

Compare Results

Old File:

**IPR2023-00037 - PTACT Patent Owners
Response.pdf**

41 pages (987 KB)

versus

New File:

CORRECTED IPR2023-00037 Final POR.pdf

41 pages (884 KB)

3/7/24, 5:00:09 PM

Total Changes

35

Text only comparison

Content

17

Replacements

16

Insertions

2

Deletions

Styling and Annotations

0

Styling

0

Annotations

[Go to First Change \(page 1\)](#)

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC.,
Petitioner,

v.

Zentian Limited
Patent Owner.

Case IPR2023-00037
Patent No. 10,971,140

PATENT OWNER'S RESPONSE
[CORRECTED]

TABLE OF CONTENTS

I.	Introduction	1
II.	Background of the '140 Patent	3
III.	Person of ordinary skill in the art	6
IV.	The POSA would not have had a reasonable expectation of success with respect to the Petition's combination of Jiang and Chen	7
V.	The Petition and Mr. Schmandt fail to demonstrate a motivation to combine Jiang with Chen	16
VI.	The Petition fails to show obviousness as to limitations 1(d) and 1(e)	27
VII.	Conclusion	35

EXHIBIT LIST

Exhibit No.	Description
2017	Deposition of Christopher Schmandt dated September 6, 2023
2018	<i>Intentionally left blank</i> 
2019	Hennessy & Patterson, <i>Computer Architecture, A Quantitative Approach</i> , Third Edition (2003) (“Hennessy & Patterson”)
2020	Declaration of  David 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 been motivated to implement Jiang's speech recognition teachings in Chen's clustered processor and memory architecture with a reasonable expectation of success.

But neither Jiang nor Chen enables such a combination, so the Petition must prove enablement through evidence outside those references. The Petition and Mr. Schmandt, however, fail to make that showing. Indeed, Mr. Schmandt admitted under cross-examination that he has **never** built or designed the processor to memory architecture for any of the speech recognition systems he identifies in the background of his declaration. Mr. Schmandt further admitted he has never supervised anyone in the process of mapping speech recognition software to a hardware architecture like Chen's.

As Dr. Anderson explains in detail, the Petition's combination would have raised a number of complications and conflicts that the POSA would not have been qualified to address or resolve. Indeed, Jiang and Chen are inherently ill-suited for combination, since Jiang's technique requires extensive communication between computational components whereas Chen does not allow communication between some of its processors and memories.

Moreover, real-world evidence from contemporaneous references teaches that the combination would **not** be expected to speed up Jiang's recognition process, and that the POSA would not have been motivated by such an expectation. To the contrary, *Mathew I* teaches that a five-processor parallel processing system **slowed down** the speech recognition process compared to a two-processor system due to the associated synchronization overhead. Mr. Schmandt's other purported motivations—alleged cost benefits and “flexibility and scalability”—are likewise baseless, as Mr. Schmandt's deposition revealed. By contrast, Dr. Anderson explains in detail why the POSA would not have been motivated to combine Jiang with Chen in view of the known challenges and disadvantages of such a combination balanced against the low likelihood of benefits.

Finally, the Petition further fails because it provides no theory as to how the *combination* of Jiang and Chen would operate to meet limitations 1(d) and 1(e). As Mr. Schmandt admitted, his theory requires that Jiang *modified* to use Chen's hardware would practice both limitations, but his declaration does not even *mention* Chen with respect to limitations 1(d) and 1(e), much less explain *how* the combination of Jiang and Chen would meet them. And although Mr. Schmandt attempted to fill the void with a new, *ad hoc* theory presented for the first time at his deposition, the evidence demonstrates that theory would not have been enabled.

Mr. Schmandt's failure to present a workable explanation as to how the POSA could have arrived at the invention of the challenged claims by combining Jiang and Chen is definitive evidence that the claims are far from obvious.

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

II. Background of the '140 Patent

U.S. Patent 10,971,140, 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, 1:18-20. The '140 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, 3:13-18. "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, 3:19-24. The '140 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, 3:62-66.

The specification of the '140 patent thus provides: “[T]he present 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. . . . Each processor within a group can access the same lexical data as any other processor in the group. The controller can thus control the parallel processing of input speech parameters in a more flexible manner. For example, it allows more than one processor to process input speech parameters using the same lexical data in a lexical memory. This is because the lexical data is segmented into domains which are accessible by multiple processors.” Ex. 1001, 3:62-4:18.

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).

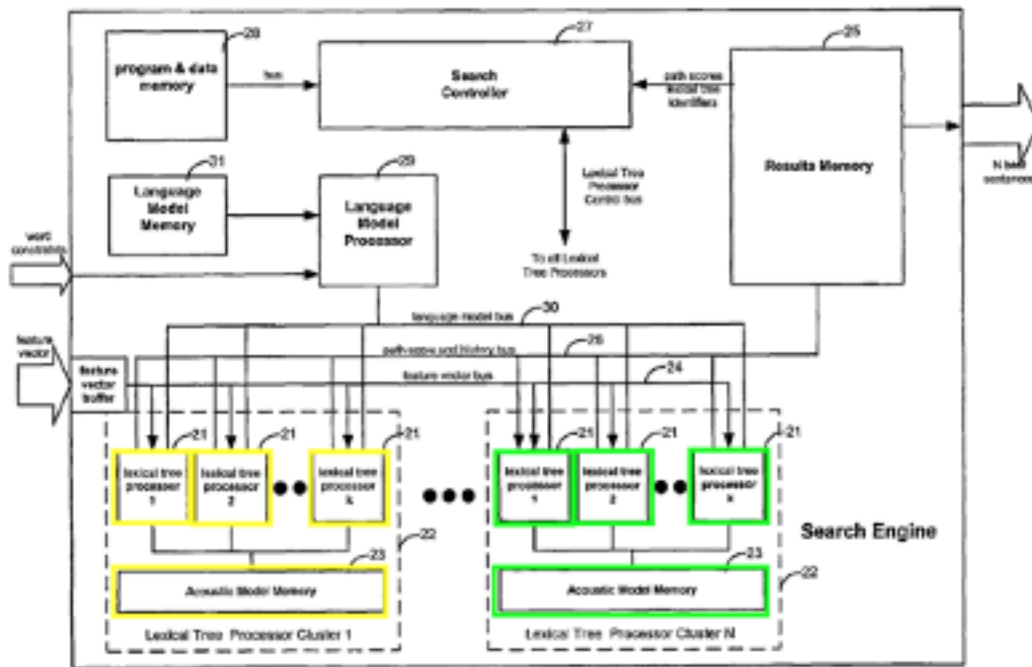


Fig 2

Moreover, the '140 patent expressly distinguishes its novel architecture from two prior known alternative designs, which the patent describes as less advantageous. In particular, the '140 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, 4:1-9. The patent thus distinguishes

its design from (1) a one-memory-to-**all**-processors design, which it describes as bandwidth limited in the processor to memory interface; and (2) a one-memory-to-**one**-processor design, which would require more memory interfaces and would be less flexible.

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 POSA would not have had a reasonable expectation of success with respect to the Petition's combination of Jiang and Chen

The Petition's theory as to all challenged claims relies on at least a combination of Jiang and Chen. In particular, the Petition and Mr. Schmandt propose that the ordinary artisan would have found it obvious "to configure a computing platform comprising clusters of processors as taught by *Chen* to perform the speech recognition techniques of *Jiang*." Ex. 1003 ¶ 68.

“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, Jiang does not enable performing its speech recognition techniques using the particular clustered processor embodiment the Petition has selected from Chen, and Chen does not enable using that clustered processor embodiment to perform Jiang’s speech recognition techniques. Ex. 2020 ¶ 24. Mr. Schmandt admitted that reality at his deposition.

9 Chen itself doesn't specifically
10 teach how to implement Jiang's methods or
11 system; right?
12 A. Correct.
13 Q. Jiang does not expressly teach
14 how to implement Chen's processor to memory
15 architecture with Jiang's methods or
16 systems; right?
17 A. Of course not.

Ex. 2017 at 86:9-17.

Accordingly, the evidence of record beyond Jiang and Chen must establish that the *ordinary* artisan could have “configure[d] a computing platform comprising clusters of processors as taught by *Chen* to perform the speech recognition techniques of *Jiang*,” as the Petition’s theory requires. Ex. 1003 ¶ 68.

Mr. Schmandt and the Petition, however, do not provide that evidence. Instead, Mr. Schmandt’s declaration simply states the sweeping and conclusory generalization that the “modifications necessary, such as configuring *Chen*’s circuitry to specifically recognize speech pursuant to *Jiang* would have required only ***software programming and well-known computing techniques and structures***, and thus would have led a POSITA to a reasonable expectation of success.” Ex. 1003 ¶ 68 (emphasis added).

As Mr. Schmandt admitted under cross-examination, his declaration is thus silent as to any underlying details to support his conclusion as to the POSA’s capability; instead, his declaration merely states the conclusion without any underlying explanation.

13 Q. In your declaration, you don't
14 specify what software programming or
15 computer techniques and structures would be
16 needed to configure Chen's circuitry to
17 recognize speech pursuant to Jiang; fair?
18 A. I don't list the details, but I
19 do express an opinion that a POSITA would
20 understand those details, would understand
21 how to do it.

Ex. 2017 at 48:13-21.

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 POSA's capabilities are also not credible in view of his deposition testimony.

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.

Moreover, Mr. Schmandt 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 146:20-24.

Mr. Schmandt thus lacks an adequate basis from his personal knowledge and experience to opine that the combination of Jiang and Chen was within the skill level of ordinary artisan in the field of speech recognition.

By contrast, as Dr. David Anderson explains, implementing Jiang's speech recognition techniques on Chen's clustered processor and memory architecture 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, without the benefit of the '140 Patent. Ex. 2020 ¶ 27. In particular, moving Jiang's speech recognition techniques to Chen's clustered, parallel processors and memories would have required coordinating multiple caches, avoiding memory conflicts, controlling task sharing in an efficient manner, resolving synchronous bottlenecks, and addressing communication bandwidth and latency issues among the various hardware components, and developing a messaging strategy to coordinate

information sharing between and within clusters, among other challenges. Ex. 2020 ¶ 27.

Such tasks were not within the skill level of the ordinary artisan in the field of speech recognition as defined by the Petition prior to the '140 Patent. Ex. 2020 ¶ 27. As Dr. Anderson explains, 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 ¶¶ 17-19, 28 (pp. 7-9, 17). Students of electrical or computer engineering specialize in one of these areas after taking some general courses. *Id.* The *ordinary* artisan in the field of speech recognition would have specialized in the area of digital signal processing (DSP), and would have further pursued studies specific to speech recognition, an advanced subspecialty of DSP, in their graduate studies. *Id.* The POSA in the field of speech recognition would not have been capable of addressing the complex challenges of combining Chen with Jiang in the manner the Petition has proposed any more than a dermatologist could be expected to also perform open heart surgery. Ex. 2020 ¶ 28 (p. 17).

Moreover, speech recognition requires evaluating the comparative likelihoods of many possible outcomes for each incoming frame of samples. Ex. 2020 ¶ 29 (p. 17). Jiang teaches the use of the Viterbi decoding algorithm. Ex. 1004 at 1:19-39, 2:1-46. The Viterbi algorithm compares the likelihoods of all possible sequences for a given input frame, given the most likely outcomes from the previous frame. *Id.*; Ex. 2020 ¶ 29 (p. 18). To perform these comparisons requires extensive communication between computational components that can have a drastic impact on the system design and performance. Ex. 2020 ¶ 29 (p. 18). By contrast, Chen teaches that the memory associated with one cluster is ***not directly or adjacently accessible*** by the processors and memories in each other cluster. See Ex. 1003 ¶ 95; Ex. 1005 at 9:10-39; Ex. 2020 ¶ 29 (p. 18). As a result, performing Viterbi decoding when the information needed by each node may be in a cluster that is not directly accessible, as in Chen, is not likely to be successful in a practical speech recognition system. Ex. 2020 ¶ 29 (p. 18).

As explained by Hennesy & Patterson and Dr. Anderson, communication between parallel processing nodes is one of several considerations that can determine the feasibility of a particular parallel architecture for use with a particular problem. Ex. 2019 at 534-535; Ex. 2020 ¶ 30 (p. 18). Hennesy & Patterson also points out that the private memory model in which memory is

connected only within a cluster may be suitable “***for applications that require little or no communication.***” Ex. 2019 at 533; Ex. 2020 ¶ 30 (p. 18). Chen implements such a private memory architecture but the speech recognition search stage is an application that requires significant communication. Ex. 2020 ¶ 30 (p. 18). Therefore, based on teachings at the time of the ’140 Patent, the POSA would not expect the proposed combination of Chen with Jiang to be successful. Ex. 2020 ¶ 30 (pp. 18-19).

Dr. Anderson’s own experience in supervising a team that attempted to implement an application within a parallel processing environment is illustrative. Ex. 2020 ¶ 31. 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). *Id.* 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. *Id.* 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. *Id.* 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). *Id.* 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. *Id.*; *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.”).

Thus, even accepting *arguendo* that Jiang and Chen both contemplate the use of a plurality of processors, a POSA would not have “been capable of simply substituting” Chen’s clustered processors to implement Jiang’s speech recognition techniques, as Mr. Schmandt’s declaration simplistically presumes. *Compare* Ex. 1003 ¶¶ 67, 74 with Ex. 2020 ¶¶ 27-32 (pp. 16-20).

The evidence of record thus does not demonstrate that an ordinary artisan in the field of speech recognition could have configured Chen’s clustered processor and memory architecture to perform Jiang’s speech recognition techniques with a reasonable expectation of success.

V. The Petition and Mr. Schmandt fail to demonstrate a motivation to combine Jiang with Chen

Apple and Mr. Schmandt allege that the POSA would have been motivated to combine Jiang with Chen due to the alleged benefits of improved speed and power, a relaxed pruning threshold, and reduced cost. Ex. 1003 ¶¶ 67, 74.

Separately, Mr. Schmandt contends that “the flexibility and scalability afforded to the MARS machine by its clustered architecture [] would have motivated a POSITA to implement clustered architecture in other speech recognition circuits.” Ex. 1003 ¶ 71. 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 Jiang's speech recognition, and thus allowed for relaxed pruning threshold. Ex. 1003 ¶¶ 67, 74. Notably, Mr. Schmandt's declaration does not undertake any quantitative or other particularized analysis to demonstrate *why* Chen would allegedly speed up Jiang and relax its pruning threshold, or *by how much*. Ex. 2020 ¶ 35. Rather, Mr. Schmandt seems to simply assume the conclusion that more processors would have been better. *Id.*

Contemporaneous real-world evidence, however, refutes Mr. Schmandt's assumptions. Ex. 2020 ¶ 36. *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. *Mathew 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. Ex. 2021 at 11; Ex. 2020 ¶ 36. 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). Ex. 2021 at 11; Ex. 2020 ¶ 36. As *Mathew I* reports, "the resulting 5 processor version **was slower** than the 2 processor version due to the **high synchronization overhead**." Ex. 2021 at 11 (emphasis added); Ex. 2020 ¶ 36.

As Dr. Anderson explains, *Mathew I* demonstrates that 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 Jiang's speech recognition system or relaxed its pruning threshold in view of the high synchronization overhead and other complications that such a system would have implicated. Ex. 2020 ¶ 37.

Mr. Schmandt's declaration testimony as to the alleged cost-based motivations of the combination are likewise entirely baseless. As Mr. Schmandt admitted at his deposition, his declaration opines that the combination of Jiang and Chen would have been less expensive, but never states what the combination would have been *less expensive than*, or on what basis.

14 Q. In paragraph 74 of your
15 declaration near the bottom of page 58, do
16 you see that you say modifying Jiang's
17 speech recognition circuit to use clusters
18 of processors as taught by Chen reduces the
19 expense incurred by grouping commercially
20 available, simple processors into clusters?

21 A. Yes.

22 Q. When you say reducing the
23 expense, you don't explain what you're
24 comparing the expense to; right?

25 A. No. You're right.

2 Q. You also don't provide a
3 quantitative analysis of how or how much the
4 expense would be reduced; right?

5 A. No. It would be silly to try to
6 perform a quantitative analysis on a
7 hypothetical and it's really not my job to
8 price out all these various parts as of 2000
9 whatever.

Ex. 2017 at 61:14-62:9. Mr. Schmandt also made clear that his declaration testimony was not based on a cost comparison between Chen's clustered processing architecture and any other allegedly alternative clustered processing architectures.

8 Q. In your declaration, you did not
9 provide an analysis showing the degree to
10 which Chen's clustered architecture would be
11 cheaper, if at all, compared to other
12 clustered architectures; right?

13 A. I cited to a number of other
14 clustered architectures, but I did not do a
15 cost comparison; no.

Ex. 2017 at 63:8-15. And Mr. Schmandt likewise admitted that his declaration provides no quantitative cost comparison between Jiang's unmodified architecture

and his proposed modified architecture in combination with Chen. Ex. 2017 at 40:23-41:4.

In sum, there is no basis for Mr. Schmandt's declaration testimony that the POSA would have been motivated to combine Jiang with Chen based on a purported cost or expense benefit from the combination. Ex. 2020 ¶ 39.

Moreover, Mr. Schmandt testified that a motivation to modify would require the ordinary artisan to first assess the cost implications of a potential modification against the imputed or actual value of the modification, considering both any specific performance improvements and the value of those improvements to end users. Ex. 2017 at 41:18-42:13. According to Mr. Schmandt: "Cost would be part of a design tradeoff and also looking at your customer base." Ex. 2017 at 42:12-42:13. Yet Mr. Schmandt readily admitted he did not undertake any qualitative or quantitative cost-benefit analysis in reaching his opinions regarding a motivation to combine.

14 Q. In your declaration, you did not
15 undertake a quantitative or qualitative cost
16 versus purported improvement in comparison
17 to customer-based analysis to get into those
18 details; fair?

19 A. Right. That's correct.

Ex. 2017 at 42:14-19.

Mr. Schmandt's declaration testimony as to a cost-based motivation thus does not meet his own criteria for what a POSA would need to consider before forming a motivation.

Nor does it meet the Federal Circuit's requirement as to proving a motivation to combine. *Winner Int'l Royalty Corp. v. Wang*, 202 F.3d 1340, 1349 (Fed. Cir. 2000) (holding that "[m]otivation to combine requires" evidence the alleged benefits of the combination, in view of the associated tradeoffs and costs, would have been "on balance, desirable.>").

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 Jiang and Chen. *Id.* 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; Ex. 2020 ¶ 40. 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. Ex. 2017

at 105:8-106:4; Ex. 2020 ¶ 40; Ex. 1006 at 24. *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 Jiang with a clustered processing architecture, much less Chen's in particular. Ex. 2020 ¶ 40. Indeed, Mr. Schmandt has made no showing that Jiang itself, without modification, was not already both flexible and scalable in the same manner that *Hon* describes with respect to MARS. Ex. 2020 ¶ 41.

Ultimately, Mr. Schmandt testified that it is ***never*** desirable to add another processor to a speech recognition system when that system can already perform adequately.

4 Could you enumerate for me
5 reasons why it's not desirable to add
6 another processor to an existing speech
7 recognition system as that consideration is
8 commonly understood in the field of art?

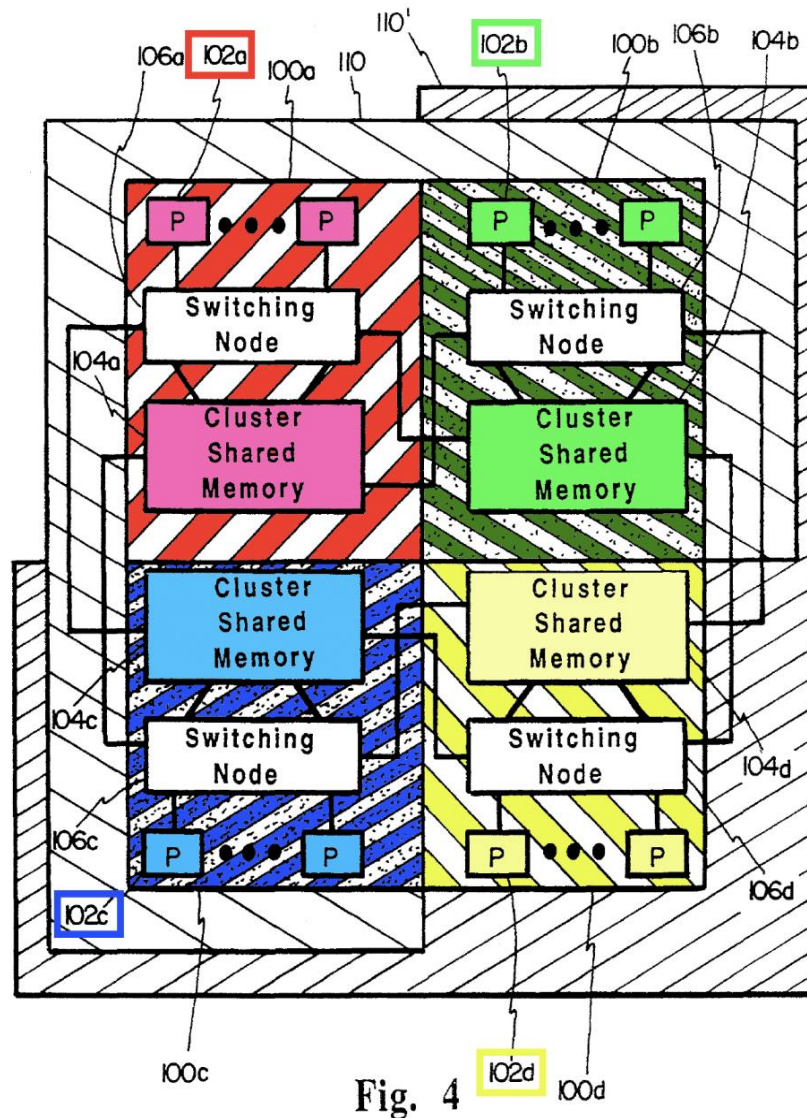
9 A. Well, it's always not desirable
10 to complicate hardware or software when an
11 existing hardware or software can perform
12 adequately.

Ex. 2017 at 163:4-12 (emphasis added). Yet Mr. Schmandt and Apple have made no showing that Jiang's system could not perform adequately without being combined with Chen. Accordingly, there is no motivation to combine here under Mr. Schmandt's own criteria.

As Dr. Anderson further explains, the POSA would not have been motivated to implement Jiang's speech recognition techniques in Chen's clustered processing architecture. Ex. 2020 ¶¶ 42-45. Jiang's existing architecture relies on a shared memory structure in which all components have access to the same memory. Ex. 1004, Fig. 1; Ex. 2020 ¶ 42. The relevant embodiment the Petition selects from Chen, however, relies on a distributed memory consisting of multiple private memories, with one memory per cluster, in which access to those private memories is extended also to directly adjacent clusters. Ex. 2020 ¶ 42. 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 ¶ 42. 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 ¶ 42. The POSA would have been aware of Chen’s warning that porting Jiang’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 Jiang’s speech recognition techniques to balance the competing problems of memory size and overhead. Ex. 2020 ¶ 43. In addition, the overhead required to perform message passing (communications for information sharing) between private-memory clusters can be onerous. Ex. 2020 ¶ 43. One example of this problem was demonstrated in *Mathew I* and cited previously. *Id.* Since speech recognition search requires information to be shared among multiple recognition paths (candidate HMM state sequences) during the Viterbi decoding, the design of the communication and memory architecture would be critical. *Id.* Chen presents a particularly poor choice for the speech recognition search stage because it

explicitly *does not* have direct communication between all the clusters. Ex. 1003 ¶ 95; Ex. 1005, 9:10-39; Ex. 2020 ¶ 43. The POSA would not have been motivated to combine Jiang with Chen in the manner proposed in view of those obstacles. Ex. 2020 ¶ 44. Indeed, the POSA would not have expected to successfully manage the overhead implications of such a combination in order to reasonably expect benefits. Ex. 2020 ¶ 44. As already explained, *Mathew I* teaches that even a five-processor parallel processing system ended up performing more slowly than a two-processor system “due to the high synchronization overhead.” Ex. 2021 at 11; Ex. 2020 ¶ 44. The Petition’s combination, by contrast, requires at least eight processors and four distributed memories running in parallel across four clusters. Ex. 1003 ¶ 73.



The ordinary artisan would have known that mapping Jiang’s speech recognition techniques to such a complex architecture would have not, on balance, have been desirable in view of the high downsides and the low likelihood of benefits. Ex. 2020 ¶ 45.

Accordingly, the POSA would not have been motivated to combine Jiang with Chen. Ex. 2020 ¶ 45.

VI. The Petition fails to show obviousness as to limitations 1(d) and 1(e)

Limitation 1(d) requires that “the speech recognition circuit is configured to generate an initial score for an audio sample.” Limitation 1(e) further requires that “the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.”

According to Mr. Schmandt's declaration, Jiang's tree search engine 74 allegedly performs limitations 1(d) and 1(e) of the challenged claims.

17 Q. So my question is with respect
18 to the so-called initial score that you
19 identify in Jiang in paragraph 87 of your
20 declaration, do you see that it is the tree
21 search engine 74 in Jiang that calculates
22 what you identify as the so-called initial
23 score?

24 A. Yes. That's what I just said in
25 answer to your previous question.

Ex. 2017 at 68:17-25.

10 Q. The teachings from Jiang that
11 you rely on for purposes of your theory as
12 to claim 1(e) are performed by Jiang's tree
13 search engine 74; right?
14 A. Yes.

Ex. 2017 at 83:10-14.

But as Mr. Schmandt admitted at his deposition, his combination theory requires **modifying** Jiang's tree search engine 74 to use Chen's clustered processor architecture. Ex. 2017 at 46:12-22.

Thus, Mr. Schmandt's declaration and the Petition were required to explain how Jiang's tree search engine 74 *modified in view of Chen* would operate to meet limitation 1(d) and 1(e). *ActiveVideo Networks, Inc. v. Verizon*, 694 F.3d 1312, 1327 (Fed. Cir. 2012) (holding that expert testimony failed to show obviousness where the expert "failed to explain . . . **how any specific combination would operate or read on the asserted claims.**") (emphasis added). In particular, since Jiang's tree search engine 74 would have been entirely modified to use Chen's clusters of processors, distributed memories, and parallel processing architecture, the Petition and Mr. Schmandt must provide an explanation of *how* Jiang modified in view of Chen would have operated to meet limitations 1(d) and 1(e). *Id.*

But as Mr. Schmandt admitted, his declaration contains no such explanation.

11 Q. In paragraphs 84 to 87 of your
12 declaration, you don't explain whether
13 Chen's clustered processors would be
14 involved in any way in the so-called initial
15 score computation; fair?

16 (Witness reviewing document.)

17 A. No, there is no mention
18 explicitly of Chen in there, but as I
19 discussed elsewhere, Chen would be used
20 advantageously to result in an improvement
21 to the tree search engine. And before you
22 ask any more question, I need a lunch break.
23 It's 1:30 here.

Ex. 2017 at 79:11-23.

8 Q. So in your discussion of claim
9 1(d) in paragraphs 84 to 87, you don't
10 provide an explanation as to how Chen's
11 clustered processors and memory would be
12 used to generate an initial score in view of
13 Jiang if they would be so used at all;
14 right?

15 A. As I have just said, I don't
16 mention Chen in those paragraphs, but I have
17 previously discussed how Chen -- how Chen's
18 teaching suggest an improvement to the
19 performance of Jiang's tree search engine,
20 and as we've just discussed, the tree search
21 engine is the focus of all that I've written
22 in those paragraphs.

Ex. 2017 at 81:8-22.

15 Q. In your paragraphs regarding
16 claim 1(e), you don't expressly discuss
17 anything about Chen; correct?

18 A. Yes. I make no mention of Chen.

Ex. 2017 at 83:15-18.

Accordingly, the Petition fails to meet its burden of explaining “**how** [the combination of Jiang and Chen] would operate or read on the [challenged] claims.”

ActiveVideo Networks, Inc. v. Verizon, 694 F.3d 1312, 1327 (Fed. Cir. 2012)

(emphasis added). The Petition's theory as to limitations 1(d) and 1(e), and thus all challenged claims, cannot prevail on that basis alone.

Yet Mr. Schmandt's attempt to salvage the problem at his deposition only further demonstrates that the Petition's failure was no trivial oversight. Under cross-examination, Mr. Schmandt testified for the first time that **one** processor in **one** cluster in Chen would have been used to arrive at **limitations 1(d) and 1(e)**. Ex. 2017 at 102:5-103:3. Mr. Schmandt further testified that the probability computations of **limitation 1(c)** would have occurred across **all** of the processors and memories disclosed in Chen's Figure 4, i.e., at least eight processors 102(a)-(d), and four memories 104(a)-(d), across four clusters 100(a)-(d).

12 Q. So you're saying that every
13 single processor in Chen's clusters of
14 processors would have been used by the POSA
15 to compute a probability with respect to
16 Jiang's tree search engine 74?
17 A. Yes.

Ex. 2017 at 73:12-17.

4 Q. And the parallel operations in
5 your theory happened across multiple
6 clusters of Chen's processors; right?
7 A. Yes.
8 Q. And multiple acoustic memories?
9 A. Yes.
10 Q. And specifically the ones
11 104a-d; right?
12 A. Yes. And the other ones in that
13 figure.
14 Q. Well, the ones in the figure I
15 think are 104a-d; no?
16 A. Oh, you're right. I'm sorry.
17 One sec.
18 Where does that line go?
19 Yes, you're right. I'm sorry.

Ex. 2017 at 103:4-19.

But as Mr. Schmandt's declaration states, Chen teaches that "not all[] the clusters are interconnected." Ex. 1003 ¶ 95. According to Mr. Schmandt, "the processors 102a in cluster 100a . . . cannot directly or adjacently access the memory 104d of cluster 100d, while the processors 102d in cluster 100d . . . cannot directly or adjacently access the memory 104a." *Id.* Thus, no one processor in one cluster can directly or adjacently access **all** of the data stored across the four

memories 104(a)-(d). Ex. 1003 ¶ 95; Ex. 1005, 9:10-39; Ex. 2020 ¶ 50. Chen's Fig. 4, as annotated by the Petition, visually illustrates that each cluster shared memory only has connections to some but not all the other cluster shared memories.

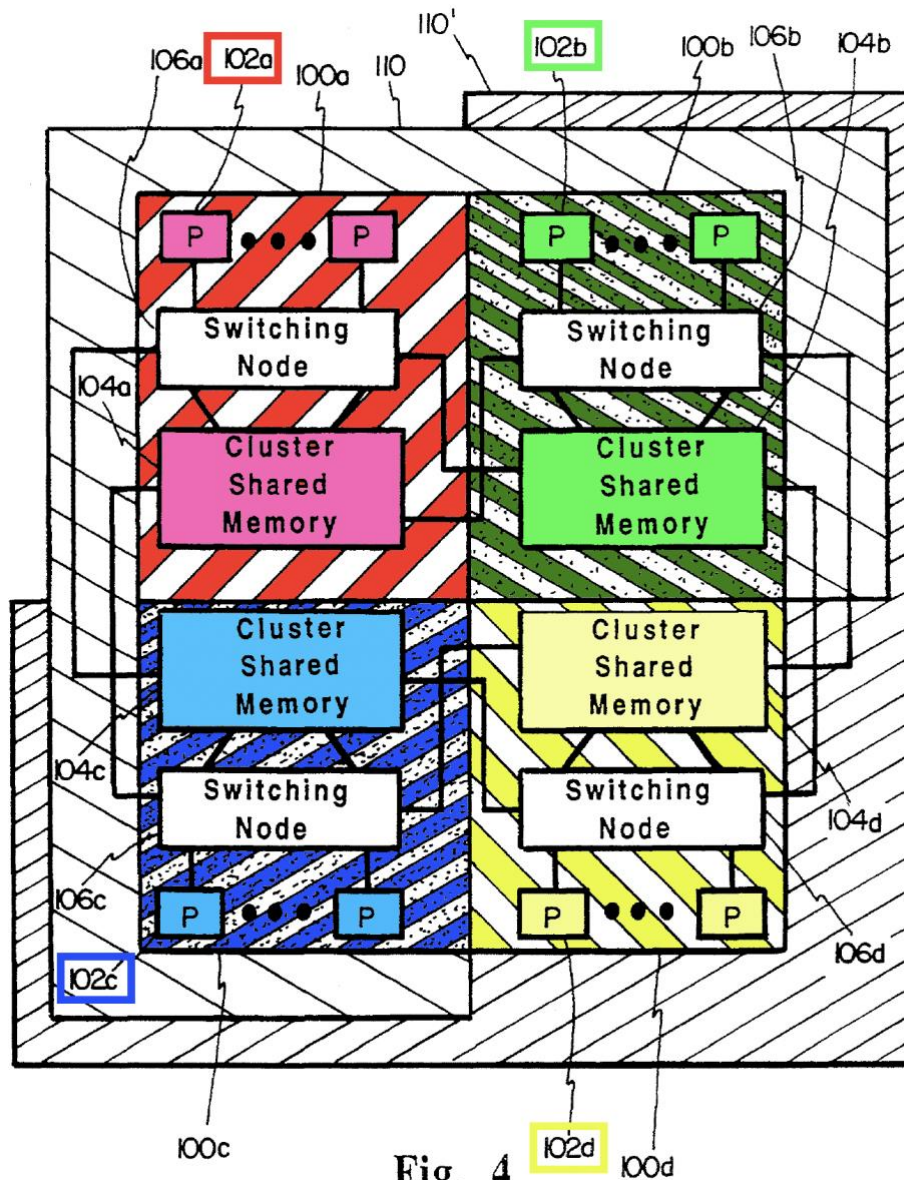


Fig. 4

Mr. Schmandt's impromptu deposition theory as to how the combination of Jiang and Chen would allegedly meet limitations 1(d) and 1(e) is therefore not enabled. Ex. 2020 ¶ 51. Since **every** processor and memory in Chen's Fig. 4 clusters would be used to make probability computations with respect to limitation 1(c), Ex. 2017 at 103:4-19, 73:12-17, **all four memories** 104(a)-(d) would contain necessary information for computing the initial score of limitation 1(d) and for making the determination of whether to compute a final score as recited in limitation 1(e). Ex. 2020 ¶ 51. But since "not all[] the clusters are interconnected," such that no one processor in one cluster can access the information contained in all four of the memories 104(a)-(d), Ex. 1005, 9:10-39; Ex. 2020 ¶ 51, it necessarily follows that the combination of Jiang and Chen could **not** readily use only **one processor in one cluster** to perform limitations 1(d) and 1(e), as Mr. Schmandt's *ad hoc* deposition theory proposed. Ex. 2017 at 102:19-103:3; Ex. 2020 ¶ 51.

Mr. Schmandt's failure to conceive a workable theory as to how the combination of Jiang and Chen would meet limitations 1(d) and 1(e) only begins to illustrate the significant complexities and complications of the proposed combination. Ex. 2020 ¶ 52. It also proves the Petition's failure to establish that "a skilled artisan could have made the claimed invention" by combining Jiang with

Chen in the manner proposed. *Raytheon Techs. Corp. v. GE*, 993 F.3d 1374, 1380 (Fed. Cir. 2021) (emphasis added). A person of *ordinary* skill in the field of speech recognition certainly would not have found it “obvious” to arrive upon the invention of claim 1 given Mr. Schmandt’s failure to do so even with the benefit of hindsight and an improperly elevated level of expertise. Ex. 2020 ¶ 53.

VII. Conclusion

For the reasons above, the Board should not find any challenged claim unpatentable.

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 **CORRECTED PATENT OWNER'S RESPONSE** specified in 37 C.F.R. § 42.24 have 5,353 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.

/Kayvan Noroozi/

Kayvan Noroozi

Counsel for Patent Owner

Date: March 7, 2024

CERTIFICATION OF SERVICE

The undersigned hereby certifies that the foregoing **CORRECTED** **PATENT OWNER'S RESPONSE** was served electronically via e-mail on **March 7, 2024**, on the following counsel of record for Petitioner:

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

Adam M. Sandwell
Adam.Sandwell@eriseip.com

Jennifer C. Bailey
Jennifer.Bailey@eriseip.com

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

/Peter C. Knops/
Peter C. Knops
NOROOZI PC
11601 Wilshire Blvd., Suite 2170
Los Angeles, CA 90025
Tel.: 310-975-7074

Counsel for Patent Owner

Date: **March 7, 2024**