admitted at his deposition, Thelen in unmodified form expresses no doubt or concern about achieving real time recognition. Ex. 2017 at 155:7-17. Mr. Schmandt's combination, however, would require multiple processors in one of Chen's clusters to simultaneously request recognition model data from one cluster memory at the same time.

8 Q. So are you saying that you
9 understand Thelen to mean that even within
10 the context of sports as a whole, as a topic
11 category, there are additional individual
12 memories, in your opinion, that store
13 language models just related to football and
14 then another memory that stores lexical
15 models for baseball and so forth?
16 A. Yes.
17 Q. That's how you understand
18 Thelen?
19 A. Yes.
20 Q. And so what you're really just
21 proposing is to take the collection of all
22 sports-related language models, for
23 instance, and keep them in one memory;
24 right?
25 A. They're in one memory, and each

2 of the processors connected to that memory
3 is looking at the particular subset of
4 vocabulary in that memory that's germane to
5 its context; yes.

Ex. 2017 at 125:8-126:5, 124:13-125:7; Ex. 1003 ¶ 83; Ex. 2020 ¶ 57.

But as explained earlier, this would cause extensive memory collisions, causing the combination system to fall far behind real time operation. Ex. 2020 ¶ 58. The POSA would not have been motivated to modify Thelen as the Petition has proposed for that reason alone. Ex. 2020 ¶ 58.

Equally without merit is Mr. Schmandt's "flexibility and scalability" motivation. Ex. 1003 ¶¶ 71, 72. Mr. Schmandt's declaration cites *Hon* as purportedly teaching that clustered processing architectures in general present the benefits of flexibility and scalability, and that this would have motivated the combination of Chen with Thelen and Bailey. But Mr. Schmandt's deposition testimony clarifies that *Hon*'s cited passages instead teach that MARS was flexible because its pipeline was programmable, and that it was scalable because more processors could be added to it. Ex. 2017 at 105:8-106:4. In both instances, *Hon* was comparing the MARS system to another system, SRI-Berkeley, which did not have a programmable pipeline or extendable processor architecture. *Id. Hon's* discussion of MARS is thus specific to two particular advantages of MARS over SRI-Berkeley, and *Hon* thus does not provide a generalized "flexibility and scalability" motivation for combining Thelen / Bailey with a clustered processing architecture, much less Chen's in particular. Ex. 2020 ¶ 59. Indeed, Mr. Schmandt

has made no showing that Thelen itself, without modification, was not already both flexible and scalable in the same manner that *Hon* describes with respect to MARS.

Ultimately, Mr. Schmandt testified that it is ***never*** desirable to complicate a speech recognition system with additional hardware or software when that system can already perform adequately.

> 9 A. Well, it's always not desirable
> 10 to complicate hardware or software when an
> 11 existing hardware or software can perform
> 12 adequately.

Ex. 2017 at 163:4-12 (emphasis added). Yet Mr. Schmandt and Apple have made no showing that Thelen's system could not perform adequately without being combined with Chen or Bailey. Rather, the evidence demonstrates that Thelen would perform ***worse***, not better, after the proposed combination. Ex. 2020 ¶ 60. Accordingly, there is no motivation to combine here under Mr. Schmandt's own criteria.

VII. Conclusion

As the Supreme Court explained in *KSR*, the ordinary artisan is a person of *ordinary* creativity. *KSR Int'l Co. v. TeleflEx. Inc.*, 550 U.S. 398, 421 (2007) ("A person of ordinary skill is also a person of ordinary creativity"). Just as a person of "***ordinary*** skill" is "not an automaton," *id.*, the person of ordinary skill is also not a person of above-average creativity in the field of art. *Id.* The line between invention and obviousness is thus not whether *any person* ***could*** have arrived at the design of the challenged claims at the time. Rather, it is whether a person of *ordinary* creativity would have been motivated to do so with a reasonable expectation of success. *Id.* The Petition and Mr. Schmandt have not made that showing here, and have instead proposed a combination that the POSA would not and could not have brought together at the time of the '319 Patent. The Board should not find any challenged claim unpatentable here.

Respectfully submitted,

/s/ *Peter C. Knops*

Peter C. Knops, Reg. No. 37,659
Kayvan B. Noroozi, *Pro Hac Vice* (To be Filed)
NOROOZI PC
11601 Wilshire Blvd., Suite 2170
Los Angeles, CA 90025
Attorneys for the Patent Owner

Date: September 26, 2023

CERTIFICATION OF WORD COUNT

The undersigned hereby certifies that the portions of the above-captioned **PATENT OWNER'S RESPONSE** specified in 37 C.F.R. § 42.24 have 5,243 words, in compliance with the 14,000 word limit set forth in 37 C.F.R. § 42.24(b)(1). This word count was prepared using Microsoft Word 365.

/Peter C Knops/
Peter C Knops

Counsel for Patent Owner

Date: September 26, 2023

CERTIFICATION OF SERVICE

The undersigned hereby certifies that the foregoing **PATENT OWNER'S RESPONSE** was served electronically via e-mail on September 26, 2023, on the following counsel of record for Petitioner:

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

/Peter C. Knops/
Peter C. Knops
Counsel for Patent Owner

Date: September 26, 2023