

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 SUR-REPLY

TABLE OF CONTENTS

I.	The Petition’s combination lacks a reasonable expectation of success	1
A.	Patent Owner’s Response presented detailed evidence directed to the <i>Petition’s combination</i> and the abilities of the <i>Petition’s</i> POSA	1
B.	Petitioner’s Reply fails to rebut Patent Owner’s evidence, and instead relies on incorrect premises	5
1.	Petitioner’s false logic 1: “a POSITA is <i>assumed</i> to have requisite knowledge” to enable the Petition’s combination	5
2.	Petitioner’s false logic 2: “a POSITA would have [] had knowledge of parallel processing architectures for speech recognition” and so a POSA would have known how to achieve <i>Petitioner’s combination</i>	11
3.	Petitioner’s false logic 3: “Skilled persons knew of various speech recognition hardware architectures,” and so “would have had knowledge to address ‘trivial’ issues arising in the combination”	14
4.	Petitioner’s false logic 4: “Zentian does not argue that the combination is inoperable, [or] could not be done. . . .”	15
5.	Petitioner’s false logic 5: “Zentian’s alleged complications are specific to a clustered processor architecture and would occur regardless of the type of information being processed.” Paper 21 at 13.	16
6.	Petitioner’s false logic 6: “Zentian’s [] argument that memory collisions would occur and thereby slow down the processing does not account for the combination having <i>multiple</i> processors.” Paper 21 at 15.	19
II.	The evidence proves no motivation to combine	22
A.	Patent Owner’s Response disproved Petitioner’s motivations	22
B.	Petitioner’s attempts to rehabilitate its motivations to combine are ineffective	24

EXHIBIT LIST

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

I. The Petition's combination lacks a reasonable expectation of success

A. Patent Owner's Response presented detailed evidence directed to the *Petition's combination* and the abilities of the *Petition's* POSA

Patent Owner's Response demonstrated through extensive evidence that *the Petition's combination* lacks a reasonable expectation of success. Paper 19 at 6-22. Each aspect of Patent Owner's argument was directed to problems and deficiencies arising out of *the Petition's specific combination*, not the '319 Patent. *Id.*

In particular, Patent Owner's Response proved:

- The Petition's combination requires grouping Thelen's recognizers and then porting the recognizer groups into clusters of processors, with each cluster connected to a dedicated memory, such that each dedicated cluster memory holds *part* of Thelen's lexical data, and all of the cluster memories together contain *all* of Thelen's lexical data, Paper 19 at 6; Pet. at 40;
- The combination must perform in ***real time***, as Thelen did in unmodified form, Ex. 1003 ¶ 84, Ex. 2020 ¶ 42, Ex. 2017 at 154:8-155:17 (Mr. Schmandt testifying that "the task most described in Thelen is dictation" and that if a dictation system had recognition latency, "usability would be pretty low");
- That in the combination, all of Thelen's sub-context models for a given context would be stored into the ***same*** single cluster memory, Ex. 1003 ¶ 83,

Ex. 2017 at 124:13-125:7, Ex. 2020 ¶ 43, thus requiring every processor in a cluster to *simultaneously* request “the particular subset of vocabulary . . . that’s germane to its context” from *a single* shared cluster memory, Ex. 2017 at 125:8-126:5, Ex. 2020 ¶¶ 43-44;

- The clustered processor and memory architecture of the Petition’s combination is not taught in Thelen, but rather is taken from Chen, Paper 19 at 7; Ex. 2017 at 120:8-18;
- Chen does not teach how to implement any speech recognition techniques on its hardware, much less Thelen’s techniques, Ex. 2017 at 120:19-121:2; Ex. 2020 ¶ 30;
- Thelen likewise does not teach how to run its techniques on groups of clustered processors like Chen’s. Ex. 2017 at 136:7-16; Ex. 2020 ¶ 31.
- Mr. Schmandt did *not*:
 - explain *how* a person of *ordinary* skill in speech recognition could implement parallel operation of Thelen’s different recognition models working simultaneously within Chen’s clusters, Ex. 2017 at 137:21-138:24;

- identify any challenges the POSA would face in making his proposed combination, Ex. 2017 at 141:23-142:2, nor even the tasks the POSA would need to undertake, *Id.* at 144:3-20;
 - analyze “the tasks that would be required” of the POSA before he reached his conclusion of a reasonable expectation of success, *Id.* at 144:22-145:3;
 - supervise anyone involved in mapping a speech recognition model to a clustered processor and memory architecture like Chen’s, *Id.* at 145:20-24;
 - build the processor to memory architecture of any speech recognition system identified in his background experience, *Id.* at 34:24-35:5, 32:3-9, 106:10-23.
- Implementing the Petition’s combination of Chen, Thelen, and Bailey would have been highly complex and well beyond the skill level of the POSA as defined in the Petition, Ex. 2020 ¶¶ 17-19, 27, because:
 - The ordinary artisan in the field of speech recognition was a specialist in digital signal processing and speech recognition, not parallel processing architectures or high performance computing, Ex. 2020 ¶ 39;

- The Petition's specific combination would have required avoiding memory conflicts, controlling task sharing, addressing communication bandwidth and latency problems across the hardware, and developing a messaging strategy to coordinate information sharing between and within Chen's clusters, Ex. 2020 ¶ 38, which the Petition's POSA could not reasonably do, Ex. 2020 ¶ 39;
- Placing all of Thelen's sub-context models for a given context into the **same** single cluster memory, Ex. 1003 ¶ 83, Ex. 2017 at 124:13-125:7, Ex. 2020 ¶ 43, and requiring every processor in a cluster to **simultaneously** request "the particular subset of vocabulary . . . that's germane to its context" from **a single** shared cluster memory, Ex. 2017 at 125:8-126:5, Ex. 2020 ¶¶ 43-44, would cause extensive memory collisions that would prevent real time speech recognition, Ex. 2020 ¶ 45;
- It was known in the prior art that adding processors to perform speech recognition in a parallel processing architecture could in fact **slow down** the system's performance, as taught in Mathew I, which reported that a "5 processor version **was slower** than the 2 processor version due to the high synchronization overhead," Ex. 2011 at 11;

- Even Dr. Anderson, who is a person of *extraordinary skill*, along with a team of engineers and a DSP expert from Texas Instruments, could not successfully transition existing signal processing software to a parallel processing architecture after six months. Ex. 2020 ¶¶ 40-41.

B. Petitioner's Reply fails to rebut Patent Owner's evidence, and instead relies on incorrect premises

In the face of the extensive evidence summarized above, Petitioner does not present a supplemental declaration from Mr. Schmandt. Moreover, Petitioner abandoned its deposition of Dr. Anderson after merely one hour, and thus has nothing substantive to introduce from Dr. Anderson's testimony. Ex. 1044.

Instead, Petitioner's Reply presents a series of incorrect conclusions built on inaccurate or inapplicable premises. Each of Petitioner's logically false arguments is addressed and refuted in turn below.

1. Petitioner's false logic 1: "a POSITA is *assumed* to have requisite knowledge" to enable the Petition's combination

Petitioner attempts to conflate *the Petition's combination* with the embodiments and claims of the '319 Patent. Petitioner argues that if the *patent* did not discuss the problems with the *Petition's combination*, it must be assumed that the POSA would have known how to overcome the problems with the *Petition's combination*. Paper 21 at 2-5. That logic is false.

“Skill in the art does **not** act as a bridge over gaps in substantive presentation of an obviousness case, but instead supplies an important guarantee of **objectivity** in the process.” *Al-Site Corp. v. VSI Int'l, Inc.*, 174 F.3d 1308, 1324 (Fed. Cir. 1999) (emphasis added). “Rarely, [] will the skill in the art component operate to supply missing knowledge or prior art to reach an obviousness judgment.” *Id.* Rather, “[t]o imbue one of ordinary skill in the art with knowledge of the invention in suit, when no prior art reference or references of record convey or suggest that knowledge, is to fall victim to the insidious effect of a hindsight syndrome wherein that which only the inventor taught is used against its teacher.” *Id.*

Here, the Petition’s combination begins with Thelen, Bailey, and Chen, and proposes to combine those references in particular ways that none of the references enables. Section I.A, *supra*. Petitioner has failed to demonstrate through actual evidence that the POSA, as defined in the Petition, would have the necessary skill or knowledge to in fact achieve *the Petition’s combination* with a reasonable expectation of success. Section I.A, *supra*. Patent Owner has demonstrated through detailed evidence that the POSA could not. Section I.A, *supra*.

Petitioner’s assertion that the POSA should be “**assumed to have requisite knowledge**” to achieve the Petition’s contorted and highly complex combination, Paper 21 at 2, is the essence of hindsight bias. Petitioner cannot weaponize the

level of *ordinary* skill in the art, which is a safeguard against hindsight bias, to fill substantive gaps in its obviousness case. *Al-Site*, 174 F.3d at 1324.

Petitioner's premise is that if the '319 *Patent* did not teach how to solve the specific problems arising out of *Petitioner's combination*, then the Board must conclude that the POSA would know how to solve *Petitioner's problems*. Paper 21 at 2-5. That reasoning is false: the '319 Patent does not teach placing all of Thelen's sub-context models for a given context into the **same** single cluster memory, Ex. 1003 ¶ 83, Ex. 2017 at 124:13-125:7, Ex. 2020 ¶ 43, thus requiring every processor in a cluster to **simultaneously** request "the particular subset of vocabulary . . . that's germane to its context" from **a single** shared cluster memory, Ex. 2017 at 125:8-126:5, Ex. 2020 ¶¶ 43-44, causing memory collisions that would prevent Petitioner's goal (and Thelen's teaching) of real time speech recognition. Ex. 2020 ¶ 45. The '319 Patent thus has no reason to teach solutions to problems arising out of *Petitioner's combination*. Paper 21 at 4. The patent's silence on those issues is certainly not evidence that the solutions were within the skill level of the Petition's POSA. The consequence of the Petition's absurd combination theory lies with Petitioner, not Patent Owner.

Petitioner's case law citations do not establish otherwise. In *Uber Techs., Inc. v. X One, Inc.*, 957 F.3d 1334, 1339 (Fed. Cir. 2020), the issue was whether a

POSA would have been motivated to select between one of only two known, finite approaches. Because the challenged patent was directed to *both* solutions, the Federal Circuit held that the appropriate choice was within the level of skill of the POSA. *Id.* Here, the issue is expectation of success, not motivation to combine, there are not finite choices, and the ’319 Patent does not teach Petitioner’s combination.

In *In re Publicover*, 813 F. App’x 527, 531-32 (Fed. Cir. 2020), the prior art and the challenged patent **both** taught identification of vestibulo-ocular movement, a type of eye movement. Patent Owner presented only “attorney argument,” not evidence, that the prior art’s disclosures failed to adequately explain how to identify that type of eye movement. *Id.* The Federal Circuit rejected that “attorney argument” because the challenged patent likewise **taught** identifying that type of eye movement **but did not** provide a detailed teaching of **how** to do so. *Id.* Here, neither Thelen, Chen, nor Bailey teaches implementing Thelen’s recognizers on a clustered parallel processing, shared cluster-memory architecture—much less in the manner Petitioner has proposed. Ex. 2017 at 120:19-121:2, 136:7-16; Ex. 2020 ¶¶ 30-31. Accordingly, the prior art does not and cannot carry any presumption of enablement **as to Petitioner’s combination**. *In re Antor Media Corp.*, 689 F.3d 1282, 1287-88 (Fed. Cir. 2012) (prior art only presumed enabling for what it

teaches). The ’319 Patent **also** does not teach implementing a speech recognition system like Thelen’s on a clustered processor architecture, much less a system in which multiple sub-context language models must be simultaneously accessed by multiple processors from a single shared cluster memory, as Petitioner’s combination requires. Thus, the ’319 Patent provides no assumption that Petitioner’s combination was within the POSA’s skill. Moreover, Patent Owner’s demonstration of a lack of a reasonable expectation of success in this proceeding is based on extensive evidence, not “attorney argument.” *Publicover* is thus inapposite.

In re Epstein, 32 F.3d 1559, 1568 (Fed. Cir. 1994), fails to support Petitioner for the same reasons. There, the Federal Circuit simply held that “one of skill in the art would have known how to implement ***the features of the references***” as taught in each reference individually. *Id.* (emphasis added). Here, Petitioner’s combination does not simply require practicing what Thelen, Bailey, or Chen taught individually—it requires a combination that none of those references taught, and that a POSA would not have been able to achieve. Ex. 2020 ¶¶ 39, 48.

Petitioner’s reliance on the Board’s decisions in *Walmart* and *Samsung* is misplaced because those decisions do not carry legal authority. Nonetheless, both decisions also fail to support Petitioner’s argument.

In *Walmart*, the Board relied on the challenged patent's teaching that "a plurality of light panels 40 may be disposed end to end," and the challenged patent's silence as to how to address challenges that would arise "***when disposing the panels in electrical series.***" IPR2022-00534, Paper 50 at 32 (emphasis added). By contrast, the problems Patent Owner has identified with Petitioner's combination do ***not*** arise out of practicing what the '319 patent itself teaches; they arise out of ***Petitioner's combination design.*** The Board's decision in *Walmart* is thus inapposite.

In *Samsung*, the prior art disclosed wireless communication, and thus was entitled to a presumption of enablement ***for its own teachings.*** IPR2022-00324, Paper 33 at 36-37. Moreover, the challenged patent itself disclosed the ***same*** wireless communication techniques as the prior art without implementation details. By contrast, the problems with Petitioner's combination here do not arise out of the individual teachings of any reference, but rather out a proposed combination that is neither disclosed by nor enabled by any of the references. Section I.A, *supra*. There is thus no presumption of enablement with respect to Petitioner's combination. Nor does the '319 patent disclose Petitioner's combination. Accordingly, the '319 Patent's lack of implementation details for Petitioner's

specific and contorted combination is not evidence that Petitioner's combination was within the skill of the POSA.

Finally, the problems Patent Owner has identified with respect to Petitioner's combination are not "generic," as Petitioner contends. Paper 21 at 5. They are highly detailed and specific to Petitioner's combination, and supported by the testimony of *both* experts. Section I.A, *supra*.

2. Petitioner's false logic 2: "a POSITA would have [] had knowledge of parallel processing architectures for speech recognition" and so a POSA would have known how to achieve *Petitioner's combination*

Petitioner argues that because Thelen and Bailey allegedly taught "parallel processing architectures," and because the POSA would allegedly have been "knowledgeable of Chen," it follows that the POSA would have known how to achieve *Petitioner's specific combination* of Chen, Thelen, and Bailey. Paper 21 at 9-10.

That logic is false. Neither Chen, Thelen, or Bailey teach or suggest Petitioner's combination, which requires:

1. grouping Thelen's recognizers and porting each recognizer group into clusters of processors, with each cluster connected to a dedicated memory, such that each dedicated cluster memory is a lexical memory holding only *part* of

Thelen's lexical data, and all of the cluster memories together contain *all* of Thelen's lexical data. Paper 19 at 6; Pet. at 40; *and*

2. storing all of Thelen's sub-context models for a given context into the *same* single cluster memory, Ex. 1003 ¶ 83, Ex. 2017 at 124:13-125:7, Ex. 2020 ¶ 43, thus requiring every processor in a cluster to *simultaneously* request "the particular subset of vocabulary . . . that's germane to its context" from *a single* shared cluster memory, Ex. 2017 at 125:8-126:5, Ex. 2020 ¶¶ 43-44, while performing speech recognition *in real time*. Ex. 1003 ¶ 84, Ex. 2020 ¶ 42.

Accordingly, neither the disclosures of Chen, Thelen, and Bailey, nor the POSA's alleged (unspecified) general knowledge of parallel processing architectures would have enabled the POSA to achieve Petitioner's combination. Section I.A, *supra*.

That reality is proven through Dr. Anderson's own experience, who explains that despite his *extraordinary* level of skill, and the help of multiple other engineers, he could not successfully transition existing signal processing software to a parallel processing architecture after many months. Ex. 2020 ¶¶ 40-41.

Petitioner argues that Dr. Anderson's testimony merely "impugns Dr. Anderson's credibility," but "does not prove a POSITA would lack the knowledge to reasonably succeed in the combination." Paper 21 at 8. That is incorrect. Dr.

Anderson's testimony demonstrates that even a person of *extraordinary skill* in the art "could not necessarily expect success with respect to far simpler combinations than the one the Petition has proposed," Ex. 2020 ¶ 41, and thus supports the credibility of Dr. Anderson's testimony that the Petition's combination was beyond the level of *ordinary skill*. Ex. 2020 ¶¶ 38, 48.

By contrast, Mr. Schmandt's lack of experience in building the processor to memory architecture of any speech recognition system, Ex. 2017 at 34:24-35:5, 32:3-9, 106:10-23, or supervising anyone involved in mapping a speech recognition model to a clustered memory and processor architecture, *id.* at 145:20-24, greatly undermines his opinions regarding a reasonable expectation of success. And Mr. Schmandt's opinion is further discredited by his testimony that his declaration does not explain *how* a POSA could have achieved his combination, Ex. 2017 at 137:21-138:24, does not identify any challenges the POSA would face, Ex. 2017 at 141:23-142:2, nor even the tasks the POSA would need to undertake, *Id.* at 144:3-20, which Mr. Schmandt himself did not analyze before reaching his conclusion of a reasonable expectation of success. *Id.* at 144:22-145:3.

Petitioner argues that Mr. Schmandt's testimony is *admissible* despite those flaws, Paper 21 at 7-8, but the issue here is not admissibility but rather credibility. Mr. Schmandt's conclusory assertions of a reasonable expectation of success

should be given no weight in view of the evidence of record, including his own admissions.

3. Petitioner's false logic 3: "Skilled persons knew of various speech recognition hardware architectures," and so "would have had knowledge to address 'trivial' issues arising in the combination"

Petitioner reasons that because a POSA knew of "various speech recognition hardware architectures," the POSA could have solved what Petitioner calls the "trivial issues arising in the combination." Paper 21 at 10-11.

Petitioner provides no evidence, however, that a POSA would have been aware of the speech recognition architecture of *Petitioner's combination*. To the contrary, the evidence is uncontroverted that Petitioner's combination appears nowhere in the prior art, but is instead a novel fabrication concocted from multiple modifications to the teachings of three different references in a manner not taught or enabled by any reference. Ex. 2017 at 120:19-121:2, 136:7-16; Ex. 2020 ¶¶ 31-32, 42-45.

Prior art is only presumptively enabled ***for what it teaches***. *In re Antor Media Corp.*, 689 F.3d 1282, 1287-88 (Fed. Cir. 2012). Mr. Schmandt has admitted that neither Chen, Thelen, nor Bailey teaches *Petitioner's combination*, and thus none of those references is entitled to a presumption of enablement with

respect to *Petitioner's combination*, as Petitioner contends. *Compare* Paper 21 at 12-13 with Ex. 2017 at 120:8-121:2, 136:7-16; Ex. 2020 ¶¶ 31-32, 42-45; *Raytheon Techs. Corp. v. GE*, 993 F.3d 1374, 1380-81 (Fed. Cir. 2021). Even assuming Chen teaches an “operable clustered processor architecture,” Paper 21 at 12-13, Chen does **not** teach using its architecture to run Thelen’s recognizers (much less in the manner the Petition has proposed), or to run *any* speech recognition. Ex. 2017 at 120:19-121:2; Ex. 2020 ¶ 31. Petitioner simply wishes to assume that Chen’s clustered processor and memory architecture was automatically enabled to run ***any speech recognition architecture imaginable***, regardless of the implementation requirements of that architecture. Paper 21 at 11-13. But Dr. Anderson’s testimony refutes Petitioner’s wishful assumptions through detailed evidence that the POSA could not successfully implement *Petitioner’s combination* using Chen’s clustered processors and memories. Ex. 2020 ¶¶ 30-48.

4. Petitioner’s false logic 4: “Zentian does not argue that the combination is inoperable, [or] could not be done. . . .”

Petitioner appears to suggest that Patent Owner must prove that Petitioner’s combination could never be achieved even with unlimited knowledge, skill, and resources. Paper 21 at 13. That is false. Patent Owner need only demonstrate that *Petitioner has not proven* that a person of *ordinary* skill in the art at the time of the

invention would have reasonably expected to succeed in combining the prior art elements in the way *Petitioner has proposed*. *KEYnetik, Inc. v. Samsung Elecs. Co.*, 841 F. App'x 219, 227 (Fed. Cir. 2021); *Arctic Cat Inc. v. Bombardier Rec. Prods.*, 876 F.3d 1350, 1360-61 (Fed. Cir. 2017); *In re Cyclobenzaprine Hydrochloride Extended-Release Capsule Patent Litig.*, 676 F.3d 1063, 1069 (Fed. Cir. 2012). Patent Owner has met and exceeded that requirement. Section I.A, *supra*. Petitioner's arguments fundamentally ignore the crucial limitation that the POSA is merely one of *ordinary* skill, not *extraordinary* skill.

5. Petitioner's false logic 5: "Zentian's alleged complications are specific to a clustered processor architecture and would occur regardless of the type of information being processed." Paper 21 at 13.

Petitioner contends that the problems arising out of its combination would occur in *any* clustered processing architecture. Paper 21 at 13. But Petitioner has ***no evidence*** in support of that assertion. *Id.* Petitioner's assertion has no basis.

Petitioner next attempts to flip its burden onto Patent Owner, arguing that "Zentian never explains why [memory collision] problems would be present in the combination but not" in the system of Claim 46. Paper 21 at 14. That argument also fails. Petitioner cannot shift its burden of proof to Patent Owner. Nonetheless, Patent Owner has demonstrated in detail that Petitioner's memory collision

problems arise out of the requirements Petitioner and its expert have identified for Petitioner's combination, as well as the teachings of the prior art references they have relied upon, Ex. 2020 ¶¶ 43-48, which are not common to the '319 Patent. In particular, Petitioner's theory requires a real time system in which all of Thelen's sub-context models for the **same** context are stored into the **same** single cluster memory, Ex. 1003 ¶ 83, Ex. 2017 at 124:13-125:7, Ex. 2020 ¶ 43, causing every processor in a cluster to **simultaneously** request "the particular subset of vocabulary . . . that's germane to its context" from **that single** shared cluster memory, Ex. 2017 at 125:8-126:5, Ex. 2020 ¶¶ 43-44, which in turn would cause memory collisions that would prevent even near real time operation. Ex. 2020 ¶ 45. The '319 Patent does not have such a design or trigger such problems.

Petitioner's argument that "the claims do not recite real-time recognition" is equally misguided. Paper 12 at 14. The question of a reasonable expectation of success *for the combination* must be assessed based on the details Petitioner has proposed for the *combination*, not the words of the challenged claims. *Arctic Cat Inc. v. Bombardier Rec. Prods.*, 876 F.3d 1350, 1360-61 (Fed. Cir. 2017) (requiring proof that "an ordinarily skilled artisan would [] believe that a **particular combination** would have a reasonable expectation of 'anticipated success') (emphasis added). Otherwise, the reasonable expectation of success

inquiry would collapse into the question of whether the claims are enabled, which is entirely wrong. Here, Petitioner's theory requires real time recognition, and is in fact motivated by alleged speed *improvements* to Thelen's unmodified system, which already performed in real time. Ex. 1003 ¶ 84; Ex. 2020 ¶ 42; Ex. 2017 at 154:8-155:6 (Mr. Schmandt testifying that "the task most described in Thelen is dictation" and stating that if a dictation system had recognition latency "usability would be pretty low"); *id.* at 155:7-17 (admitting unmodified Thelen had no issues operating in real time). Accordingly, the POSA must have a reasonable expectation of success with respect to that combination.

Petitioner likewise falsely asserts that Patent Owner's arguments "assume[] that a POSITA would have thoughtlessly appended processors to the system[.]" Paper 21 at 14. Patent Owner did not assume the POSA to be thoughtless at all. Rather, it was Petitioner and its expert that thoughtlessly proposed a fanciful and contrived combination of elements without so much as analyzing whether the person of *ordinary* skill in speech recognition could have reasonably achieved the combination. Ex. 2017 at 137:21-138:24, 141:23-142:2 (Schmandt admitting his declaration does not explain *how* the POSA could implement his combination, nor any challenges the POSA would face, nor even the tasks the POSA would need to undertake); *id.* at 144-22-145:3 (Schmandt admitting he did not analyze "the tasks

that would be required” of the POSA before concluding a reasonable expectation of success).

6. Petitioner’s false logic 6: “Zentian’s [] argument that memory collisions would occur and thereby slow down the processing does not account for the combination having *multiple* processors.” Paper 21 at 15.

Petitioner argues that Patent Owner’s evidence does not account for the fact that the combination would include “multiple processors” from Chen. Paper 21 at 15. That assertion is entirely false. Dr. Anderson’s declaration expressly states that in Petitioner’s combination, “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 ***at the same time***, causing a problem known in the art as ‘memory collisions.’” Ex. 2020 ¶ 45. Dr. Anderson explains that it is the very fact of multiple processors deployed in the manner Petitioner’s combination requires that would cause Petitioner’s system to fall behind real time, Ex. 2020 ¶¶ 44-45, and to slow down compared to Thelen in its unmodified form. *Id.* at ¶¶ 59-60.

Petitioner asserts that Chen’s multiple processors “would more than account for any memory contention issues.” Paper 21 at 15. That is a quantitative assertion for which Petitioner has no evidence. Indeed, Petitioner cites no evidence

demonstrating that its combination of Chen, Thelen, and Bailey would overcome memory contention problems and operate at real time. Dr. Anderson's testimony on that point was neither rebutted through a supplemental declaration from Mr. Schmandt, nor even explored during Dr. Anderson's deposition. Instead, Petitioner cites to teachings from Chen, which do not address (and did not contemplate) the Petition's combination with Thelen and the ensuing memory collisions that would arise from Petitioner's combination design. As Petitioner itself states, "Chen is enabled **for what it teaches.**" Paper 21 at 16. Chen does not teach how to address memory contention problems arising from storing all of Thelen's sub-context models for a given context into the **same** single cluster memory and requiring every processor in a cluster to **simultaneously** request lexical data from that single shared cluster memory. Thus, Petitioner cannot rely on Chen to solve problems arising out of a speech recognition system design that Chen did not teach.

For the same reason, Petitioner wrongly argues that if Chen teaches it did not "have to sacrifice performance" in one context, it follows that the *Petition's combination* would not sacrifice performance in *any* context. Paper 21 at 16. Once again, Petitioner relies on attorney argument. Dr. Anderson, by contrast, has explained in detail why Petitioner's combination theory would cause Petitioner's system to slow down, not speed up. Ex. 2020 ¶¶ 44-45, 59-60. And Dr. Anderson

further demonstrated that it was known 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,” as taught in Mathew I. Ex. 2020 ¶ 46. In particular, Mathew I disclosed that a five-processor parallel processing speech recognition system “was slower than the 2 processor version due the high synchronization overhead.” Ex. 2002 ¶¶ 53-55, Ex. 2021 at 11.

Although Petitioner itself relied upon Mathew I in related proceedings, Petitioner now attempts to discredit Mathew I. Paper 21 at 17. Petitioner states that Mathew I discussed using multiple processors for the HMM stage only, whereas the combination would use multiple processors “for all three stages of the speech recognition.” *Id.* Petitioner fails to explain, however, why that distinction makes a difference, and Petitioner has **no evidence** to support such a conclusion. Nor is there any reason to conclude that providing even more tasks to the combination system’s processors would make them operate more quickly. Petitioner’s attempt at elevating Chen above Mathew I’s is likewise baseless. Mathew I reports the results from **actually implementing and testing** a speech recognition system performed in parallel across two processors and then across five processors. Chen has nothing to say about speech recognition, Ex. 2017 at 120:19-121:2, much less a

comparative analysis between the speeds achieved from running speech recognition in parallel across more or fewer processors.

Finally, Petitioner's arguments rest on the false premise that the combination would have increased processing power. To the contrary, Petitioner's expert initially testified that the combination system would have the **same** number of processors as in Thelen, Ex. 2017 at 130:23-131:4, and ultimately stated that he did not know whether it would or not. *Id.* at 136:3-6.

II. The evidence proves no motivation to combine

A. Patent Owner's Response disproved Petitioner's motivations

Patent Owner's Response refuted each of the Petition's motivation through the testimony of both experts. Paper 19 at 22-31.

With respect to increasing processing power, Mr. Schmandt admitted at his deposition that his combination would **not** add processing power beyond what Thelen already had, or else that he did not know if it would. Ex. 2017 at 130:23-131:4, 136:3-6.

As for processing speed, Mathew I establishes that the POSA would have known at the time that adding processors to a parallel processing speech recognition system would not have necessarily improved recognition speed. Ex. 2021 at 11; Ex. 2020 ¶¶ 51-53. To the contrary, Mathew I taught that a five-

processor system was slower than a two-processor system due to high synchronization overhead. *Id.* Moreover, Thelen in unmodified form was a real time speech recognition system, and Mr. Schmandt admitted that Thelen had no problems achieving real time recognition. Ex. 2017 at 155:7-17. As discussed, however, Mr. Schmandt's combination of Thelen with Chen and Bailey would cause extensive memory collisions that would cause the system to fall far behind real time operation, and thus ***slow down*** the system compared to Thelen in unmodified form. Ex. 2020 ¶¶ 59-60. As Dr. Anderson testified, that fact alone would cause the POSA to ***not*** have been motivated to combine Thelen with Chen and Bailey in the manner the Petition has proposed. *Id.* at ¶¶ 60, 62.

Regarding the cost implications of the combination, Mr. Schmandt further admitted that he did not undertake any qualitative or quantitative cost-benefit analysis in reaching his opinions regarding a motivation to combine. Ex. 2017 at 42:14-19, 63:8-15.

As for “flexibility and scalability,” Mr. Schmandt's deposition and Dr. Anderson's testimony established that there is no such advantage to the combination compared to Thelen alone. Ex. 2017 at 105:8-106:4; Ex. 2020 ¶ 61.

In addition, Dr. Anderson demonstrated that Thelen's shared memory architecture would have been a poor fit for Chen's multiple private memories, and

would have required significant reprogramming effort that would have caused the POSA to not form a motivation to combine. Ex. 2020 ¶¶ 56-58.

B. Petitioner's attempts to rehabilitate its motivations to combine are ineffective

Petitioner argues extensively that its combination system would have increased processing power and speed, and that a POSA would have been motivated accordingly. Paper 21 at 13-19, 25. Those assertions are false and refuted by the evidence. Section I.B.5-6, *supra*. Surprisingly, Petitioner later concedes the point, stating that “increasing processing power via adding more processors beyond what is taught in Thelen” is not a motivation. Paper 21 at 24.

Petitioner also alleges, in passing, that Mr. Schmandt's declaration testimony “that Chen's clustered processors utilizing commercially-available processors strikes a balance between cost and processing power” is “sufficient to support a motivation of the combination.” Paper 21 at 18. The evidence, however, proves that the combination system would operate more slowly than Thelen's, would not meet Thelen's realtime recognition purpose or the Petition's realtime recognition requirement, would entail significant reprogramming efforts that the POSA would not desire to undertake, and would not necessarily have any more

processing power than Thelen alone. Section I.A, *supra*. The evidence thus weighs against a motivation to combine regardless of Petitioner's "cost" motivation.

Moreover, Mr. Schmandt admitted he did not compare the cost of Chen's architecture to any other clustered processing architectures, Ex. 2017 at 63:8-15, nor did he perform a cost-benefit analysis that would allegedly motivate the combination. *Id.* at 42:14-19. Thus, there is no evidence to support a cost motivation. Indeed, Petitioner admits it cannot state what its combination would have been ***cheaper than***, or what costs would have been "reduced." Paper 21 at 23, 25. Petitioner's circular and conclusory allegations cannot establish obviousness. *ActiveVideo Networks, Inc. v. Verizon Communs., Inc.*, 694 F.3d 1312, 1327-28 (Fed. Cir. 2012) (rejecting evidence that "bears no relation to any specific combination of prior art elements").

Petitioner next resurrects its "flexibility and scalability" theory in passing, Paper 21 at 23, but does not even attempt to address Patent Owner and Dr. Anderson's refutation of that alleged motivation. Paper 21 at 23.

Petitioner argues that Chen's teachings are intended to solve challenges from distributed memory models. Paper 21 at 20. But Dr. Anderson acknowledged Chen's solution, which requires a complex process, and which there is no evidence the POSA would have been motivated to undertake. Ex. 2020 ¶¶ 56-58.

Petitioner concedes that Chen teaches complications from using distributed memory models, but contends the Petition's combination does not use a distributed memory model because its combination allegedly does not give memory access across adjacent clusters. Paper 21 at 21-22. Petitioner's assertions are refuted by its own Petition. The Petition relies on the embodiment of Chen's Fig. 4. Pet. at 38-39. Chen teaches that in the embodiment of Fig. 4, "the processors 102a in cluster **100a** . . . can **adjacently access** the cluster shared memory 104b and 104c of clusters **100b and 100c**, respectively." Ex. 1005 at 9:24-28 (emphasis added). Thus, the Petition's combination system employs a distributed memory model according to Petitioner's own terminology. Paper 21 at 21.

Petitioner further argues that "the clustered processor architecture provides increased benefits **over a single-processor architecture**." Paper 21 at 20. That assertion is both unproven, and entirely irrelevant. Petitioner admits that unmodified Thelen is not a single-processor architecture. *See, e.g.*, Paper 21 at 10 (alleging that Thelen is a "parallel processing architecture"). Thus, there is no basis for comparing the combination to a single-processor architecture.

Finally, Petitioner contends that the POSA would have been motivated to pursue the combination to "reduc[e] the duplication of lexical data." But Petitioner's own Reply states that its combination would **require duplication of**

lexical data. Paper 21 at 21-22 (stating that “a core vocabulary [would be] **replicated** in each local memory”) (emphasis added). In any case, the evidence establishes that the POSA would know that the combination system would be slower than Thelen, would fail to operate at real time, would require significant reprogramming effort, would cause extensive memory collisions, and would ultimately perform **worse**, not better, than Thelen alone. Ex. 2020 ¶¶ 49-62. There is no motivation to combine on this record.

Respectfully submitted,

/s/ Kayvan B. Noroozi

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: February 20, 2024

CERTIFICATION OF WORD COUNT

The undersigned hereby certifies that the portions of the above-captioned **PATENT OWNER'S SUR-REPLY** specified in 37 C.F.R. § 42.24 have 5,586 words, in compliance with the 5,600 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: February 20, 2024

CERTIFICATION OF SERVICE

The undersigned hereby certifies that the foregoing **PATENT OWNER'S SUR-REPLY** was served electronically via e-mail on February 20, 2024, 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: February 20, 2024