

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

KRISP TECHNOLOGIES, INC.,
Petitioner,

v.

SANAS.AI INC.,
Patent Owner.

Case No. IPR2026-00272
Patent No. 11,948,550

PATENT OWNER'S PRELIMINARY RESPONSE

TABLE OF CONTENTS

I. INTRODUCTION.....	1
II. THE '550 PATENT.....	6
III. CLAIM CONSTRUCTION	11
IV. THE PETITION FAILS TO DEMONSTRATE MOTIVATION, REASONABLE EXPECTATION OF SUCCESS, AND PARTICULARIZED MAPPING OF CLAIM ELEMENTS.....	15
A. The Petition Does Not Demonstrate a Reasoned Motivation to Combine Prior Art References	18
B. The Petition Does Not Demonstrate a Reasonable Expectation of Success	23
C. The Petition Does Not Make a Particularized Showing of Where Each Claim Element is Disclosed in the Asserted Prior Art.....	25
D. Institution Should Be Denied as an Inefficient Use of Office Resources.....	27
V. PETITIONER’S CHALLENGES FAIL ON THE MERITS	30
A. Ground 1 Fails on the Merits.....	31
1. Ground 1 Independently Fails Because Sonsaat is Not Analogous Art.....	32
2. The Petition Fails to Establish that Loots’ Decision Tree Can Be Inserted into Fan’s Pipeline Without Disrupting the Model	34
3. Petitioner Fails to Show an Operable Combination of Loots’ Symbolic P2P Decision Trees with Fan’s Neural PPG-to-Mel Pipeline.....	37
4. Petitioner Fails to Show That Sonsaat’s Forced Alignment is Compatible with Fan’s Architecture.....	40
5. Ground 1 Dependent-Claim Challenges Do Not Cure the Independent-Claim Defects	43
B. Grounds 2-4 Fail on the Merits	44
1. Grounds 2-4 Fail Because They Rely on Ground 1.....	44

2.	Ground 4 Fails Because Neither Fan Alone nor Dalce in View of Fan Teaches the Missing User Input Selection Limitation ..	45
3.	Ground 3 Fails Because Osuna’s Accent Conversion Pipeline Is Architecturally Incompatible with Fan’s Seq2seq Neural Network.....	49
4.	Ground 2 Fails Because Feinauer’s Real-Time Architecture is Fundamentally Different from the Architecture of Fan.....	51
VI. CONCLUSION.....		53

TABLE OF AUTHORITIES

	Page(s)
CASES	
<i>ActiveVideo Networks, Inc. v. Verizon Commc'ns, Inc.</i> , 694 F.3d 1312 (Fed. Cir. 2012)	19
<i>Apple Inc. v. Ginko LLC</i> IPR2025-01388, Paper 14 (March 2, 2026).....	27
<i>In re Bigio</i> , 381 F.3d 1320 (Fed. Cir. 2004)	32
<i>Chevron Oronite Co. v. Infineum USA</i> , IPR2018-00923, Paper 9 (November 7, 2018)	28
<i>CyberSecure IPS, LLC v. Network Integrity Systems, Inc.</i> , IPR2025-01441, Paper 18 (May 26, 2026).....	27
<i>Honeywell Int'l Inc. v. 3G Licensing, S.A.</i> , 124 F.4th 1345 (Fed. Cir. 2025)	23
<i>Intelligent Bio-Sys., Inc. v. Illumina Cambridge Ltd.</i> , 821 F.3d 1359 (Fed. Cir. 2016)	23
<i>In re Kahn</i> , 441 F.3d 977 (Fed. Cir. 2006)	19
<i>Johns Manville Corp. v. Knauf Insulation, Inc.</i> , IPR2018-00827, Paper 9 (PTAB Oct. 16, 2018)	20
<i>KSR Int'l Co. v. Teleflex Inc.</i> , 550 U.S. 398 (2007).....	15,19
<i>PAR Pharm., Inc. v. TWI Pharms., Inc.</i> , 773 F.3d 1186 (Fed. Cir. 2014)	23
<i>Phillips v. AWH Corp.</i> , 415 F.3d 1303 (Fed. Cir. 2005) (en banc)	12
<i>Xerox Corp. v. Bytemark, Inc.</i> , IPR2022-00624, Paper 9 (PTAB Aug. 24, 2022).....	19

STATUTES

35 U.S.C. § 103	1
35 U.S.C. § 312(a)(3).....	10, 25
37 C.F.R. § 42.104(b)(4).....	25
37 C.F.R. § 42.107	1

PATENT OWNER’S EXHIBIT LIST

<u>EXHIBIT NO.</u>	<u>DESCRIPTION</u>
Ex2001	Complaint filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Jul. 7, 2025) (D.I. 1) (“Complaint”)
Ex2002	Answer filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Sep. 8, 2025) (D.I. 28) (“Answer and Counterclaims”)
Ex2003	Amended Complaint filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Sep. 22, 2025) (D.I. 35) (“Amended Complaint”)
Ex2004	Order on First Motion to Dismiss filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Sep. 29, 2025) (D.I. 37) (“Order on First Motion to Dismiss”)
Ex2005	Order Denying Krisp Motion to Dismiss First Amended Complaint and Granting in Part and Denying in Part Motion to Dismiss Defendant Krisp’s Counterclaims filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Dec. 1, 2025) (D.I. 58) (“Order Denying Krisp Motion to Dismiss First Amended Complaint”)
Ex2006	Answer to Amended Complaint filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Dec. 15, 2025) (D.I. 60) (“Answer to Amended Complaint”)

<u>EXHIBIT NO.</u>	<u>DESCRIPTION</u>
Ex2007	Order Denying Motion to Dismiss filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Dec. 19, 2025) (D.I. 63) (“Order Denying Motion to Dismiss”)
Ex2008	Order Denying [Renewed] Motion for Judgment on Pleadings filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Feb. 23, 2026) (D.I. 72) (“Order Denying [Renewed] Motion for Judgment on Pleadings”)
Ex2009	Joint Case Management Statement filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Mar. 19, 2026) (D.I. 83) (“Joint Case Management Statement”)
Ex2010	Case Management Scheduling Order filed in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal. Mar. 26, 2026) (D.I. 92) (“Case Management Scheduling Order”)
Ex2011	Docket Sheet in <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i> , No. 3:25-cv-05666 (N.D. Cal.) (“Docket Sheet”)
Ex2012	Coke Morgan Stewart, <i>Memorandum re Interim Processes for PTAB Workload Management</i> (Mar. 26, 2025), https://www.uspto.gov/sites/default/files/documents/InterimProcesses-PTABWorkloadMgmt-20250326.pdf . (“2025 Memo.”)

<u>EXHIBIT NO.</u>	<u>DESCRIPTION</u>
Ex2013	John A. Squires, <i>Memorandum re Additional Discretionary Institution Considerations—U.S. Manufacturing and Small Business Use of AIA Proceedings</i> (Mar. 11, 2026), https://www.uspto.gov/sites/default/files/documents/Additional_Discretionary_Institution_Considerations_US_Manufacturing_and_Small_Business_Use_of_AIA_Proceedings.pdf . (“2026 Memo.”)
Ex2014	<i>Magnolia Med. Techs., Inc. v. Kurin, Inc.</i> , No. IPR2026-00097, Paper 17 (PTAB May 14, 2026) (precedential) (“ <i>Magnolia</i> ”)
Ex2015	Scott R. Boalick, <i>Guidance on USPTO's Rescission of "Interim Procedure for Discretionary Denials in AIA Post-Grant Proceedings with Parallel District Court Litigation"</i> (Mar. 24, 2025)
Ex2016	[Intentionally Left Blank]
Ex2017	Sanas, LinkedIn, https://www.linkedin.com/company/sanastech/about/ (last visited May 13, 2026)
Ex2018	U.S. Patent No. 12,205,609 B1 (issued Jan. 21, 2025)
Ex2019	U.S. Patent No. 12,223,979 B1 (issued Feb. 11, 2025)
Ex2020	U.S. Publication No. 2022/0358903 A1 (published Nov. 10, 2022)

<u>EXHIBIT NO.</u>	<u>DESCRIPTION</u>
Ex2021	U.S. Publication No. 2024/0347070 A1 (published Oct. 17, 2024)
Ex2022	U.S. Publication No. 2024/0363135 A1 (published Oct. 31, 2024)
Ex2023	U.S. Patent No. 12,125,496 (issued Oct. 22, 2024)
Ex2024	<i>About Sanas</i> , Sanas, https://www.sanas.ai/about-us (last visited May 16, 2026).
Ex2025	Krisp, LinkedIn, https://www.linkedin.com/company/krisphq (last visited May 13, 2026)
Ex2026	Krisp Team, <i>What’s It Like to Work at Krisp?</i> (Sep. 20, 2021), https://krisp.ai/blog/working-at-krisp/ (last visited May 13, 2026) (“What’s it like to work at Krisp?”)
Ex2027	Arto Minasyan, <i>Krisp Voice AI Lab: Revolutionizing Voice Technology</i> , LinkedIn (posted Mar. 2026), https://www.linkedin.com/posts/artominasyan_eight-years-ago-we-started-with-a-simple-activity-7430219416912629760-8l6t (last visited May 13, 2026).
Ex2028	Krisp, <i>Introducing Accent Conversion for the Listener</i> , https://krisp.ai/blog/introducing-accent-conversion-for-the-listener/ (last visited May 20, 2026)

<u>EXHIBIT NO.</u>	<u>DESCRIPTION</u>
Ex2029	Arka Telecom, <i>YSU–Krisp AI Lab Opens at Yerevan State University</i> , https://arkatelecom.am/en/news/development/ysu_krisp_ai_lab_opens_at_yerevan_state_university/ (last visited May 20, 2026).
Ex2030	Careers at Krisp, Krisp, https://krisp.ai/careers/ (last visited May 14, 2026)
Ex2031	Krisp, Google Maps, https://www.google.com/maps (last visited May 13, 2026)
Ex2032*	U.S. Patent & Trademark Office, Enforcement and Non-Waiver of 37 C.F.R. § 42.104(b)(4) and Permissible Uses of General Knowledge in Inter Partes Reviews (July 31, 2025), https://www.uspto.gov/sites/default/files/documents/aa_pa_memo_final__signed.pdf (“Element Particularity Memo”)
Ex2033*	Defendant’s Proposed Terms for Construction, <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i>, No. 3:25-cv-05666-RS (KAW) (N.D. Cal. June 9, 2026)
Ex2034*	Defendant’s Invalidity Contentions, <i>Sanas.AI Inc. v. Krisp Technologies, Inc.</i>, No. 3:25-cv-05666-RS (KAW) (N.D. Cal. May 26, 2026)

*** New Exhibits submitted herewith.**

I. INTRODUCTION

Pursuant to 37 C.F.R. § 42.107, Patent Owner Sanas.AI Inc. submits this Preliminary Response to the Petition for inter partes review of Claims 1-22 of U.S. Patent No. 11,948,550 (the “’550 Patent”).

Institution here should be denied on the merits.

The ’550 Patent claims a complete, end-to-end, **real-time accent conversion** pipeline in which a machine-learning model is trained, that trained model is applied to received speech, fourth audio data is synthesized, and the synthesized data is converted into output speech. *See* Ex1001. The claims recite a specific, integrated sequence of operations required for execution of real-time accent conversion. *See id.* The Examiner understood exactly that and found that no prior art taught or suggested the claimed “[**1(c)**] **training and [1(d)] application of the machine learning model and the [1(e), 1(e)(i)] synthesizing of fourth audio data using the method described in the claims.**” Ex1002 at 660 (element labels added).

To prevail on obviousness, the Petition must show that the prior art teaches or suggests each limitation of Claim 1 (including elements [**1(c)**]-[**1(e)(i)**]), that a PHOSITA would have been motivated to combine the references with a reasonable expectation of success, and that the claimed invention as a whole would have been obvious. *See* 37 C.F.R. §42.104(b); 35 U.S.C. § 103. Here, the Petition does not do

so. Petitioner fails to identify any prior-art system—or coherent combination—that performs the claimed pipeline.

Faced with the reality that this invention is not taught in the prior art, as the Office previously found, Petitioner instead offers a combination of at least three disparate references built for different purposes and that operate on different types of data to achieve different results, using the patent as a hindsight-based roadmap and failing to offer any real motivation or reason to combine. For Ground 1, which all successive grounds rely upon, the Petition unsuccessfully cobbles together Fan, [Sonsaat](#), and [Loots](#).

Fan (Ex1004) is used as a superficially relevant backdrop because it generally concerns voice or accent conversion. However, Fan does **not** disclose the integrated pipeline the Office found to be missing in the prior art: it says nothing about training and applying the claimed machine-learning models or synthesizing fourth audio data in the claimed manner.

In an attempt to fill that gap, Petitioner imports [Sonsaat](#) (Ex1005) for training-related preprocessing, but [Sonsaat](#) is **not** directed to accent conversion—the relevant field of the invention. This non-analogous art uses a **different** data set and concerns corpus preparation, forced alignment, and mispronunciation annotation.

Then, Petitioner further imports [Loots](#) (Ex1006) for symbolic phoneme mapping, even though [Loots](#) does **not** use [Sonsaat](#) data, does **not** operate in Fan's neural conversion architecture, and does **not** synthesize audio data.

The resulting combination is not a technical roadmap but a parts-bin reconstruction. Petitioner fails to sufficiently explain:

- **how** Fan's deployed, real-time neural conversion architecture would absorb [Sonsaat](#)'s corpus-annotation workflow;
- **why** a system built on continuous acoustic features would be retrofitted with [Loots](#)' dictionary-level symbolic phoneme mapping; or
- **how** the hybrid would still train and apply the claimed machine-learning models and synthesize fourth audio data as the claims require.

Silence on these incompatibilities defeats any motivation to combine, forecloses any reasonable expectation of success, and leaves the required particularized mapping of claims unsupported.

And even if a PHOSITA were to somehow try to put together these disparate pieces, **the result would still *not* be the claimed invention.** It would be *chaos*. The combination is similar to taking a French teacher, retraining her to recognize abstract color patterns, and then expecting her students to demonstrate fluency by composing music—the pieces exist, but the combination makes no functional sense.

The remaining independent claim challenges to Claims 11 and 19 similarly fail at least on these same defects. Claims 1, 11, and 19 recite the same operative

invention across different classes—a system, a computer-readable medium, and a method—and each requires the same integrated pipeline of training a first machine-learning algorithm, applying it to microphone-received speech to derive non-text phoneme representations, synthesizing fourth audio data with a second trained machine-learning algorithm, mapping one phoneme representation to a different phoneme representation, and converting that synthesized data into second-accent output speech. *See* Ex1001 Cl. 1, 11, 19. The Examiner allowed the claims because the prior art failed to teach that integrated training, application, and synthesis sequence, and the Petition offers no separate theory for Claims 11 and 19 beyond its defective Claim 1 mapping. Ex1002 at 660. Because the same deficiencies infect the Petition’s inadequate reading on all three independent claims, institution must be denied.

Grounds 2-4 offer **no** cure for the flaws of Ground 1. They depend on the same deficient Fan-[Sonsaat-Loots](#) base combination and merely pile on even more disparate and irrelevant references that have no teachings to cure this defect. *See* Feinauer (Ex1008), Osuna (Ex1009), and Dalce (Ex1007) (offered for dependent-claim details). Putting aside the independent deficiencies of those grounds, for present purposes what matters is that none are offered to supply the missing Examiner-identified limitations or convert Petitioner’s disconnected fragments into a coherent real-time accent-conversion pipeline.

This failure is dispositive. The Petition **does not satisfy its burden** to establish a reasonable likelihood that Petitioner would prevail. Institution should therefore be denied.

II. THE '550 PATENT

The '550 Patent is directed to “techniques for real-time accent conversion.” Ex1001 Abstract. Claim 1 recites a specific pipeline for real-time accent conversion which trains and applies machine-learning models to convert received speech from a first accent into synthesized speech with a second accent by mapping phoneme representations. *Id.* Cl. 1.

The claimed invention operates by training a first machine-learning algorithm using audio data captured from multiple speakers having a first accent, with training that aligns and classifies frames of captured speech. Ex1001 Cl. [1(c)]. The system then applies that first machine-learning algorithm to microphone-received speech to derive a non-text linguistic representation of phonemes associated with a first pronunciation. *Id.* Cl. [1(d)]. Based on that non-text linguistic representation, the system synthesizes intermediate audio data representative of the speech content having a second accent using a second machine-learning algorithm trained with audio data comprising both first- and second-accent speech. *Id.* Cl. [1(e)]. The synthesis includes mapping a non-text representation of a first phoneme associated with the first pronunciation to a non-text representation of a different phoneme associated with a second pronunciation. *Id.* Cl. [1(e)(i)]. The system then converts the synthesized intermediate audio data

into output speech having the second accent and the updated phoneme set. *Id.* Cl. [1(f)].

Claim 1 requires:

[1(Pre)] A system, comprising:

[1(a)] at least one processor; and

[1(b)] non-transitory computer-readable medium comprising program instructions stored thereon that are executable by the at least one processor to cause the system to:

[1(c)] train a first machine-learning algorithm with first audio data comprising speech content captured from a plurality of different speakers having a first accent, wherein the training comprises aligning and classifying each of a first plurality of frames of the captured speech content corresponding to respective ones of the plurality of different speakers;

[1(d)] apply the first machine-learning algorithm to speech content received via at least one microphone, and comprising a set of phonemes associated with a first pronunciation of the received speech content, to derive a non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content;

[1(e)] based on the derived non-text linguistic representation of the set of phonemes associated with the first pronunciation of the received speech content, synthesize, using a second machine-learning-algorithm trained with (i) second audio data comprising the first accent and (ii) third audio data comprising a second accent, fourth audio data representative of the received speech content having the second accent,

[1(e)(i)] wherein synthesizing the fourth audio data comprises mapping at least a first non-text linguistic representation of a first phoneme of the set of phonemes associated with the first pronunciation of the received speech content to a second non-text linguistic representation of a second phoneme of an updated set of phonemes associated with a second pronunciation of the received speech content that is different from the first pronunciation of the received speech content, wherein the first and second phonemes are different phonemes; and

[1(f)] convert the synthesized fourth audio data into a synthesized version of the received speech content having the second accent, wherein the synthesized version of the received speech content having the second accent comprises the updated set of phonemes associated with the second pronunciation of the received speech content.

Ex1001 Cl. 1 (emphasis added to show elements recognized as inventive by the Office).

These ordered combinations of elements define **an integrated real-time pipeline that must operate cohesively** to achieve accent conversion by using a processor and computer-readable medium, training a machine-learning algorithm, applying the machine learning algorithm, synthesizing intermediate audio data, mapping non-text representations of phonemes, and converting intermediate audio data into output speech.

Critically, elements **[1(c)-1(e)(i)]** (bolded above)—training, application of the first machine learning model, and the synthesizing of fourth audio data through phoneme mapping—were recognized by the Examiner as not taught or suggested

by the prior art and as a basis for allowance. *See* Ex1002 at 660.¹ This finding *reflects the Office’s determination that the prior art failed to disclose the very feature that differentiates the claimed real-time pipeline.*

The Petition simply ignores this dispositive finding and attempts to reconstruct the claimed invention by cobbling together three disparate references—Fan, [Sonsaat](#), and [Loots](#)—without identifying any acceptable combination that addresses the Examiner’s stated reasons for allowance. The Petition:

- **does not show** the asserted art teaches or suggests the integrated sequence of training the first machine-learning algorithm, applying that trained model to received speech, and synthesizing fourth audio data through the claimed non-text phoneme mapping; and
- **does not explain how** [Sonsaat](#)’s forced-alignment corpus workflow or [Loots](#)’ symbolic dictionary-level phoneme mapping could be incorporated into Fan’s neural real-time accent-conversion pipeline to produce the claim-required system.

By failing to identify where these limitations are disclosed in the prior art, and by failing to explain how the proposed combination would perform those

¹ The Office recognized that this novel pipeline applied to all the pending claims, including the other independent Claims 11 and 19. Ex1002 at 660. In a similar manner, the Petition applies essentially identical (albeit defective) analysis to all three independent claims. Pet. at 16-44, 55-56, 57-58.

limitations as an integrated whole, the Petition lacks the particularized showing required by 35 U.S.C. § 312(a)(3) and should be rejected.

III. CLAIM CONSTRUCTION

Claims 1, 11, and 19 require synthesizing “fourth audio data representative of the received speech content having the second accent,” and then converting that synthesized fourth audio data “into a synthesized version of the received speech content having the second accent.” Ex1001 Cl. 1, 11, 19.

Petitioner proposes construing “fourth audio data representative of” as “synthesized acoustic feature data generated as an output of a speech conversion process and used as input to a subsequent conversion process that produces output waveform audio.” Pet. at 5. Petitioner further asserts that “[t]he specification consistently describes a multi-stage pipeline in which the ‘[voice conversion] VC engine’ outputs synthesized audio data in the form of acoustic features (e.g., mel spectrograms), which are then subsequently input into the ‘output speech generation engine’ and converted into waveform audio.” *Id.* Petitioner states that “[t]his confirms that the claimed ‘fourth audio data representative of’ is an intermediate output acoustic representation rather than a hearable waveform.” *Id.*

Patent Owner does not agree with this claim construction. Petitioner’s construction imports limitations the claim does not contain and should be rejected. The claim specifies what the fourth audio data represents and how it functions in the claimed pipeline, but it does not require that the fourth audio data be limited to “acoustic feature data,” “mel spectrograms,” or any particular data or

representation. *Id.* Nor does the claim language require importing Petitioner’s specific “output of a speech conversion process” and “input to a subsequent conversion process” formulation. *Id.* Nor does the claim language require that the “output” be “waveform audio”—to the contrary, the claim expressly provides that “fourth audio data” is converted “into a synthesized version of the received speech content having a second accent.” *Id.* Petitioner conspicuously ignores this express claim language and does not seek its construction. That is because it is the claim language that governs, not exemplary embodiments set forth in the specification. *Phillips v. AWH Corp.*, 415 F.3d 1303, 1323 (Fed. Cir. 2005) (“[A]lthough the specification often describes very specific embodiments of the invention, we have repeatedly warned against confining the claims to those embodiments.”). At bottom, the two phrases for which construction is requested—“fourth audio data” and “representative of”²—are plain English terms that the patent does not imbue

² Petitioner requested that these two phrases be combined and construed together is arbitrary and confusing, and yet another reason why the construction should be rejected. Among other things the inclusion of “representative of” without more text merely begs the question: “representative of what?” But the claim language itself answers this question (it is “representative of the received speech content

with any specialized or unique meaning. Their plain and ordinary meaning should govern.

In addition, the same claim term (“fourth audio data representative of”) is currently the subject of claim construction proceedings in the parallel district court litigation between the parties. Ex2033 at 2. Notwithstanding Petitioners’ assertion that their “broad *Sotera* stipulations confirm[] Petitioner will not bring any such challenges which could create any overlap,” Petitioner has requested that the district court construe the very same term at the *Markman* hearing scheduled for October 2026—after the expected institution decision in this case but almost one year before any final written decision would be due. *See* Petitioner’s Response to Patent Owner’s Discretionary Denial Brief, Paper 8 at 8; Ex2033 at 2. The district court proceedings will address the scope and meaning of the asserted claims based on a more developed record, including full briefing, discovery, and expert testimony. Proceeding with claim construction in parallel here risks inconsistent interpretations, duplicative efforts and decisions on an incomplete record, further underscoring that institution would not promote efficient resolution of the dispute.

having the second accent”)—underscoring that Petitioner’s proposed construction is incorrect and unnecessary.

In any event, for purposes of institution, *it does not matter which construction the Board adopts because Petitioner's analysis fails either way.*

Accordingly, Patent Owner submits that no express construction is necessary at this preliminary stage. To the extent construction is required, the Board should reject Petitioner's attempt to import exemplary compression and dimensionality-reduction embodiments, and exemplary outputs, into the claims. Patent Owner reserves the right to propose express constructions should trial be instituted.

IV. THE PETITION FAILS TO DEMONSTRATE MOTIVATION, REASONABLE EXPECTATION OF SUCCESS, AND PARTICULARIZED MAPPING OF CLAIM ELEMENTS

The Board may deny institution where a petitioner fails to satisfy its threshold burdens, including demonstrating:

- (1) a reasoned motivation to combine prior art references,
- (2) a reasonable expectation of success, and
- (3) a particularized showing of where each claim element is disclosed in the asserted prior art.

35 U.S.C. §§ 103, 312(a)(3), 314(a); *see also KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398, 418 (2007).

Each of these requirements is independently dispositive, and failure to satisfy any one of them warrants denial as they each independently show that there is not a reasonable likelihood of success that the petitioner would prevail. 35 U.S.C. §§ 103, 312(a)(3), 314(a).

Here, the Petition fails since it never shows the claimed integrated real-time accent-conversion pipeline. The reason the Petition cannot make this showing is because **the combination is inoperable at each step, and *no amount of additional explanation can reconcile these architectural defects.***

For training step [(1(c))], the proposed combination of the forced-alignment process in **Sonsaat** (not relating to the same field of endeavor as the '550 Patent-in-suit), with the architecture of Fan is simply not functional.

Sonsaat provides **annotated speech recordings: the corpus contains orthographic transcriptions, forced-aligned phoneme transcriptions, and manual annotations.** Ex1005 at 1, 3. In other words, **Sonsaat** gives recording mark-ups that describe **what was said and how it was pronounced.** Fan, *by contrast*, trains its dialect-conversion model **using matched input/label examples**, where “the input is Sichuanese, and the label is Mandarin with the same content as Sichuanese,” or vice versa. Ex1004 ¶[0140]. Put simply, **Sonsaat** provides **labeled examples of speech**, while Fan requires **paired “before-and-after” examples** showing the same content in two dialect forms. Because **Sonsaat** does **not** provide those paired, same-content input/label examples, it does **not** supply the type of training data Fan uses to train its dialect-conversion model.

Thus, **Sonsaat**’s workflow is **not a deployable training component** for Fan’s system and **does not provide the infrastructure needed** to prepare training data within Fan’s architecture for dialect conversion. The proposed Fan-**Sonsaat** combination therefore **does not** and **cannot** coherently produce the trained first machine-learning algorithm required by the claims.

For application step [1(d)], the proposed combination also fails. It cannot coherently perform the claimed training step; it yields no first machine-learning algorithm with defined properties that can later be applied to microphone-received

speech. Nor does it yield a non-text linguistic representation having any defined relationship to Fan’s dialect conversion process.

The proposed combination of Loots’ decision tree being inserted into Fan’s neural pipeline also fails to teach the claimed synthesis and mapping steps [1(e)]-[1(e)(i)].

Loots works with phoneme strings—that is, written sequences of phoneme symbols. It performs accent conversion by converting “one phoneme string ... to another” using decision trees trained on pronunciation dictionaries that use ARPABET phoneme symbols. Ex1006 at 2-3, 5-6. **Fan works differently.** Fan does **not** convert one phoneme string into another; **instead**, it extracts continuous PPG features from speech, feeds those features into a seq2seq neural network, generates Mel-spectrogram features, and then converts those features into playable audio. Ex1004 ¶¶[0097], [0099], [0126]-[0127]. Put simply, Loots **operates on discrete phoneme labels, while Fan operates on continuous speech features.** Because those are different types of data, Loots’ **output cannot simply be plugged into Fan’s PPG-to-Mel neural pipeline.** Doing so would require additional conversion steps between phoneme labels and speech features, which Fan does not disclose or suggest.

For conversion step [1(f)], the further cobbling together of the annotations from the non-analogous Sonsaat, with the symbolic phoneme strings of Loots,

does not make any sense as the input/output of Fan's neural PPG-to-Mel-to-speech pipeline. Claim [1(f)] requires converting the synthesized fourth audio data into synthesized speech having the second accent and the updated set of phonemes. Ex1001 Cl. [1(f)]. But the proposed combination never produces coherent synthesized fourth audio data in the first place: **Sonsaat** supplies only annotations (Ex1005 at 1-3), **Loots** supplies only symbolic phoneme strings (Ex1006 at 1-2), and the Petition does not provide an operable way to integrate those outputs into Fan's neural PPG-to-Mel-to-speech pipeline (Ex1004 ¶[0097]). Without functional synthesized fourth audio data that reflects the claimed phoneme mapping, there is nothing for the final conversion step to convert into output speech having the claimed updated phoneme set.

Each of these incompatibilities separately render the proposed combination non-operable as the claimed pipeline. The result is not an integrated system that trains, applies, synthesizes, maps, and converts as the claims require but a set of disconnected components that cannot be assembled into the claimed invention without rebuilding the system. These failures are carried through to all the grounds which rely upon this defective analysis.

A. The Petition Does Not Demonstrate a Reasoned Motivation to Combine Prior Art References

First, obviousness *requires more than identifying references* that could be combined; a petitioner *must articulate a reasoned motivation to combine those*

references, supported by a rational underpinning. *KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 418 (2007); *In re Kahn*, 441 F.3d 977, 988 (Fed. Cir. 2006).

Conclusory assertions—such as invocations of “known methods”, reprogramming, or “predictable results”—are insufficient where, as here (*see e.g.* Pet. at 26, 27, 43, 62, 77), *the petition does not explain how the specific disclosures would be integrated into a working combination.* *See ActiveVideo Networks, Inc. v. Verizon Commc’ns, Inc.*, 694 F.3d 1312, 1327-28 (Fed. Cir. 2012) (rejecting generic expert testimony that “failed to explain how specific references would be combined, which combination(s) of elements in specific references would yield a predictable result, or how any specific combination would operate or read on the asserted claims”).

Petitioner’s declarant Mr. Schmandt (Ex1003) should be accorded **little** weight since he **merely** repeats unsupported assertions **without** providing additional supporting analysis or evidence. *Xerox Corp. v. Bytemark, Inc.*, IPR2022-00624, Paper 9 at 15-17 (PTAB Aug. 24, 2022) (precedential) (holding that the expert testimony should be “entitled to little weight” because the declaration “is conclusory and unsupported” and “adds little to the conclusory assertion for which it is offered to support”).

Petitioner’s argument that it is not offering expert testimony to fill the gap in the prior art is disingenuous. Petitioner’s Response to Patent Owner’s

Discretionary Denial Brief, Paper 8 at 14. Petitioner asserts that “every limitation is directly mapped” and that Patent Owner has not identified any missing limitation supplied by expert testimony. *Id.* But that is not true. Patent Owner calls out the Petition’s failure to directly map any specific prior art for Examiner identified elements [1(c)] and [1(e)(i)]. *See* Pet. at 24-27, 42-44; *see also* Patent Owner’s Request for Discretionary Denial at 19-20.

During prosecution, the Examiner noted that no cited prior-art reference teaches the pipeline as arranged in the challenged claim. Ex1002 at 660. Petitioner relies on Mr. Schmandt’s testimony to bridge the gap—asserting that **Sonsaat**’s forced-alignment process and English-language recordings could be imported into Fan’s training data and model-training pipeline and then further combined with **Loots**’ phoneme mapping. Petition at 24, 42; Ex1003 ¶¶[108]-[114], [136]-[139]. However, the references themselves **do not provide that combined architecture or explain why** a PHOSITA would have stitched those disparate teachings together. That is not a teaching in the prior art; it is expert reconstruction. Besides being nonsense, this is exactly the type of expert gap filling that the PTAB cautions against. *Johns Manville Corp. v. Knauf Insulation, Inc.*, IPR2018-00827, Paper 9 at 12-14, 18-20 (PTAB Oct. 16, 2018) (informative) (denying institution and according to the expert declaration little weight where conclusory assertions and expert reliance did not sufficiently explain why the asserted prior-art elements

would have been selected and combined to achieve the claimed invention with predictable results).

Here, as discussed in Section V(A), Petitioner does not provide a reasoned explanation for **how** Fan, [Sonsaat](#), and [Loots](#) would be combined into a coherent system that performs the claimed real-time pipeline. While the Petition identifies purported reasons why a PHOSITA would have combined Fan, [Sonsaat](#), and [Loots](#), those assertions are conclusory and **fail** to address the technical **incompatibilities** noted above that such a combination entails or **how** such combination would purportedly obtain the claimed invention.

With respect to [Sonsaat](#), the Petition argues that Fan generally contemplates corpus-based training and that incorporating [Sonsaat](#)'s MFA would be a “simple design choice” involving only “minor programming adjustment[s].” Pet. at 27. But the Petition never explains **how** [Sonsaat](#)'s forced-alignment and corpus-generation workflow could function as a training component within Fan's neural dialect conversion architecture. That failure is fatal. The Petition only proffers an unsupported assertion that [Sonsaat](#)'s teachings are “readily applied” to other accents and languages, without sufficient detail and logic (or evidentiary support) to make a sufficient bridge to justify institution. Pet. at 24.

The same defect of failing to address technical incompatibilities appears in the Petition's reliance on [Loots](#). [Loots](#) performs symbolic phoneme mapping

through a decision tree (Ex1006 at 6), whereas Fan processes continuous PPG representations and generates Mel features through a seq2seq neural network (Ex1004 ¶¶[0097], [0099], [0126]-[0127]). The Petition asserts without support or analysis that “the phonemes mapped (per Loots) is the PPG...feature.” Pet. at 42. But it never explains how Loots’ symbolic phoneme-mapping logic could be (let alone would be) inserted into Fan’s continuous PPG-to-Mel pipeline. The PPG in Fan is a continuous, frame-by-frame probability distribution across an entire phoneme set, whereas Loots operates on discrete, certain phoneme symbols at the dictionary entry level—Loots and Fan represent phonetic information in **fundamentally different formats**, and Loots’s fixed phoneme labels **cannot** be fed into a pipeline designed to process Fan’s PPG features. Thus, the Petition **fails** to sufficiently explain not only that the combination would make the claimed result (in other words what happens when the symbolic phoneme is put into Fan instead of the PPG) but **fails** to provide a proper motivation to make such a combination in the first instance.

Petitioner’s theory is like taking a machine designed to turn flour, water, and yeast into bread, changing the input to raw gelatin, and assuming the same machine will still produce bread. It will not. Likewise, Fan’s neural pipeline was designed for continuous PPG inputs, not Loots’ symbolic phoneme strings. The Petition

never explains why Fan would still work—much less produce the claimed output—after that substitution.

Because the Petition relies on conclusory assertions of convenience and ease of modification rather than reasoned analysis addressing the incompatibility, it fails to establish the required motivation to combine.

B. The Petition Does Not Demonstrate a Reasonable Expectation of Success

Second, a petitioner must demonstrate *a reasonable expectation of success for the proposed combination, not merely a motivation to attempt it*. *Intelligent Bio-Sys., Inc. v. Illumina Cambridge Ltd.*, 821 F.3d 1359, 1367-68 (Fed. Cir. 2016); *PAR Pharm., Inc. v. TWI Pharms., Inc.*, 773 F.3d 1186, 1193 (Fed. Cir. 2014). While a combination need not be “the preferred, or the most desirable, combination,” the petitioner must show “there is something in the prior art as a whole to suggest the desirability, and thus the obviousness, of the claimed invention.” *Honeywell Int’l Inc. v. 3G Licensing, S.A.*, 124 F.4th 1345, 1356 (Fed. Cir. 2025) (quotation omitted).

Here, as discussed in Section V, the Petition gives a PHOSITA no adequate basis to expect that the proposed combination would succeed in producing the integrated pipeline that Claims 1, 11 and 19 require.

Petitioner does not explain *how* the disparate systems of Fan, [Sonsaat](#), and [Loots](#) would **function together** to achieve the claimed integration of pipeline components required by Claims 1, 11 and 19.

Nor does the Petition explain *how* these fundamentally different architectures could be integrated (let alone *why* they would be integrated) to perform those operations, particularly the Examiner-identified limitations that were found missing from the prior art.

Fan is a deployed, real-time neural speech-conversion system that processes continuous PPG features through a seq2seq PPG-to-Mel pipeline, while [Sonsaat](#) is a corpus-annotation and forced-alignment workflow dependent on transcripts and ARPABET labeling, and [Loots](#) performs symbolic phoneme-string mapping using decision-tree logic applied to discrete pronunciation-dictionary entries. *See* Ex1004; Ex1005; Ex1006.

These are not merely different references but fundamentally incompatible architectures, and the Petition never explains *why* a skilled artisan at the time of the invention (without the patent as a roadmap) would have expected continuous, learned representations and discrete, symbolic operations to interoperate within a single end-to-end system. An expectation of success asserted at the broad level of generality of the Petition, untethered from any substantive analysis bridging these

architectural differences, is no expectation of success at all, and cannot sustain the obviousness ground.

C. The Petition Does Not Make a Particularized Showing of Where Each Claim Element is Disclosed in the Asserted Prior Art

Third, the petition must identify “**with particularity**” where each claim element is disclosed in the prior art and **explain how** the asserted grounds render the claims unpatentable. 35 U.S.C. § 312(a)(3); 37 C.F.R. § 42.104(b)(4). The Office has made clear that missing claim limitations cannot be supplied by general knowledge, common sense, or conclusory expert testimony. Ex2032 at 1.

Here, the Petition maps the individual elements of independent Claims 1, 11, and 19 to scattered disclosures across three references—application to Fan, training to Fan as supplemented by **Sonsaat**, synthesizing to Fan, and phoneme mapping to **Loots**—but never makes a particularized showing that those pieces combine into the claimed pipeline as an operable whole. Pet. at 16-45. That failure is especially clear for element [1(e)(i)]. That limitation does not merely require phoneme mapping; it requires that the mapping occur **as part of “synthesizing the fourth audio data”** representative of the received speech content having the second accent.

Loots does not do that. It provides symbolic, dictionary-level phoneme mapping only. It does **not** synthesize audio data, does **not** produce fourth audio

data, and does **not** show how its dictionary-level phoneme strings become the audio data that Fan later converts into speech. Thus, pointing to [Loots](#) for phoneme mapping is not sufficient. The Petition needed to explain **how** [Loots](#)' mapping is **integrated** into Fan's **synthesis of fourth audio data** in the proposed Fan-[Sonsaat-Loots](#) pipeline. **The Petition provides no such explanation.**

That omission is fatal under § 312(a)(3). The claim requires the training, application, synthesis, and non-text phoneme mapping steps of elements **[1(c)]-[1(e)(i)]** to operate **within a single integrated system**. In allowing the claims, the Examiner found that integrated sequence was missing from the prior art. The Petition never supplies the missing integration. It instead identifies isolated parts in separate references and assumes the claimed system follows.

Identifying where each limitation appears in some reference, standing alone, **does not discharge Petitioner's burden**: the claim requires that training, application, synthesis of the intermediate audio data, and non-text phoneme mapping (elements **[1(c)]-[1(e)(i)]**) be performed within a single, integrated system, and it is that integration—the very limitation the Examiner identified as missing from the prior art and that drove allowance—for which the Petition offers no particularized analysis.

Thus, Petitioner cobbles the limitation together from separate and disparate sources **without showing**, element to element, **how** Fan's continuous neural

pipeline, [Sonsaat](#)'s alignment workflow, and [Loots](#)' discrete symbolic mapping combine to form a coherent pipeline that performs the claimed sequence of operations, let alone that such a combination would result in the claimed pipeline.

Apple Inc. v. Ginko LLC underscores that, where, as here, a petition fails to map claim elements with particularity, denial at institution is warranted. IPR2025-01388, Paper 14 at 3 (March 2, 2026). There, the Director emphasized that a petitioner must, “from the onset,” provide an understandable, element-by-element explanation grounded in the prior art. Petitioner’s showing here falls short for the same reasons: it neither identifies a prior-art disclosure of the claimed **training, application of the machine learning model and synthesizing of fourth audio data**, nor explains how the proposed combination performs **that requirement within the claimed pipeline**.

D. Institution Should Be Denied as an Inefficient Use of Office Resources

Finally, denial is also warranted where, as here, institution would not serve the efficient administration of the Office. The Director has confirmed that the Board may deny institution where only a subset of claims or grounds has potential merit and proceeding to trial would waste resources. *CyberSecure IPS, LLC v. Network Integrity Systems, Inc.*, IPR2025-01441, Paper 18 at 2-3 (May 26, 2026)

(citing *Chevron Oronite Co. v. Infineum USA*, IPR2018-00923, Paper 9 at 10-11 (November 7, 2018) (informative)).

Petitioner's reliance on the *Sotera*-style stipulation (Petitioner's Response to Patent Owner's Discretionary Denial Brief, Paper 8 at 6-11) **does not meaningfully mitigate** this inefficiency. Although Petitioner purports that it will not pursue in the district court litigation "the specific grounds raised in IPR2026-00272" (Ex1046 at 2), the submitted invalidity contentions confirm that Petitioner continues to assert broad combinations of references (including the art presented in this proceeding, along with other art), system-based theories, and expert-driven PHOSITA reasoning that fall outside the scope of any estoppel. Ex2034 at 4-13. For example, Petitioner expressly reserves the ability to mix and match prior art across references and systems, expanding the scope of its invalidity case beyond the grounds asserted in the Petition and undermining the intended narrowing effect of *Sotera*-style stipulations:

The references in the below charts may be combined with one or more references in the same chart, another chart, or another piece of prior art identified in these contentions, for at least the same exemplary reasons provided in the motivation to combine sections of the charts and **Krisp's IPR and PGR Petitions, which are incorporated by reference**. IPR2026-00275 ('457 Patent); PGR2026-00032 ('561 Patent); IPR2026-00273 ('496 Patent); IPR2026-00272 ('550 Patent); IPR2026-00274 ('745 Patent); PGR2026-00033 ('756 Patent).

Ex2034 at 10 (emphasis added).

Thus, the same underlying factual disputes, expert testimony, and technical theories will proceed in parallel regardless of institution. The stipulation alters only the labeling of Petitioner's theories—not their substance—leaving intact duplicative proceedings that do not simplify the issues before the parties or the Board. Moreover, Petitioner continues to assert invalidity and has reserved broad rights to amend and supplement its invalidity positions. Ex2034 at 2-3. These reservations ensure that the district court litigation will proceed in full irrespective of any stipulation, further confirming that institution would not reduce the scope of the dispute and would result in inefficient use of the Board's resources.

Here, all asserted grounds depend on the same defective Fan, **Sonsaat**, and **Loots** combination, and the additional references in Grounds 2-4 **do not cure the deficiencies** in the independent claims.

Instituting trial would therefore require the Board to expend resources on multiple grounds that fail to address the dispositive issue.

V. PETITIONER’S CHALLENGES FAIL ON THE MERITS

Petitioner asserts four obviousness grounds:

Ground	Claims Challenged	References Relied Upon	Petition Citation
Ground 1	1-4, 7, 10-13, 16, 19, 21-22	Fan, Sonsaat , Loots	Pet. at 16
Ground 2	8-9, 17-18, 20	Fan, Sonsaat , Loots , Feinauer	Pet. at 58
Ground 3	5, 14	Fan, Sonsaat , Loots , Osuna	Pet. at 70
Ground 4	6, 15	Fan, Sonsaat , Loots , Dalce	Pet. at 73

This Preliminary Response focuses on Ground 1 because Ground 1 supplies the base combination that every other asserted ground, and every claim, requires. Ground 1 asserts that Claims 1-4, 7, 10-13, 16, 19, and 21-22 are obvious over Fan, [Sonsaat](#), and [Loots](#). Pet. at 16. Grounds 2-4 merely add Feinauer, Osuna, or Dalce for dependent-claim limitations, while retaining the same Fan-[Sonsaat](#)-[Loots](#) combination for the independent-claim limitations. Pet. at 58, 70, 73. Thus, if Ground 1 fails to show the claimed pipeline in independent Claims 1, 11, and 19, Grounds 2-4 fail as well.

A. Ground 1 Fails on the Merits

Ground 1 fails because Petitioner attempts to reconstruct the '550 Patent through hindsight rather than identify a teaching or reasoned combination that a PHOSITA at the time of the invention would have recognized in the prior art.

As noted, the Petition relies on three very disparate references that perform meaningfully different functions using materially distinct architectures:

- (1) Fan is a voice conversion or “human voice beautification” system with selectable style conversion, dialect conversion, and voice enhancement modes. Ex1004 ¶¶[0004]-[0007], [0045], [0055]-[0060].
- (2) **Sonsaat** is a non-native English speech corpus with forced-aligned transcriptions and manual mispronunciation annotations, **not accent conversion** like the '550 Patent. Ex1005 Abstract, 1, 3-4.
- (3) **Loots** is a pronunciation-dictionary conversion method using deterministic decision trees to convert one phoneme string to another. Ex1006 Abstract, 6, Fig. 2.

The Petition never bridges those differences. It **does not identify** a coherent combined architecture, **explain** which components are retained or discarded, or **show** how outputs from one system become usable inputs to another. In particular, the Petition **does not sufficiently explain** how Fan’s accent conversion model would implement **Loots**’ decision tree or **Sonsaat**’s forced alignment. Those deficiencies defeat both motivation to combine and reasonable expectation of success.

Accordingly, the challenges to independent Claims 1, 11, and 19 fail on the merits.

1. Ground 1 Independently Fails Because Sonsaat is Not Analogous Art

Petitioner's Ground 1 fails at the threshold because **Sonsaat** (Ex1005) is not analogous art. A reference is analogous only if it is from the same field of endeavor or is reasonably pertinent to the problem faced by the inventor. *In re Bigio*, 381 F.3d 1320, 1325 (Fed. Cir. 2004) ("Two separate tests define the scope of analogous prior art: (1) whether the art is from the same field of endeavor, regardless of the problem addressed and, (2) if the reference is not within the field of the inventor's endeavor, whether the reference still is reasonably pertinent to the particular problem with which the inventor is involved"). **Sonsaat** satisfies neither requirement.

Sonsaat is **not in the same field of endeavor** as the '550 Patent. The '550 Patent is directed to a real-time accent-conversion pipeline that trains and applies machine-learning algorithms to microphone-received speech, derives non-text linguistic representations of phonemes, synthesizes fourth audio data having a second accent based on mapping the non-text linguistic representations to a different second phoneme representation, and converts the synthesized fourth audio data into second-accent output speech. **Sonsaat**, by contrast, is directed to generating a speech corpus of non-native English and discloses preparing and

annotating recorded speech using transcriptions, forced alignment, and manual mispronunciation labels. **Sonsaat** does not disclose a deployed real-time accent-conversion system, does not synthesize accent-converted audio, and does not train or apply the claimed machine-learning pipeline to microphone-received speech.

Sonsaat is also not reasonably pertinent to the problems addressed by the '550 Patent. The '550 Patent addresses problems related to real-time accent conversion including “different pronunciations of certain sounds that are inherent to a given accent,” preservation of “prosody or emotion of the speaker” and latency. Ex1001 1:40-2:3. **Sonsaat** addresses the different problem of creating an annotated non-native English corpus for later research and evaluation. Its forced-alignment workflow is used only to label a static corpus after recordings are collected, edited, and segmented; it assumes pre-recorded speech paired with transcripts and produces time-aligned word and phoneme annotations for later analysis. A PHOSITA seeking to build the claimed real-time training, application, synthesis, mapping, and conversion pipeline would not have looked to **Sonsaat**'s offline corpus-annotation process as a pertinent solution.

Because **Sonsaat** is not analogous art, it cannot supply the claimed chunking or context limitations in Ground 1. Petitioner's reliance on **Sonsaat** is hindsight-driven, and Ground 1 fails at the threshold.

2. The Petition Fails to Establish that Loots' Decision Tree Can Be Inserted into Fan's Pipeline Without Disrupting the Model

The Petition asserts that [Loots](#)' P2P decision tree is “incorporated into the dialect conversion model” of Fan (Pet. at 47) but never explains:

- *where* in Fan's architecture that tree would actually reside,
- *how* it would function alongside Fan's existing components, or
- *why* a continuous acoustic feature system would be altered to include [Loots](#)' symbolic phoneme-mapping.

That omission is significant given the fundamentally different processing pipelines of Fan and [Loots](#).

The Petition does not merely leave unexplained where [Loots](#)' phoneme decision tree would be placed within Fan's neural voice-conversion pipeline; the omission masks a more fundamental incompatibility. There are only two possible placements for [Loots](#)' decision tree, and each confirms that the proposed combination is illogical and if made, **would not result in the claimed invention.**

Petitioner provides no reason why a system based on continuous acoustic feature representations would be modified to include [Loots](#)' symbolic, dictionary-level phoneme mapping, particularly where such a modification would require collapsing probabilistic PPG representations into discrete phoneme labels—discarding information Fan relies on and undermining its stated robustness. Fan's system is an end-to-end neural architecture that processes learned

acoustic/phonetic feature representations through a seq2seq model to generate converted speech features. Ex1004 ¶¶[0097], [0099], [0126]-[0127]. [Loots'](#) decision tree, by contrast, is a symbolic phoneme-classification mechanism that operates on discrete phoneme labels and fixed true/false comparisons. Ex1006 at 3.1-3.3. The Petition **does not explain how Loots' decision tree would be** integrated into Fan and combining them would require converting between continuous speech features and discrete phoneme labels—a process that neither reference alone or in combination discloses—and would disrupt, rather than improve, Fan's system.

If the tree were placed upstream of Fan's seq2seq neural network—*i.e.* between the voice feature extraction model and the encoder—then Fan's system would convert the source-accent phonetic content into target-accent pronunciations before Fan's encoder processes the input, resulting in the receipt of inputs different from the data on which the system was trained. In Fan, the dialect conversion model receives PPG features extracted from the source-dialect speech along with style and timbre information used to preserve characteristics of the speaker's voice during conversion. Ex1004 ¶¶[0097], [0126]-[0127]. “Upstream” placement necessarily requires [Loots'](#) decision tree to intervene between the voice feature extraction model and the encoder by converting the source-accent phonetic content in the PPG into target-accent pronunciations before the encoder processes the

input. Thus, feeding the network already-converted pronunciations would alter that input structure and would require retraining the model on a new input. Moreover, because [Loots](#) performs phoneme conversion independently, introducing it here would bypass, rather than complement, the neural network's role. The alternative placement is no more workable. If [Loots](#)' decision tree were inserted downstream of Fan's seq2seq network, the tree would receive Fan's output Mel spectrogram features. Ex1004 ¶¶[0053], [0126]. But [Loots](#)' decision tree operates only on discrete symbolic phoneme labels using fixed true/false phoneme comparisons. Ex1006 at 4, 6. Because Fan's mel spectrogram is a continuous acoustic representation rather than a symbolic phoneme sequence, [Loots](#)' phoneme-based decision tree has no disclosed mechanism for processing it.

The Petition fails to specify either option, let alone how the combination would result in the claimed invention. These architectural problems directly undermine the Petition's assertion that the proposed combination "would have required only minor reprogramming" using "hardware already present." Pet. at 43. Retraining Fan's end-to-end neural conversion model on a materially different input structure is **not minor reprogramming**, and the Petition never identifies a workable mechanism for integrating [Loots](#)' symbolic decision tree into Fan's existing neural pipeline without substantial modification. In either placement of

Loots' decision tree, the proposed system **would not produce the claimed synthesized fourth audio data**:

- upstream placement destroys the input assumptions of Fan's trained neural model,
- while downstream placement gives Loots data it cannot process at all.

3. Petitioner Fails to Show an Operable Combination of Loots' Symbolic P2P Decision Trees with Fan's Neural PPG-to-Mel Pipeline

Loots is incompatible with Fan in the Petition's proposed combination. Loots works word-by-word on dictionary entries. It converts one symbolic phoneme string into another using decision trees. Ex1006 Abstract, 4, 6, Fig. 2. Its output is a dictionary-style pronunciation sequence—**not** an acoustic representation, **not** a mel spectrogram, and **not** continuous speech. Ex1006 at 2-3.

By contrast, Fan is a real-time system that takes received speech and converts it into another dialect, while trying to make the output sound natural. Fan receives audio, extracts PPGs, feeds representations into a seq2seq neural network, and outputs a mel feature that is converted into playable speech. Ex1004 ¶[0126]-[0127], [0140]-[0141].

Petitioner relies on Loots—particularly its statement that converting a word's pronunciation from one accent to another using P2P conversion is more accurate than using G2P conversion—in arguing that a PHOSITA would have added Loots' P2P decision tree to Fan's dialect-conversion model. Pet. at 42-43;

Ex1006 at 8. But that argument treats Loots and Fan as compatible merely because both involve accents and phonetic information. In reality, they are **not similar in how they function**. Loots converts one phoneme string into another using pronunciation dictionaries and decision trees. Ex1006 at 2, 6. Fan does something different: it takes speech, extracts continuous PPG features, processes those features through a seq2seq neural network, generates Mel-spectrogram features, and then converts those features into speech. Ex1004 ¶¶[0097], [0099], [0126]-[0127]. Thus, Loots improves a system that changes **phoneme labels into other phoneme labels**, while Fan uses a neural network to process **continuous speech features**. The Petition **does not explain** why an accuracy benefit in Loots's dictionary-based phoneme-mapping system would apply to Fan's different neural speech-conversion system.

The Petition **does not sufficiently explain** any motivation to combine these disparate references beyond hindsight reconstruction based on the '550 Patent. Moreover, the Petition does not adequately explain how Loots' decision tree mapping of a "focus" phoneme and "current phoneme" would be inserted into Fan's continuous PPG-to-mel conversion pipeline without redesigning Fan. **The proposed combination would not result in the claimed invention.**

Integrating Loots into Fan would require more than "minor reprogramming." Cf. Pet. at 43. To feed Fan's PPG into Loots, the system would

first have to collapse the PPG into a phoneme label. But Fan specifically used a PPG as opposed to “convert[ing] the voice into text” (akin to the phoneme strings of [Loots](#)) “to retain the information content of the voice”, further allowing “robustness to be improved”. Ex1004 ¶[0007]. Implementing phonemes into Fan would discard information that its neural model relies on, thus precluding the ability of Fan to implement an audio output. Thus, the Petitioner needed to explain how [Loots](#)’ symbolic decision-tree mapping could operate on Fan’s PPG representation without first collapsing the PPG into a phoneme sequence. Petitioner’s assertion that “the phonemes mapped (per [Loots](#)) is the PPG feature” merely conflates a probability distribution with a discrete phoneme string and fails to show why a PHOSITA would have been motivated to make the proposed modification or reasonably expect it to preserve Fan’s stated robustness advantage. Thus, the Petition fails to demonstrate that even if there was motivation to make such a combination (which there is not), that the combination would result in the claimed invention, let alone a workable system.

Furthermore, after applying [Loots](#)’ decision tree to phoneme labels, the system would then need to reconstruct a continuous representation suitable for Fan’s downstream seq2seq network. **The Petition identifies no such reconstruction step.**

Without a disclosed mechanism to convert Fan's PPG features into Loots-compatible symbolic phoneme strings and then convert Loots' symbolic output back into the continuous representation required by Fan's seq2seq model, the combined system has no operable data path from speech input to synthesized fourth audio data. The claimed synthesis and conversion steps therefore cannot be performed and **would not result in the claimed system**. The proposed combination would disrupt Fan's existing pipeline rather than improve it.

4. Petitioner Fails to Show That Sonaat's Forced Alignment is Compatible with Fan's Architecture

The Petition asserts that a PHOSITA would have implemented Sonaat's MFA into the voice feature extraction model of Fan as a minor programming adjustment. Pet. at 27. That assertion misstates the role of forced alignment, mischaracterizes Fan's architecture, and proposes a combination that contradicts Fan's own teachings about where and when training occurs.

The Petition misstates MFA by treating a forced-alignment tool used to label pre-recorded, transcribed speech data as if it were a trainable, terminal-based component of Fan's real-time speech-conversion architecture. Sonaat uses MFA as a preprocessing tool: after its corpus was recorded, edited, and segmented, MFA was run on the pre-recorded audio to produce speech labels (phonetic transcriptions), identifying where words and individual speech sounds begin and

end within a pre-recorded speech dataset. Despite the Petition's cursory treatment of the Fan-Sonsaat combination, the MFA in Sonsaat is fundamentally different from the components of Fan's architecture: it is merely a preprocessing tool, **not** part of a trained model, **not** part of an inference pipeline, and **not** part of a deployed real-time conversion system.

Furthermore, the Petition's proposed placement of MFA **at the terminal** contradicts Fan's teaching that training is **at the server**, **precluding the combination from resulting in the claimed invention**.

The Petition asserts that "implementing the Montreal forced-aligner of Sonsaat into the terminal of Fan would have constituted a simple design choice," and "implementing the Montreal forced-aligner of Sonsaat into the programming of Fan's terminal allows **the terminal to train** the voice feature extraction model." Pet. at 26-27.

But Fan teaches the opposite. Fan's voice feature extraction model "is trained in advance by using a large quantity of corpuses" and, once trained, "does not participate in the parameter update of the entire network." Ex1004 ¶[0113]. Fan is explicit about **where** that advance training occurs: "training and update of each model may be performed **by a server**. After **the server trains** a voice processing neural network, the terminal downloads the voice processing neural network to a local end." Ex1004 ¶[0143]. In Fan, the terminal receives live input

speech, extracts PPG features using the **already-trained** model, feeds those features to a dialect conversion model, and outputs converted speech through a vocoder. Ex1004 ¶¶[0126]-[0127], Fig. 9.

The Petition thus proposes installing a training-time corpus-annotation tool inside the very component of Fan's system where, by Fan's own teaching, **training does not occur**—and asserts that this contradiction of Fan's architecture is a “simple design choice.” It makes no sense, and is far from simple.

The Petition's assertion that implementing MFA into the voice feature extraction model of Fan would be a minor programming adjustment understates the architectural change required. It fails to explain how **Sonsaat's** forced-alignment workflow becomes part of the claimed machine-learning algorithm in Fan, as opposed to merely being an external step. Improperly, the Petition treats **Sonsaat's** MFA as though it were already trained to act on Fan's model architecture.

Moreover, if **Sonsaat's** Montreal Forced Aligner is used during training it cannot be implemented into the terminal as Petitioner asserts (Pet. at 27; Ex1003 ¶114), because Fan confirms training occurs “in advance” as a separate phase from the terminal's runtime application of the already-trained model, making forced alignment part of data preparation rather than a deployed terminal component. Ex1004 ¶ [0113]. If instead the MFA is used during inference, it cannot process live microphone speech as claimed, because forced alignment structurally requires

a pre-existing corpus paired with transcripts and a trained acoustic model — none of which exist for the live, untranscribed microphone input Fan receives. Furthermore, the MFA's output is time-aligned phoneme annotations (Ex1005 at 3), not the non-text linguistic representation (PPG) that Fan's first machine-learning model derives (Ex1004 ¶[0113]).

Thus, the proposed modification breaks Fan's division between training and inference: the terminal would be asked to perform training-stage corpus alignment on live input even though Fan's terminal is designed only to apply pre-trained models. The result is a system lacking an identified trained first machine-learning algorithm that can be applied as claimed. *Thus, contrary to the Petition, the combination makes no sense, and would not result in the claimed invention.*

5. Ground 1 Dependent-Claim Challenges Do Not Cure the Independent-Claim Defects

Ground 1 also challenges independent Claims 11 and 19, along with dependent Claims 2-4, 7, 10, 12-13, 16, and 21-22. Because Petitioner fails to establish the limitations of Claim 1, its challenges to all dependent claims necessarily fail. *In re Fine*, 837 F.2d 1071, 1076 (Fed. Cir. 1988). Moreover, the Petition's dependent-claim theories do not cure the deficiencies in the independent claims.

Claims 2-4, 7, and 10 depend from Claim 1; Claims 12-13 and 16 depend from Claim 11; and Claims 21-22 depend from Claim 19. Petitioner must first show the claimed base pipeline, including the training and application of the first machine-learning algorithm, the non-text linguistic representation of phonemes associated with a first pronunciation, the synthesis using a second machine-learning algorithm, and the mapping to different phonemes in an updated pronunciation. It has not done so.

Accordingly, Ground 1 fails for all claims challenged in that ground.

B. Grounds 2-4 Fail on the Merits

1. Grounds 2-4 Fail Because They Rely on Ground 1

Grounds 2-4 fail at least because they depend on Ground 1's defective base combination. Ground 2 adds Feinauer for Claims 8-9, 17-18, and 20; Ground 3 adds Osuna for Claims 5 and 14; and Ground 4 adds Dalce for Claims 6 and 15. Pet. at VI-VIII.

Because Petitioner has not shown that Ground 1 satisfies independent Claims 1, 11, and 19, Petitioner cannot show that the dependent claims are unpatentable in Grounds 2-4. The additional references are used only for additional dependent limitations. **They do not cure Ground 1's failures** regarding mapping as well as alignment and classification of frames. The Board can deny Grounds 2-4 because they depend on Ground 1.

2. Ground 4 Fails Because Neither Fan Alone nor Dalce in View of Fan Teaches the Missing User Input Selection Limitation

Ground 4 challenges Claims 6 and 15 and adds Dalce for manual selection of a target dialect teaching. The Petition advances two theories: that Fan alone teaches the limitation, and, “[t]o the extent it is argued Fan does not teach both” selections, that Dalce supplies them. Pet. at 75. Both theories fail. **Under either theory, Ground 4 does not offer any fix for the failures of Ground 1 discussed above.**

In any event, Fan **does not teach** two user inputs each indicating a selection of an accent. The Petition maps the limitation onto Fan's dialect-mode interface: selecting “Add accent” is said to be the first user input selecting the first accent and choosing from the “Select a place” menu the second input selecting the second accent. Pet. at 73-75. But “Add accent” is a selection of an *operation*—a conversion mode—**not** “a user input indicating a selection of the first accent.” In Fan's add-accent mode, the input accent is Mandarin by system design: Fan teaches that Mandarin is the assumed base from which an accent is added (Ex1004 ¶[0120]), so the user never selects it; the mode *presupposes* it. See Ex1003 ¶227 (showing Ex1004 Fig. 11A GUI presenting no option for selecting an input accent). The Petition attempts to remedy this gap with the expert's inference that, because Fan lacks accent-recognition software, “the user selection must specify

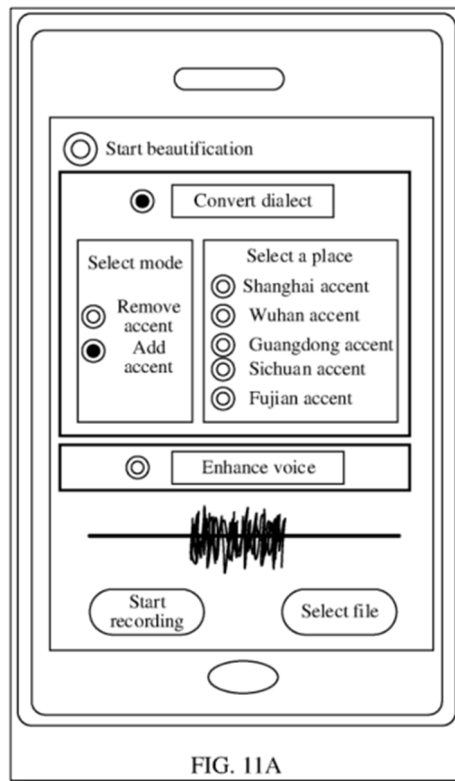
both the input and output accents.” Ex1003 ¶227. That inference is contradicted by Fan’s disclosed workflow; Fan confirms the opposite of the Petition’s inference, as the system presupposes one accent based on the selected mode and allows the user to select only the other. **Implementing Fan’s teachings would not result in the claimed invention.**

With respect to the Dalce fallback theory, this reference’s dialect-selection mechanism is structurally and functionally different from Fan’s accent/dialect selection. Dalce is a universal language translator whose pipeline recognizes the speaker's words, **translates them**—substituting vocabulary and restructuring grammar and syntax into the target language or dialect. Ex1007 Abstract, ¶¶[0016]-[0017]. Its dialect selection determines the language or language variant into which the system translates the speaker’s words and then synthesizes *new* speech mimicking the speaker's auditory attributes. **It is not about accent conversion; it is about translation.**

Fan’s selection mechanism operates in a different setting. Fan performs same-language dialect or **accent conversion**, such as adding or removing a regional accent, while preserving the content and style of the input voice. Ex1004 Abstract, ¶¶[0016]-[0019], [0120]-[0127].

Neither of the Petition's two motivations for incorporating Dalce holds, and accordingly, implementing Dalce’s teachings **would not result in the claimed**

invention. The first—that Fan's users may not always know the base accent and might select the wrong conversion model (Pet. at 76; Ex1003 ¶[231])—is unsupported speculation. Nothing in Fan, Dalce, or any other record evidence identifies any user confusion; Fan's own interface presents the choice in plain terms (“Remove accent” “Add accent” “Select a place” Ex1004 Figs. 11A-11B).



The second motivation—that Fan's system would be improved by letting the user “simply identif[y] the input accent and the target accent”—is no better. The Petition mischaracterizes Dalce’s teachings. While it is cited for its manual selection disclosure, the reference actually teaches the opposite design direction. Dalce's usability improvement is *automation*: its voice recognition module

“preferably **detects** the speaking language and/or the speaking dialect of the speaker” **automatically**, with manual selection merely an interface that “may be provided”. Ex1007 ¶[0013]. Its target dialect is selected **by device location** or **by auto-detecting the listener's dialect**, with a user's manual selection listed last. Ex1007 ¶[0015]. There is no motivation to modify Dalce’s preferred automated dialect selection to a less preferred manual selection as proposed by the Petition.

A PHOSITA motivated by Dalce to improve the usability of Fan would therefore have moved toward *eliminating* manual accent selections through automatic detection. The Petition erroneously reached the limitation only by lifting out of context Dalce's expressly non-preferred fallback—**one Dalce reserves for a scenario Fan does not present**—because it matches the claim language.

Thus, the combination’s failure rests on the relationship between user input and model selection. Fan’s interface selects a conversion mode and, at most, a target place/accent; Dalce’s interface selects language/dialect information for a translation pipeline. Those selections control different operations, and appending Dalce's selection interface onto Fan **would not make Fan perform the claimed first-accent and second-accent selections**, because Fan's source accent remains presupposed by mode, not selected by the user. Furthermore, a PHOSITA motivated by Dalce to improve Fan's usability would therefore have moved toward eliminating manual accent selection through automatic detection—away from, not

toward, the two manual selections the claims require. The Petition reaches the limitation only by lifting Dalce's expressly non-preferred fallback because it matches the claim language, which is impermissible hindsight rather than a teaching that would have led a PHOSITA to the claimed invention.

3. Ground 3 Fails Because Osuna's Accent Conversion Pipeline Is Architecturally Incompatible with Fan's Seq2seq Neural Network

Ground 3 challenges Claims 5 and 14 and adds Osuna for audio data of a particular accent corresponding to a single speaker. The Petition's showing rests on a mischaracterization of what Osuna is, asserts a motivation that contradicts both Fan's stated design objective and the Petition's own reasoning elsewhere, and supports its expectation of success with a single conclusory sentence. **Ground 3 does not offer any fix for the failures of Ground 1 discussed above.**

In any event, the Petition misdescribes Osuna, asserting that “Osuna teaches ‘accent conversion’ using a neural network to perform non-native to native accent conversion.” Pet. at 70. Osuna's conversion model is a Gaussian Mixture Model: it “train[s] a Gaussian Mixture Model (GMM) to model the joint distribution of source and target Mel-Cepstral Coefficients (MCEPs),” then maps source to target features “using maximum likelihood estimation of spectral parameter trajectories.” Ex1009 at 1. However, **Osuna is express about why it does *not* use a neural network for conversion:** “We use a GMM instead of a DNN in this study to focus

on low-resource accent conversion scenarios—in real pronunciation training applications, we generally have a limited amount of data from the non-native speakers.” Ex1009 at 5. The only neural network in Osuna is used to extract PPGs to determine frames (Ex1009 at 3-4), and this component corresponds, if to anything in Fan, to Fan's first machine-learning algorithm (the voice feature extraction model), **not** to the dialect conversion model at issue.

The Petition's related observation that, “[s]imilar to Fan, Osuna utilizes PPGs,” is equally superficial. Pet. at 70.

Osuna and Fan both mention PPGs, but they use them in different ways. Osuna uses PPGs as a matching tool: it compares speech frames from two speakers and pairs frames that sound phonetically similar before training a separate conversion model. Ex1009 at 3. Fan, by contrast, uses PPGs as the input to the dialect-conversion model itself: Fan feeds the PPG feature into a seq2seq neural network, which generates Mel features that are then converted into playable speech. Ex1004 ¶[0126]. Accordingly, Osuna uses PPGs to match data, while Fan uses PPGs to drive the neural model. The Petition therefore relies only on the fact that both references mention PPGs, without explaining why Osuna's frame-matching use of PPGs would improve Fan's different neural conversion pipeline

The Petition's assertion of using a “known technique to improve a similar device” (Pet. at 72) thus rests on the mere presence of PPGs in Osuna, without

explaining why Osuna's frame-matching use of PPGs would improve Fan's different neural conversion pipeline.

4. Ground 2 Fails Because Feinauer's Real-Time Architecture is Fundamentally Different from the Architecture of Fan

Ground 2 challenges Claims 8-9, 17-18, and 20 and adds Feinauer for its real-time data packet receipt and transmission. Feinauer's real-time conversion is tightly coupled to its context-selection architecture. Its system does **not** simply convert from one accent to another based on a user-selected target; rather, it detects accents and dialects, identifies context, selects an optimal target dialect/accents based on that context, determines whether the first dialect/accents differs from the target by a threshold, and modifies appropriate portions of the voice communication. Ex1008 Abstract, ¶¶[0007]-[0015], [0035]-[0040], [0053]-[0057].

Ground 2 does not offer any fix for the failures of Ground 1 discussed above.

In any event, implementing Feinauer **would not result in the claimed invention**. Fan lacks the infrastructure needed to implement Feinauer's approach. Fan does **not** perform automatic accent detection, does **not** select a target accent based on conversational context, and does **not** use threshold-based gating to decide whether or where to modify speech. Instead, Fan implements a conversion mode and feeds PPG features and optionally style/timbre features through a learned

dialect conversion model. Ex1004 ¶¶[0055]-[0060], [0120]-[0127]. Through this process, input speech representations are transformed into a target accent.

Additionally, the Petition attempts to isolate Feinauer's real-time data packet receipt and transmission and graft it onto Fan. However, Feinauer's real-time operation is **not a free-floating module**; it is **embedded in a broader selective-modification system** that determines target accent, modification necessity, and modification implementation.

The Petition does not explain how Fan's seq2seq dialect conversion model could be restructured to perform Feinauer's selective, threshold-driven, context-dependent modification. This is because the combination lacks the decision architecture of the conversion pipeline. Feinauer's real-time packet handling depends on a context-aware selection-and-modification framework that is absent from Fan's seq2seq dialect conversion model. Without that missing decision infrastructure, Feinauer's packet handling cannot simply be bolted onto Fan as an operable real-time conversion module. Ground 2 therefore fails.

VI. CONCLUSION

Ground 1 is the foundation of the Petition, and Petitioner does not show a supported motivation to combine Fan, [Sonsaat](#), and [Loots](#) or a reasonable expectation of success. Because Grounds 2-4 depend on Ground 1's defective base combination, those grounds fail as well. Furthermore, Grounds 2-4 each fail for their own defective combinations.

For the foregoing reasons, Patent Owner respectfully requests that the Board deny institution of inter partes review of Claims 1-22 of the '550 Patent.

Respectfully submitted,

AMSTER, ROTHSTEIN & EBENSTEIN LLP
Attorneys for Patent Owner
405 Lexington Avenue, 48th Floor
New York, NY 10174
(212)336-8000

Dated: June 30, 2026
New York, New York

By: /Charles R. Macedo/
Charles R. Macedo (Reg. No. 32,781)

Of counsel:
Gary J. Gershik (Reg. No. 39,992)
Lewis Derenzis III (Reg. No. 84,132)

CERTIFICATE OF WORD COUNT

The undersigned certifies that the foregoing PATENT OWNER'S PRELIMINARY RESPONSE complies with the type-volume limitation in 37 C.F.R. § 42.24(b)(1). According to the word-processing system's word count, the brief contains approximately **10,299** words, excluding the parts exempted by 37 C.F.R. § 42.24(a).

Respectfully submitted,

AMSTER, ROTHSTEIN & EBENSTEIN LLP
Attorneys for Patent Owner
405 Lexington Avenue, 48th Floor
New York, NY 10174
(212)336-8000

Dated: June 30, 2026
New York, New York

By: /Charles R. Macedo/
Charles R. Macedo (Reg. No. 32,781)

CERTIFICATE OF SERVICE

I hereby certify that on June 30, 2026, a true and correct copy of the foregoing PATENT OWNER'S PRELIMINARY RESPONSE was served via email, by consent, to Petitioner's counsel of record at the correspondence addresses of record in this proceeding.

Eric A. Buresh (Reg. No. 50,394)
Chris R. Schmidt (Reg. No. 63,982)
Nick R. Apel (Reg. No. 84,017)
Adam M. Sandwell (Reg. No. 72,484)
Erise IP, P.A.
7015 College Blvd., Suite 700
Overland Park, Kansas 66211
Telephone: (913) 777-5600
Fax: (913) 777-5601
chris.schmidt@eriseip.com
nick.apel@eriseip.com
adam.sandwell@eriseip.com
PTAB@eriseip.com

Respectfully submitted,

AMSTER, ROTHSTEIN & EBENSTEIN LLP
Attorneys for Patent Owner
405 Lexington Avenue, 48th Floor
New York, NY 10174
(212) 336-8000

Dated: June 30, 2026
New York, New York

By: /Charles R. Macedo/
Charles R. Macedo (Reg. No. 32,781)