

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION,
Petitioner,

v.

DIALECT, LLC,
Patent Owner.

IPR2025-01229
Patent: 7,634,409
Issued: December 15, 2009
Application No. 11/513,269
Filed: August 31, 2006

Title: DYNAMIC SPEECH SHARPENING

PETITION FOR
INTER PARTES REVIEW OF U.S. PATENT NO. 7,634,409

TABLE OF CONTENTS

	Page
LIST OF EXHIBITS.....	v
MANDATORY NOTICES UNDER 37 C.F.R. § 42.8	viii
1. Real Party-In-Interest.....	viii
2. Related Matters	viii
3. Lead And Back-Up Counsel, And Service Information.....	viii
I. INTRODUCTION	1
II. GROUNDS FOR STANDING PER 37 C.F.R. § 42.104(A).....	2
III. IDENTIFICATION OF CHALLENGE	2
A. Statement Of The Precise Relief Requested / Statutory Grounds.....	2
IV. LEVEL OF SKILL IN THE ART AND STATE OF THE ART.....	2
A. Person Of Ordinary Skill In The Art.....	2
B. State Of The Art	3
1. Linguistic Units In Speech Recognition	3
a) Phones and Phonemes Were Known As Interchangeable Acoustic Elements In Speech Recognition Systems	4
b) Syllables (Each Having An Onset, Nucleus, and Coda) Were Known And Used Widely In Speech Recognition.....	5
2. Automatic Speech Recognition (“ASR”) Systems	6
3. Lexicons and Probabilistic Grammars Were Well Known In ASRs.....	9
4. Finite State Transducers (“FSTs”) Were Well Known.....	10

V.	THE '409 PATENT	11
A.	The '409 Patent's Specification	12
B.	The Prosecution History	13
1.	Prosecution Of The '409 Patent	13
2.	European Counterpart Prosecution	14
C.	Claims Listing	17
VI.	CLAIM CONSTRUCTION	17
A.	Claim 1 Preamble	18
B.	“acoustic grammar”	20
VII.	GROUND 1: CLAIMS 1, 2, 3, AND 6 ARE OBVIOUS OVER BAZZI	22
A.	Bazzi	22
1.	Claim 1	24
a)	Element [1.1]	24
b)	Element [1.2]	27
c)	Element [1.3]	29
d)	Element [1.4]	36
e)	Element [1.5]	44
2.	Claim 2	45
3.	Claim 3	47
4.	Claim 6	49
a)	Element [6.1]	49
b)	Element [6.2]	51
c)	Element [6.3]	53

VIII. GROUND 2: CLAIMS 2 & 3 ARE OBVIOUS OVER BAZZI IN FURTHER VIEW OF SABOURIN.....	55
A. Sabourin.....	55
B. The Bazzi-Sabourin Combination.....	58
1. Claim 1	60
2. Claim 2	60
1. Claim 3	61
IX. GROUND 3: CLAIM 6 IS OBVIOUS OVER BAZZI IN FURTHER VIEW OF EPSTEIN	63
A. Epstein	63
A. The Bazzi-Epstein Combination	64
1. Claim 1	68
2. Claim 6.....	68
a) Element [6.1]	68
b) Element [6.2]	69
c) Element [6.3]	69
X. NO OBJECTIVE INDICIA OF NON-OBVIOUSNESS.....	70
XI. CONCLUSION.....	71
CERTIFICATE OF COMPLIANCE.....	72
CERTIFICATE OF SERVICE	73

TABLE OF AUTHORITIES

Page(s)

Cases

<i>Bristol-Myers Squibb Co. v. Ben Venue Lab 'ys, Inc.</i> , 246 F.3d 1368 (Fed. Cir. 2001)	18
<i>Catalina Mktg. Int'l, Inc. v. Coolsavings.com, Inc.</i> , 289 F.3d 801 (Fed. Cir. 2002)	18
<i>Eli Lilly & Co. v. Teva Pharms. Int'l GmbH</i> , 8 F.4th 1331 (Fed. Cir. 2021)	20
<i>In re Hirao</i> , 535 F.2d 67 (CCPA 1976)	18
<i>Novartis AG v. Torrent Pharms. Ltd.</i> , 853 F.3d 1316 (Fed. Cir. 2017)	70
<i>Realtime Data, LLC v. Iancu</i> , 912 F.3d 1368 (Fed. Cir. 2019)	22
<i>Vivid Techs., Inc. v. Am. Sci. & Eng'g, Inc.</i> , 200 F.3d 795 (Fed. Cir. 1999)	22

Statutes

35 U.S.C. § 102	22, 56, 63
35 U.S.C. § 103	2

Other Authorities

37 C.F.R. § 42.104	2
37 C.F.R. § 42.24	72
37 C.F.R. § 42.6	73

LIST OF EXHIBITS

No.	Description
1001	U.S. Patent No. 7,634,409 (“ 409 patent ”)
1002	File History of U.S. Patent No. 7,634,409
1003	Declaration of Paul Jacobs, dated July 18, 2025 (“Jacobs”)
1004	Provisional U.S. Patent Application No. 60/712,412
1005	PCT Patent Application No. WO 2007/027989 A2 to Di Cristo et al.
1006	EP Patent Application No. 06814053.2 to Robert A. Kennewick
1007	Claim Comparison Chart Of '409 Patent Claims 1-3 And 6 Against May 14, 2008, Amended Claims 9-11 And 14 of EP Application No. 06814053.2
1008	Bazzi, I., & Glass, J., <i>Heterogeneous Lexical Units For Automatic Speech Recognition: Preliminary Investigations</i> , Proc. of 2000 IEEE Int'l Conf. of Acoustics, Speech, and Signal Processing, 1257-1260 (2000) (“ Bazzi ”)
1009	Huang, X., et al., <i>Spoken Language Processing – A Guide to Theory, Algorithm, and System Development</i> , Prentice-Hall, Inc. Publ. (2001) (excerpts) (“ Huang ”)
1010	Jurafsky, D., & Martin, J., <i>Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition</i> , Prentice-Hall, Inc. Publ. (2000) (excerpts) (“ Jurafsky ”)
1011	U.S. Patent No. 7,818,176 to Freeman et al. (“ 176 patent ”)
1012	Claim Construction Order, <i>VB Assets, LLC v. Amazon.com, Inc.</i> , No. 19-1410 (MN) (D. Del. June 23, 2021)
1013	Institution Decision, IPR2020-01390, Paper 7 (PTAB Mar. 11, 2021)

No.	Description
1014	Livescu, K. et al., <i>Subword Modeling for Automatic Speech Recognition: Past, Present, and Emerging Approaches</i> , IEEE Signal Proc. Mag. 29:6, 44-57 (2012)
1015	Schmandt, C., <i>Voice Communications With Computers</i> , Int'l Thomson Publ. (excerpts) (1994)
1016	Ostendorf, M., & Roukos, S., <i>A Stochastic Segment Model For Phoneme-Based Continuous Speech Recognition</i> , IEEE Transactions On Acoustics, Speech, And Signal Proc., 1857-1869, 37:12 (Dec. 1989)
1017	Ravishankar, M., <i>Efficient Algorithms for Speech Recognition</i> , Thesis, CMU-CS-96-143, Carnegie Mellon Univ. (May 15, 1996)
1018	Rabiner, L., & Juang, B., <i>Fundamentals of Speech Recognition</i> , Prentice-Hall, Inc. Publ. (1993) (excerpts)
1019	U.S. Patent No. 6,085,160 to D'hoore et al.
1020	U.S. Patent No. 6,154,722 to Jerome R. Bellegarda
1021	U.S. Patent No. 7,146,319 to Melvyn J. Hunt
1022	U.S. Patent Application Pub. No. 2004/0186714 to James K. Baker
1023	U.S. Patent No. 5,806,032 to Richard William Sproat
1024	U.S. Patent No. 6,108,627 to Michael Sabourin (" Sabourin ")
1025	U.S. Patent Application Pub. No. 2005/0055209 A1 to Epstein et al. (" Epstein ")
1026	Schwartz, R., & Chow, Y., <i>The N-Best Algorithm: An Efficient And Exact Procedure For Finding The N Most Likely Sentence Hypotheses</i> , IEEE Int'l Conf. of Acoustics, Speech, and Signal Processing, 81-84 (1990)
1027	U.S. Patent No. 5,241,619 to Richard M. Schwartz
1028	Mohri, M., et al., <i>Weighted Finite-State Transducers In Speech Recognition</i> , Computer Speech and Language, 16, 69-88 (2002)

No.	Description
1029	Declaration of Gordon MacPherson, signed June 23, 2025
1030	Proceedings, 2000 IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (June 2020) (excerpts)
1031	Appendix of Challenged Claims 1-3 and 6

MANDATORY NOTICES UNDER 37 C.F.R. § 42.8

1. Real Party-In-Interest

Microsoft Corporation is the sole real party-in-interest.

2. Related Matters

The '409 patent (EX1001) has been asserted in the following district court litigation (the “district court case”):

- *Dialect, LLC v. Microsoft Corporation*, Case No. 2:24-cv-01067 (E.D. Tex.), filed December 20, 2024.

3. Lead And Back-Up Counsel, And Service Information

Lead Counsel	Back-up Counsel
Andrew M. Mason Reg. No. 64,034 andrew.mason@klarquist.com	Cameron D. Clawson, Reg. No. 73,509 cameron.clawson@klarquist.com Amy Haspel, Reg. No. 78,385 amy.haspel@klarquist.com Frank Morton-Park, Reg. No. 80,750 frank.morton-park@klarquist.com
KLARQUIST SPARKMAN, LLP 121 SW Salmon Street, Suite 1600 Portland, Oregon, 97204 503-595-5300 (phone) 503-595-5301 (fax)	

Petitioner consents to service via email at the above email addresses and the email address of Msft-Dialect@klarquist.com.

Pursuant to 37 C.F.R. § 42.10 (b), concurrently filed with this Petition is a Power of Attorney executed by Petitioner and appointing the above counsel.

Petitioner authorizes Account No. 02-4550 to be charged for any fees, including those enumerated in 37 C.F.R. § 42.15.

I. INTRODUCTION

Microsoft Corporation (“Petitioner”) respectfully requests *inter partes* review (“IPR”) of claims 1-3 and 6 of U.S. Patent No. 7,634,409 (“the ’409 patent”) (EX1001), allegedly assigned to Dialect, LLC (“Patent Owner”). For the reasons set forth below, these claims should be found unpatentable and cancelled.

The ’409 patent relates to speech interpretation methods that involve recognizing basic sub-word units that correspond to distinct speech sounds (known in the art as “phonemes”) and applying a grammar to map the recognized phonemes to syllables. The ’409 patent was subject to no substantive rejections during prosecution and the Examiner allowed the claims because the cited art failed to teach mapping phonemes to syllable grammars.

The grounds presented herein rely on prior art Bazzi (EX1008), which was not before the Examiner during prosecution. Bazzi *was* cited, however, during subsequent prosecution of a European counterpart that included claims with elements nearly identical to ’409 patent claims 1, 2, 3, and 6 challenged in this Petition. The EPO rejected those nearly-identical claims based on Bazzi, finding Bazzi disclosed matching phonemes against syllable grammars (the element that had led to allowance of the challenged claims). In the European proceeding, the applicants failed to distinguish Bazzi, leading to eventual abandonment of the

European counterpart. As shown below, and consistent with the findings of the EPO, Bazzi renders all challenged claims obvious.

II. GROUNDS FOR STANDING PER 37 C.F.R. § 42.104(a)

Petitioner certifies that the '409 patent is available for IPR and that Petitioner is not barred or estopped from requesting an IPR challenging the patent claims on the grounds identified in this petition.

III. IDENTIFICATION OF CHALLENGE

A. Statement Of The Precise Relief Requested / Statutory Grounds

Petitioner requests *inter partes* review of claims 1-3 and 6 (the “Challenged Claims”) of the '409 patent, on the following statutory grounds:

	Reference(s)	Basis	Claims
Ground 1	Bazzi (EX1008)	35 U.S.C. § 103	1-3, 6
Ground 2	Bazzi and Sabourin (EX1024)	35 U.S.C. § 103	2, 3
Ground 3	Bazzi and Epstein (EX1025)	35 U.S.C. § 103	6

The Petition presents evidence showing a reasonable likelihood that the Petitioner will prevail in establishing that each Challenged Claim is unpatentable.

IV. LEVEL OF SKILL IN THE ART AND STATE OF THE ART

A. Person Of Ordinary Skill In The Art

The person of ordinary skill in the art in August 2006 (“POSITA”) would have had a bachelor’s degree in electrical engineering, computer science, computer

engineering, or equivalent field, and two years of experience working with speech recognition and natural language processing systems. Additional work experience could make up for less education and vice versa. This definition would not differ meaningfully were the date in question August 2005. Jacobs, ¶36.

B. State Of The Art

Certain concepts and processes reflect the state of the art for automatic speech recognition systems during the relevant 2005-2006 time period, as described in textbooks such as Huang (EX1009, published 2001) and Jurafsky (EX1010, published 2000). Jacobs, ¶37.

1. Linguistic Units In Speech Recognition

Automatic Speech Recognition (ASR) systems aim to identify the most likely sequence of words corresponding to a given speech input, accounting for uncertainties such as pronunciation variability, ambient noise, and spontaneous speech disfluencies (e.g., spoken “uhs” and “ums,” stutters, and word repetitions). *See* EX1010, 194-95; EX1009, xxii. “Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics.... Each phoneme has distinguishable acoustic characteristics and, in combination with other phonemes, forms larger units such as syllables and words.” EX1009, xxii. Jacobs, ¶38.

a) **Phones and Phonemes Were
Known As Interchangeable Acoustic
Elements In Speech Recognition Systems**

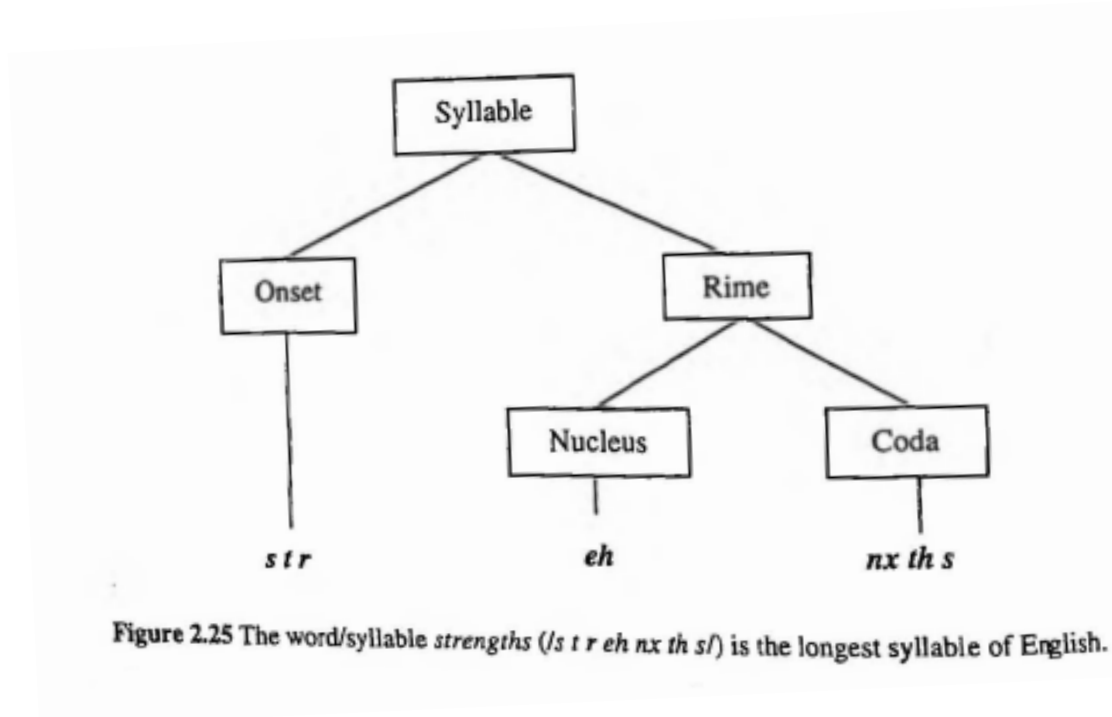
The basic acoustic-linguistic segment in speech recognition is the phoneme. EX1009, 37. (“In speech science, the term *phoneme* is used to denote any of the minimal units of speech sound in a language that can serve to distinguish one word from another.”). A phoneme is either a consonant or a vowel, with the English language containing 16 vowel and 24 consonant phonemes. *Id.*, xxii. When spoken, a phoneme may be articulated differently based on its surrounding phonemes, and the spoken articulation of a given phoneme is called a *phone*. *Id.*, 37 (“We conventionally use the term *phone* to denote a phoneme’s acoustic realization.”); *see also* EX1010, 104; Jacobs, ¶39.

As a *phone* is just an articulation of a given *phoneme*, POSITAs used and understood the terms “phoneme” and “phone” interchangeably, as reflected in the prior art literature. *See* EX1015, 15, (“**For most practical purposes, phone and phoneme may be considered to be synonyms.**”) (emphasis added); EX1009, 37 (“We will use the terms phone or phoneme interchangeably to refer to the speaker-independent and context-independent units of meaningful sound contrast.”); EX1014, 45 (“In speech recognition research, these terms [phone and phoneme] are often used interchangeably, and recognition dictionaries often include a mix of phones and phonemes.”); EX1016, 1858 (“In this paper, we are not rigorous in

distinguishing between the two terms phone and phoneme, which are used interchangeably.”); EX1017, 3 (“The *phoneme* (or phone) has been the most commonly accepted sub-word unit.”); Jacobs, ¶40.

**b) Syllables (Each Having An Onset, Nucleus, and Coda)
Were Known And Used Widely In Speech Recognition**

To further constrain recognition and improve interpretability, speech recognition systems may segment phoneme/phone streams into syllables. Syllables are intermediate sub-word units “that interpose between the phones and the word level.” EX1009, 51. Syllables are comprised of an onset, nucleus, and coda, which refer to the central vowel or vowels (nucleus) and its preceding (onset) and following (coda) consonants. *See, e.g.*, EX1010, 102 (“A syllable is usually described as having an optional initial consonant or set of consonants called the onset, followed by a vowel or vowels, followed by a final consonant or sequence of consonants called the coda.”). Huang provides an example segmenting the syllable “strengths” into the phonemes corresponding to its onset, nucleus, and coda components:



EX1009, 52, Fig. 2.25; *see also* EX1001, 2:50-52 (“Portions of a word may be represented by a syllable, which may be further broken down into core components of an onset, a nucleus, and a coda.”). Huang recognizes that “syllables are often used” as subword models for large-vocabulary speech recognition systems. EX1009, 608; *see also* EX1018, 436-37 (describing using syllables as one of “several possible choices for subword units that can be used to model speech”); Jacobs, ¶41.

2. Automatic Speech Recognition (“ASR”) Systems

As described in Huang, a typical speech recognizer comprises a computing platform including components for input signal processing and a decoder module driven by acoustic and language models, as shown in Huang’s Figure 1.2, reproduced below. EX1009, 5, Fig. 1.2; Jacobs, ¶42.

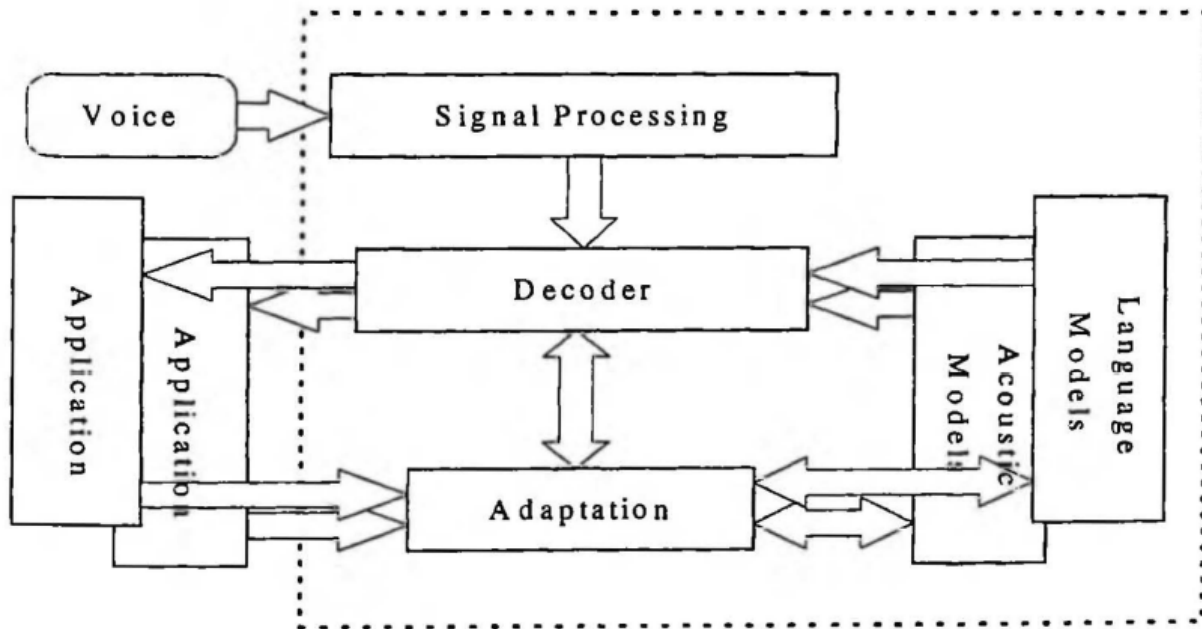
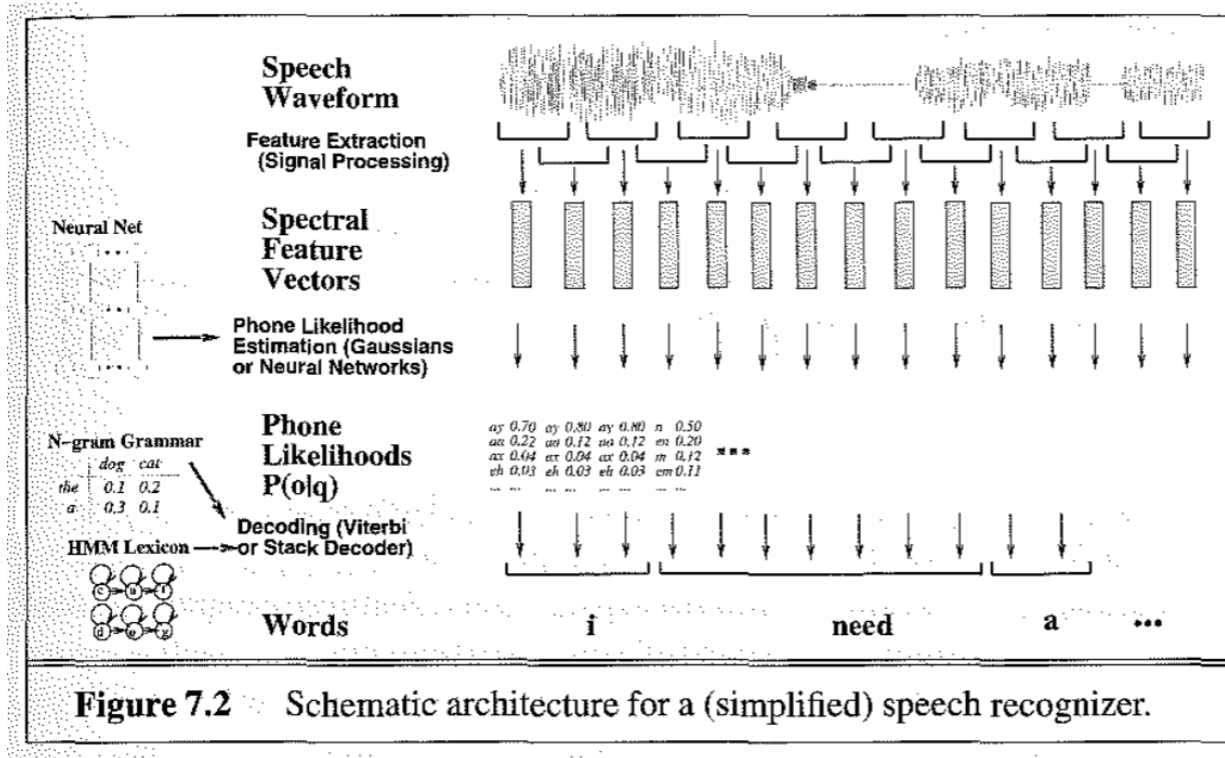


Figure 1.2 Basic system architecture of a speech recognition system [12].

ASR systems are typically structured around a processing pipeline involving steps of signal processing, phonetic analysis, and decoding. *See* EX1010, 240-41, Fig. 7.2. Jurafsky depicts these functional stages of a typical ASR system in Figure 7.2, reproduced below.



Id., 241. These functional stages are described in more detail below. *See also* Jacobs, ¶43.

Signal Processing – In the first stage, signal processing, the acoustic signal is received as a waveform “which [is] transformed into spectral features which give information about how much energy in the signal is at different frequencies. EX1010, 240.

Subword Recognition – In the second stage, subword or phone recognition, statistical techniques are used “to tentatively recognize individual speech sounds” and thereby identify the probabilities of what subword units are present in each time frame of the input signal. *Id.*

Decoding – In the final stage, decoding, the ASR system combines the sub-word probability data determined in the previous stage with “a dictionary of word pronunciations and a language model (probabilistic grammar)” using a decoding algorithm “to find the sequence of words which has the highest probability given the acoustic events.” *Id.*, 241.

3. Lexicons and Probabilistic Grammars Were Well Known In ASRs

As explained in the preceding sub-section, typical ASR systems identify a sequence of words corresponding to an input utterance by analyzing probable sub-word units in combination with two data structures: 1) a pronunciation dictionary (also called a lexicon) and 2) a probabilistic grammar (also called a language model). *See* EX1010, 270 (describing how a typical ASR decoder “takes 3 inputs (the observation likelihoods . . . , the HMM lexicon, and the *N*-gram language model) and produces the most probable string of words”). Jacobs, ¶44.

A pronunciation dictionary maps words to their pronunciations as a set of phones/phonemes. EX1010, 135 (“[Pronunciation dictionaries] give the pronunciation of words as strings of phones, sometimes including syllabification and stress.”). The most common data structure type for representing a pronunciation dictionary is the Hidden Markov Model (HMM), which is built “by taking an off-

the-shelf pronunciation dictionary” and mapping each phone/phoneme in the dictionary to a state in the HMM data structure. *Id.*, 272; Jacobs, ¶45.

The probabilistic grammar, meanwhile, models the phonotactic constraints of a language. That is, sub-word units such as phonemes and syllables in natural languages are not freely combinable; languages impose phonotactic rules that restrict the permissible sequences of sub-word units. Likewise, given words in a language are more or less likely to follow other words. *See* EX1010, 191-92. (“Guessing the next word (or word prediction) is an essential subtask of speech recognition.... [L]ooking at previous words can give us an important cue about what the next ones are going to be.”). Language models are commonly modeled using a data structure called an N-gram. “An *N*-gram model uses the previous *N*-1 words to predict the next one.” *Id.*, 193. For example, a word “bigram” models the probability of two words being in sequence while a “trigram” models the probability for three sequential words. *Id.*, 197-98. Such language models help reduce ambiguity, provide tolerance for word misrecognitions caused by noise, and improve recognition accuracy by favoring more probable phoneme sequences and by pruning likely invalid phoneme sequences from the search space. Jacobs, ¶46.

4. Finite State Transducers (“FSTs”) Were Well Known

A finite-state transducer (FST) is a mathematical model used to represent mappings between sequences—typically in the form of input-output symbol pairs,

where each mapping may be associated with a cost or weight (called a weighted FST). EX1028, 69. “[A] path through the transducer encodes a mapping from an input symbol sequence to an output symbol sequence.” *Id.* In the context of speech recognition, FSTs are used to model and combine various components of the recognition process in a unified framework. FSTs allow for probabilistic transitions from one state to another and support “composition,” i.e., the combination of probabilities and constraints among multiple FSTs. EX1028, 69-70, 73. FSTs are therefore widely employed to encode and combine probabilistic models such as pronunciation lexicons, language models, and observed acoustic feature probabilities. *Id.*; Jacobs, ¶47. These components are represented as individual FSTs and then composed together into a single search graph, which is searched using an appropriate search algorithm. *See* EX1008, 1258(1:24-37). This integrated transducer enables the system to efficiently evaluate possible interpretations of a speech input and identify the most likely hypothesis. Jacobs, ¶47.

V. THE '409 PATENT

The '409 patent, titled “Dynamic speech sharpening” issued December 15, 2009, from Application No. 11/513,269 (the “'269 application”), filed August 31, 2006. While the '409 patent claims the benefit of a provisional application filed Aug.

31, 2005, that provisional application does not support claim 1 of the '409 patent.¹ Thus, the challenged claims are not entitled to a priority date earlier than August 31, 2006.

A. The '409 Patent's Specification

The '409 patent describes a system for interpreting natural language speech. Jacobs, ¶¶29-31. The system is configured for “receiving a user verbalization,” “identifying one or more phonemes in the verbalization,” and using an “acoustic grammar ... to map the phonemes to syllables” to generate “preliminary interpretations of the verbalization.” EX1001, Abstract.

The '409 patent states that its acoustic grammar may represent the phonotactic rules of the English language. For example, the patent describes that for an “acoustic grammar representing the phonotactic rules of English” syllables may be “divided into core components of an onset, a nucleus, and a coda.” *Id.*, 6:21-26. The patent further describes that the phonotactic constraints for a given acoustic model may be used to restrict the “available transitions” between acoustic elements. *Id.*, 7:26-29.

¹ For example, the provisional application does not mention phonemes or syllables (EX1004) and thus cannot support the challenged claims' method that uses “phonemes” and “syllables.” Jacobs, ¶ 28.

The system includes a module that generates multiple candidate interpretations of the phoneme sequence and assigns to each candidate a “confidence or interpretation score ... representing a likelihood that a particular candidate interpretation is a correct interpretation of the verbalization.” *Id.*, 9:15-19. Based on the scores, the system selects the candidate with the highest or lowest value as a probable interpretation of the utterance. *Id.*, 9:19-23.

B. The Prosecution History

1. Prosecution Of The '409 Patent

The '269 application was filed on August 31, 2006. EX1002, 1-38. No substantive action occurred until January 23, 2008, when the Applicant submitted a preliminary amendment, cancelling all then-pending claims (1-45) and adding nineteen new claims (46-64). *Id.*, 131-37. On June 3, 2009, the Examiner and Applicant conducted a telephone interview wherein the Applicant made a provisional election to prosecute pending claims 46-61 and withdraw claims 62-64. *Id.*, 146-49 (summarizing interview). On June 22, 2009, the Examiner issued a notice for allowance for claims 46-61. *Id.*, 142-45. The examiner provided the following reasons for allowance, distinguishing the allowed claims over U.S. Patent No. 7,146,319 to Hunt:

Hunt fails to specifically disclose mapping the recognized stream of phonemes to an acoustic grammar that

phonemically represents one or more syllables, the recognized stream of phonemes mapped to a series of one or more of the phonemically represented syllables; and wherein the generated interpretation includes the series of syllables mapped to the recognized stream of phonemes.

In other words, Hunt fails to teach matching phonemes against syllable grammars.

Id., 148 (emphasis added); Jacobs, ¶32.

On July 2, 2009, the Applicant submitted a Request for Continued Examination in conjunction with an Information Disclosure Statement. *Id.*, 182-83. On August 24, 2009, the Examiner issued a further notice of allowance for claims 46-61 and restated the same reasons for allowance over Hunt. *Id.*, 186-93. The '409 Patent subsequently issued on December 15, 2009. *Id.*, 216.

2. European Counterpart Prosecution

On the '269 application's filing date of August 31, 2006, the Applicant also filed corresponding PCT application no. PCT/US2006/034184 (published March 8, 2007, as WO 2007/027989 A2; EX1005), which shares the same drawings and written description as the '269 application. The PCT application entered the European stage on March 26, 2008, as European application number 06814053.2 ("the EP application,"). EX1006, 156-60. The applicant filed an initial amendment to its claims on May 14, 2008, before initial examination by the EPO. *Id.*, 137-42.

Notably, the as-amended claims 9-11 and 14 pending at the EPO were substantially the same as challenged claims 1-3 and 6 of the '409 Patent, as shown in EX1007, which presents an element-by-element comparison of those claims. As shown in EX1007, every limitation of the '409 Patent's method claims 1-3 and 6 has a substantially identical step in claims 9-11 and 14 of the EPO application as amended on May 14, 2008. EX1007; Jacobs, ¶34. Following that initial amendment, the EPO filed a European search opinion on September 28, 2010, indicating that the examined claims were unpatentable as not novel based, in part, on the Bazzi reference relied on in this Petition. EX1006, 100-04. Specifically, the European search opinion determined that Bazzi disclosed the following:

- A system for providing out-of-vocabulary interpretation capabilities and for tolerating noise when interpreting natural language speech utterances;
- at least one input device that receives an utterance from a user and generates an electronic signal corresponding to the utterance;
- a speech interpretation engine that receives the electronic signal corresponding to the utterance operable to:
 - recognize a stream of phonemes contained in the utterance;
 - map the recognized stream of phonemes to an acoustic grammar that phonemically represents one or more syllables, the

recognized stream of phonemes mapped to a series of one or more of the phonemically represented syllables; and

- generate at least one interpretation of the utterance, wherein the generated interpretation includes the series of syllables mapped to the recognized stream of phonemes.

Id., 102-03.

The Applicant amended the claims on April 12, 2011 (*id.*, 79-98 (cancelling pending claims 1-8, amending pending claims 9-19 and renumbering them to 1-11, and adding new claims 12-15). On May 26, 2014, the EPO issued an examination communication rejecting the pending independent claims over Bazzi and rejecting the Applicant's arguments made in its December 4 amendment that Bazzi "does not disclose using both phoneme and syllable recognition." *Id.*, 67-70.

On November 20, 2014, the Applicant again amended the claims. *Id.*, 59-62. On May 29, 2017, the EPO issued a communication again rejecting the amended claims based on Bazzi and summoning the Applicant to oral proceedings concerning the pending application. *Id.*, 30-38. Oral proceedings were held in the Applicant's absence on October 23, 2017, whereafter the EPO reissued its rejections of the pending claims. *Id.*, 11-16. Finally, the European counterpart application was abandoned and the application was closed on July 3, 2018. *Id.*, 1-2; Jacobs, ¶¶33-35.

C. Claims Listing

Independent claim 1 is listed in the table below, with element numbering added in the lefthand column. A full listing of the challenged claims (including the challenged dependent claims) is submitted as EX1031.

[1.1]	A method for providing out-of-vocabulary interpretation capabilities and for tolerating noise when interpreting natural language speech utterances, the method comprising:
[1.2]	receiving an utterance from a user;
[1.3]	recognizing a stream of phonemes contained in the utterance on an electronic device;
[1.4]	mapping the recognized stream of phonemes to an acoustic grammar that phonemically represents one or more syllables, the recognized stream of phonemes mapped to a series of one or more of the phonemically represented syllables; and
[1.5]	generating at least one interpretation of the utterance, wherein the generated interpretation includes the series of syllables mapped to the recognized stream of phonemes.

VI. CLAIM CONSTRUCTION

For purposes of this IPR only, Petitioner applies the plain and ordinary meaning of all claim terms. Petitioner reserves the right to argue in any district court case or other proceeding that terms in the '409 patent are indefinite or otherwise, and to raise additional issues of claim construction.

A. Claim 1 Preamble

The preamble of claim 1 recites the following: “A method for providing out-of-vocabulary interpretation capabilities and for tolerating noise when interpreting natural language speech utterances.”

Claim 1’s preamble is non-limiting. Generally, a preamble is not limiting unless it recites essential structures or steps, or is “necessary to give life, meaning, and vitality to the claim.” See *Catalina Mktg. Int’l, Inc. v. Coolsavings.com, Inc.*, 289 F.3d 801, 808 (Fed. Cir. 2002). The preamble of claim 1 merely states the intended purpose or result of the claimed method—namely, to “provid[e] out-of-vocabulary interpretation capabilities” and to “tolerate[] noise” in interpreting speech. EX1001, 11:62-12:9 (claim 1).

As explained below, the claimed steps of method claim 1 stand on their own and the preamble’s recitation of intended purpose therefore is unnecessary to give the claim meaning. See *Bristol-Myers Squibb Co. v. Ben Venue Lab’ys, Inc.*, 246 F.3d 1368, 1375–76 (Fed. Cir. 2001) (finding method claim preamble non-limiting where it is “only a statement of purpose and intended result” and “[t]he expression does not result in a manipulative difference in the steps of the claim”); *In re Hirao*, 535 F.2d 67, 70 (CCPA 1976) (“[T]he preamble merely recites the purpose of the process; the remainder of the claim . . . does not depend on the preamble for completeness, and the process steps are able to stand alone.”).

Here, nothing in the body of claim 1 references the goals of tolerating noise or handling out-of-vocabulary input. The claimed method begins with “receiving an utterance” and proceeds through steps involving phoneme recognition and syllabic mapping. These steps are described independently of any functional result like noise tolerance or vocabulary adaptation. Moreover, the specification discusses those results as motivating factors behind the design, not as limitations on the method itself. Specifically, the patent’s Background of the Invention describes purported problems in existing speech recognition systems relying on word-based grammars:

[S]peech interpretation engines still have substantial problems with accuracy and interpreting words that are not defined in a predetermined vocabulary or grammar context. Poor quality microphones, extraneous noises, unclear or grammatically incorrect speech by the user, or an accent of the user may also cause shortcomings in accuracy, such as when a particular sound cannot be mapped to a word in the grammar.

EX1001, 1:65-2:9. The specification proceeds to explain that out-of-vocabulary interpretation and noise reduction is merely an intended benefit that “may” be provided or offered by phoneme recognition. *Id.*, 2:40-44 (“Phoneme recognition may disregard the notion of words, instead interpreting a verbalization as a series of phonemes, which may provide out-of-vocabulary (OOV) capabilities, such as when

a user misspeaks or an electronic capture devices drops part of a speech signal.”; *see also id.*, 6:6-9 (“Phonemic recognition provides several benefits, particularly in the embedded space, such as offering out-of-vocabulary (OOV) capabilities.”). The specification provides no other descriptions of out-of-vocabulary or noise reduction capabilities. Nor does it explicitly identify any particular steps for providing such capabilities. Moreover, the preamble provides no antecedent basis for terms appearing in the body of claim 1 or any dependent claims. Claim 1’s preamble is therefore distinguishable from other method preambles found to be limiting for giving life and meaning to the claimed method steps. *See e.g., Eli Lilly & Co. v. Teva Pharms. Int’l GmbH*, 8 F.4th 1331, 1342-43 (Fed. Cir. 2021) (finding method preamble limiting where it was necessary to give life and meaning to the term “effective amount” recited in the method step and where it provided antecedent basis for later claim term).

Because the preamble language merely expresses aspirational benefits of the method—without being required to interpret or enable any element of the claim body—it is non-limiting and does not affect the scope of the claim. Jacobs, ¶¶48-49.

B. “acoustic grammar”

Claims 1, 2, and 3 of the ’409 patent recite the term “acoustic grammar.” Petitioner submits that this term does not require construction and should be given its plain and ordinary meaning. However, as shown in the table below, the ’409

patent shares the term “acoustic grammar” with a non-family member patent²—filed by the same Applicant (VoiceBox) but assigned to a different Patent Owner (VB Assets)—that has been construed in litigation before the District Court for the District of Delaware³ (the “Amazon Litigation”) and by the PTAB in IPR2020-01390 (the “Amazon IPR”).

Term	Claim Construction Order⁴ in D. Del.	Institution Decision⁵ (Paper 7) in IPR2020-01390.
Acoustic grammar (as recited in claims of ’176 patent)	“grammar of phonotactic rules of the English language that maps phonemes to syllables”	“collection of the phonemes, or distinct units of sound of a spoken language, linked together to form syllables, which are linked together to form the words of the language”

² U.S. Patent No. 7,818,176 (the “’176 patent”). EX1011. The ’176 patent does not claim priority to the ’409 patent or any family member patent of the ’409 patent, nor does the ’176 patent share any inventors with the ’409 patent. *Id.* Title Page. The ’176 patent’s specification incorporates the ’409 patent by reference. *Id.* 3:46-51.

³ *VB Assets, LLC v. Amazon.com, Inc.*, No. 19-1410 (MN) (D. Del.).

⁴ EX1012. In the Amazon Litigation, the parties agreed to the construction of “acoustic grammar.”

⁵ EX1013 at 12. In the Amazon IPR, the Board determined its own construction of “acoustic grammar” based on its analysis of the ’409 patent’s specification.

To the extent the Board determines that either of the constructions of “acoustic grammar” as interpreted in the Amazon Litigation or Amazon IPR are appropriate, this Petition explains why the Grounds presented herein also satisfy the language of the claims under such constructions. Jacobs, ¶¶50-51.

Unless otherwise discussed below, Petitioner applies the plain and ordinary meaning of all claim terms. *See, e.g., Realtime Data, LLC v. Iancu*, 912 F.3d 1368, 1375 (Fed. Cir. 2019) (“The Board is required to construe ‘only those terms . . . that are in controversy, and only to the extent necessary to resolve the controversy.’”) (quoting *Vivid Techs., Inc. v. Am. Sci. & Eng'g, Inc.*, 200 F.3d 795, 803 (Fed. Cir. 1999)). Petitioner does not waive its right to raise additional issues of claim construction in any litigation, nor does it waive any argument in any litigation that claim terms are indefinite or otherwise invalid.

VII. GROUND 1: CLAIMS 1, 2, 3, AND 6 ARE OBVIOUS OVER BAZZI

As explained below, claims 1, 2, 3, and 6 are obvious over Bazzi.

A. Bazzi

Bazzi was published by IEEE in 2000 and is thus prior art under at least 35 U.S.C. § 102(b). EX1008, 1257 (showing 2000 copyright to IEEE and ISBN number). Bazzi was presented and distributed at the Proceedings of the 2000 IEEE

International Conference on Acoustics, Speech, and Signal Processing and was subsequently also publicly available through IEEE's digital library as of August 6, 2002. *E.g.*, EX1029 (IEEE declaration describing distribution and publication of Bazzi); EX1030 (cover page and table of contents excerpt of the Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing showing inclusion of Bazzi, IEEE insignia, 2000 copyright, and ISBN and ISSN numbers for published proceedings). The EPO's reliance on Bazzi in the European counterpart prosecution is further evidence of Bazzi's public availability. Jacobs, ¶53.

Bazzi describes a speech recognition system with a two-stage recognizer where, based on a graph of phonetic probabilities, "syllable graphs are computed in the first stage, and passed to the second stage to determine the most likely word hypotheses." EX1008, 1257-58. This is in contrast to more conventional single-stage recognition systems (Section IV.B.2, *supra*), which determine probable phones/phonemes from a speech signal and convert those directly to a word

hypothesis. Bazzi calls its two-stage recognizer a “Syllable Recognizer” (EX1008, Section 2.2, 1258(2:22)-1259(1:33)).⁶

Bazzi builds on the SUMMIT segment-based speech recognition system developed by MIT’s Spoken Language Systems Group. *Id.*, 1258(1:25-28). Jacobs, ¶54.

1. **Claim 1**

a) **Element [1.1]**⁷

[1.1] A method for providing out-of-vocabulary interpretation capabilities and for tolerating noise when interpreting natural language speech utterances, the method comprising:

As explained in Section VI.A., *supra*, the preamble of the ’409 patent is non-limiting.

To the extent the preamble is limiting, Bazzi renders it at least obvious. Bazzi discloses natural language speech processing “methods that can be used to model out-of-vocabulary and partial words.” EX1008, 1257(2:21-22). It provides these

⁶ In addition to its syllable recognizer, Bazzi also proposes an alternative two-stage Phone Recognizer (EX1008, Section 2.1, 1258(1:38-2:21)), but the grounds presented in this Petition rely on Bazzi’s teachings regarding its Syllable Recognizer.

⁷ Reference numbers in the format of [claim#.limitation#] are added throughout for ease of reference.

methods to address problems posed by the phenomena of “out-of-vocabulary words” and “partially spoken words, which are typically produced in more conversational or spontaneous speech applications.”⁸ *Id.*, (2:17-19). Jacobs, ¶56.

Bazzi’s methods, which “can be used to model out-of-vocabulary ... words use “more flexible sub-word units (such as phones or syllables, which are not constrained to match the active word vocabulary).” EX1008, 1257(2:22-24). Bazzi teaches using these methods both “within a domain-dependent word-based recognition architecture” and as “a separate first stage, operating independently of a given vocabulary.” *Id.*, 1257(2:26-28). Bazzi recognizes that reliance on a word-only lexicon can lead to erroneous interpretations of conversational speech. EX1008, 1257(2:18-20) (“These phenomena also tend to produce errors since the recognizer matches the phonetic sequence with the best fitting words in its active vocabulary.”) And Bazzi recognizes that sub-word units provide out-of-vocabulary capabilities in that they are a closed set, as opposed to an open-ended word vocabulary. *Id.*, 1257(2:23-25) (“**Sub-word units such as phones and syllables have the attractive property of being a closed set, and thus will be able to cover new words, and can conceivably cover most partial word utterances as well.**”) (emphasis added). The ’409 patent’s specification does not explain how its methods specifically

⁸ Years later, the ’409 patent noted similar problems. EX1001, 1:65-2:9.

perform out-of-vocabulary recognition, instead treating such functionality to be a result of phoneme recognition. EX1001, 2:40-46. (“Phoneme recognition may disregard the notion of words, instead interpreting a verbalization as a series of phonemes, which may provide out-of-vocabulary (OOV) capabilities.”). Accordingly, Bazzi’s disclosure of sub-word (including phones and syllables) based recognition, and its statements that its methods using such units can model out-of-vocabulary words, satisfy the claimed out-of-vocabulary recognition capabilities. Jacobs, ¶¶57-58.

A POSITA would further understand that Bazzi’s more flexible speech recognition methods tolerate noise at least because they better process partial word inputs caused by a noisy environment (just as they process partial words caused by “partial word utterances”). Jacobs, ¶59. Bazzi recognized that its use of sub-word units enables coverage of partial word utterances (EX1008, 1257(2:23-25)), which a POSITA would recognize to include words partially recognized due to noise. Jacobs, ¶59. Moreover, any speech recognition method (including those described by Bazzi) would be expected to tolerate some level of noise. For example, Bazzi utilizes bigram and trigram language models (EX1008, 1258(2:32-33)), which provide noise tolerance by predicting potentially misrecognized syllables and words based on recognized preceding syllables and words. Jacobs, ¶¶46, 59. The ‘409

patent itself provides no meaningful discussion of “noise” much less an explanation of how its methods tolerate noise.

Thus, if the preamble were limiting, Bazzi renders obvious a method that satisfies the preamble.⁹ As explained in the following sub-sections, Bazzi renders obvious methods that include each element of claim 1.

b) Element [1.2]

[1.2] receiving an utterance from a user;

Bazzi discloses that its two-stage speech recognizer receives spoken utterances from users. For example, it describes a “two-stage recognizer configuration” in which “**a user** interacting with several different spoken dialogue domains (e.g., weather, travel, entertainment), **might have their speech initially processed by a domain-independent first stage, and then subsequently**

⁹ The obviousness of the claims is unchanged by the fact that Bazzi’s “initial investigation” did not model or examine system behavior on out-of-vocabulary or partial words. *Id.*, 1258(1:15-17). Bazzi provides teachings that render the claims obvious and expressly states that its methods are intended to address out-of-vocabulary and partial word phenomena (*id.*, 1257-58) and that “preliminary results are quite encouraging” (*id.*, 1260(2:8-9)). Jacobs, ¶ 60.

processed by domain dependent recognizers.” EX1008, 1257(2:39)-1258(2:8) (emphasis added); *see also id.* (describing handling of “partial word utterances”). This is consistent with a POSITA’s understanding of the art that speech recognition systems are fundamentally designed and intended to receive a user’s speech and process that speech into recognizable language. *See* EX1009, 5 (depicting and describing the “Basic system architecture of a speech recognition system,” including starting with receiving a user’s voice into a signal processing module) []. Throughout, Bazzi describes its test implementation of a “weather information system” that receives words spoken by a user. *E.g.*, EX1008, 1257(1:34-35) (“in our weather information system, we are constantly faced with new words spoken by users”); *id.*, 1258(1:16-18) (describing recognizing “within-vocabulary utterances”); *id.*, 1259(2:19) (describing a “training set” of utterances for the recognizer); *id.* 1260(2:23-24) (describing experimental results that varied “depending on the length of the utterance”). Accordingly, Bazzi discloses receiving speech (i.e., an utterance) from a user. At the very least, Bazzi renders receiving an utterance obvious because its speech recognizer is premised on receiving spoken utterances from users, and then ultimately recognizing words based on those received spoken utterances. Jacobs, ¶61.

c) **Element [1.3]**

[1.3] recognizing a stream of phonemes contained in the utterance on an electronic device;

Bazzi teaches recognizing a stream of phonemes contained within a given user utterance in that Bazzi teaches its Syllable Recognizer traversing a “scored phonetic graph” to derive a series of phonemes contained in the utterance. With respect to its Syllable Recognizer, Bazzi teaches creating a scored phonetic graph P that includes phones (i.e., phonemes) recognized in the received user utterance. Bazzi represents its two-stage syllable-based recognizer as an FST with the following formula, where S constitutes the search space (i.e., all possible paths in the FST graph) and is generated by composing various component FSTs.

$$S = P \circ L_s \circ G_s \circ L_w \circ G \quad (4)$$

EX1008, 1258(2:25). Bazzi calls the first composed FST, P , the “scored phonetic graph.” *Id.*, 1258(1:31) A scored phonetic graph constitutes a weighted graph of probable “phonetic units” corresponding to the phonetic representation of the received acoustic signal. *Id.*, 1258(2:23-29) As explained in for the State of the Art (Section IV.B.2., *supra*), such a step of subword recognition to identify probabilities of subword units was well-known in the art. *See also* EX1010, 240-41, Fig. 7.2 (disclosing a “Phone Likelihoods” graph structure for conventional frame-based recognizer system). By “scoring” the probable phonetic units contained in the

received signal, probabilities at each transition can be further constrained by (and combined with) relevant lexicons and grammar to arrive at a best overall interpretation. *See* EX1010, 240-41 (combining scored phone likelihood graph with grammar and lexicon). The phonetic units derived in the scored phonetic graph *P*, are subsequently composed with a syllable lexicon, designated as “*L_s*.” EX1008, 1258(2:23-26). The syllable lexicon represents a mapping of phonetic units to syllables created from the relevant word lexicon by “partition[ing] the phone sequence into syllables using an automatic syllabification procedure.” *Id.*, 1258(2:26-29), indicating that the syllable lexicon takes phones (i.e. phonemes) as input units. Because the syllable lexicon takes phonemes as an input to map to syllables, a POSITA would have understood the scored phonetic graph to output corresponding phonemes. Jacobs, ¶62.

While Bazzi does not explicitly use the term “phoneme” a POSITA would have understood that the phones of the received utterance derived by the scored phonetic graph *P* satisfy the constitute “phonemes” as claimed by the ’409 patent for a number of reasons. First, as explained with regard to the State of the Art in Section IV.B.1.a), *supra*, a POSITA understood that “phone” and “phoneme” were used interchangeably in the prior art literature and that, therefore, Bazzi’s disclosure of phones corresponded to the claim’s recitation of phonemes. Jacobs, ¶63. Indeed, during prosecution of the ’409 patent’s counterpart EP application, the Applicant did

not object to the Examiner's interpretation of Bazzi's disclosed use of "phone" as satisfying the claims' use of "phoneme." *See* EX1006, 95 (arguing that, "[a]ssuming that the words phone and phoneme in this context are equivalent, [Bazzi] does not disclose using both phoneme and syllable recognition in the first ... stage of interpretation").

Second, for the English language many phones and phonemes are the same. A phone constitutes the acoustic realization of a given phoneme, and a phoneme may—but many phonemes do not—have multiple phone variants (called *allophones*) that are expressed differently depending on context. But many phones map to only a single phoneme in English. *See* EX1018, 436 (explaining that for "[p]honelike units," in "cases in which the acoustic and phonetic similarities are roughly the same ... then the phoneme and [phonelike unit] will be essentially identical."). Put another way, for any given allophone, unless that allophone can be mapped to multiple phonemes, recognition of that allophone would also result in recognition of the corresponding phoneme. Accordingly, a POSITA would understand that, by recognizing the probable phones of the input utterance via the scored phonetic graph, Bazzi would also recognize a string of phonemes corresponding to such phones. Jacobs, ¶64.

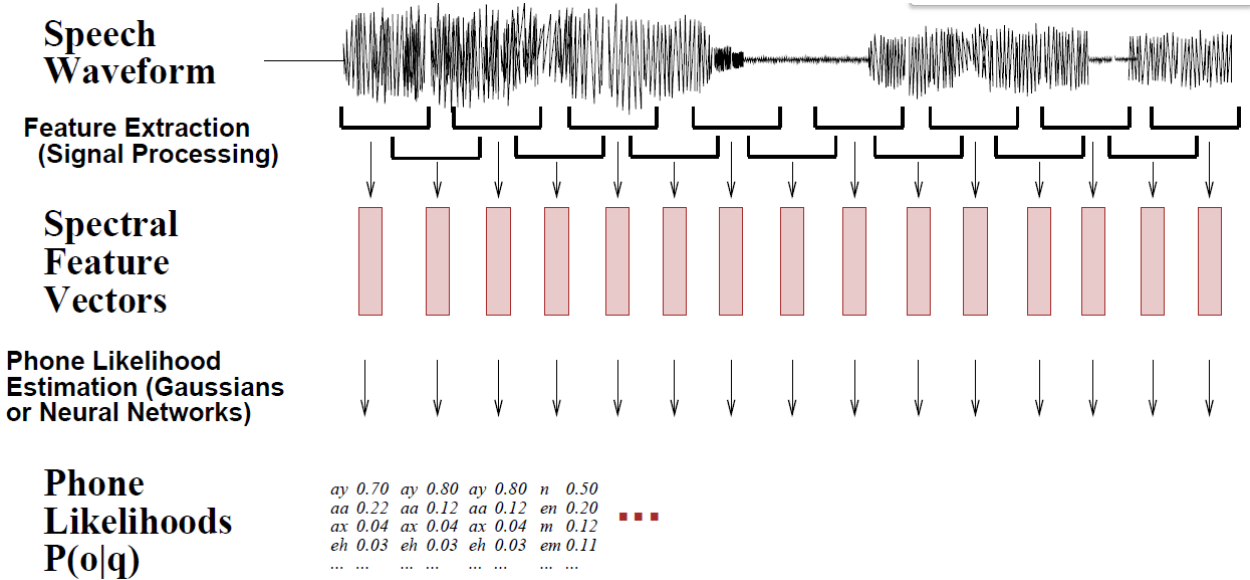
Third, even if a POSITA would have not understood the "phones" disclosed by Bazzi to constitute "phonemes," it would have been an obvious design choice of

the POSITA to implement the “phonetic units” of Bazzi’s scored phonetic graph as phonemes instead of phones. Jacobs, ¶65. It was known to a POSITA that “recognition dictionaries often include a mix of phones and phonemes.” EX1014, 45. And, as explained by Jurafsky, phonemes are often equated with the lexical level in the art, with lexicons thought of “as containing transcriptions expressed in terms of phonemes.” EX1010, 104. Jurafsky presents two ways in which pronunciations of words can be transcribed at a lexical level: “When we are transcribing the pronunciations of words we can choose to represent them at this broad phonemic level; such a broad transcription leaves out a lot of predictable phonetic detail. We can also choose to use a narrow transcription that includes more detail, including allophonic variation.” *Id.* It was therefore well known to a POSITA to present phonemes at the lexical level as an alternative to phones, and it would have been a simple design choice for the POSITA to generate the syllable lexicon (“ L_s ”) based on phoneme-level transcriptions of the available lexicon, so as to map recognized phonemes to syllables. *See* EX1008, 1258(2:26-27) (“The syllable lexicon, L_s , is created from the word lexicon, L , through a direct mapping from phonetic units to syllabic units.”). Implementing the scored phonetic graph P as a graph of probable phonemes (as opposed to probable phones) to map to the phoneme-based L_s would have been well within the skill of a POSITA because acoustic models for modeling probable phonemes from an input acoustic signal were well known in the art. *See*

e.g., EX1018, 45-46 (describing performing feature measurement, detection, and segmentation to generate a “phoneme lattice” from which syllable and word lattices can be derived by integrating vocabulary and syntax constraints); EX1019, 2:42-3:2 (describing “prior art system” of generating acoustic model “which represent each phoneme [] in [a] language” and using the acoustic model to link speech parameters in speech signal to phonemes); EX1020, Fig. 2, 5:14-40 (describing using a set acoustic models to “compute the probability of an acoustic sequence given a particular word sequence,” with an acoustic model “used for each phoneme in [a] particular language.”); Jacobs, ¶66. Accordingly, implementing Bazzi such that its scored phonetic graph and syllable grammar utilize acoustic units of phonemes instead of phones would have been an obvious design choice to a POSITA. Jacobs, ¶¶65-66.

For the reasons explained above, a POSITA would have understood that the scored phonetic graph of Bazzi represents a stream of phonemes. Moreover, Bazzi’s generation and processing of the scored phonetic graph constitutes *recognizing* a phoneme stream. As explained above, the scored phonetic graph constitutes a graph structure representing the most likely phones contained in the received utterance. A POSITA would have understood such identification of the most likely phones to constitute “recognition” of those phones because in the field of automatic speech recognition (ASR), the term “recognizing” does not imply perfect certainty or final

determination. Instead, it refers to the process by which the system analyzes an acoustic signal and produces a representation of the most likely linguistic units. This recognition process is inherently probabilistic, given the variable and noisy nature of spoken input. This understanding by a POSITA is supported by Jurafsky, which, similar to Bazzi, describes a probabilistic graph structure of probable phones as the output of a “**phone recognition stage.**” EX1010, 240-41 (emphasis added). Jurafsky explains that after initial signal processing of the input acoustic waveform, “we use statistical techniques like neural networks or Gaussian models to tentatively recognize individual speech sounds like *p* or *b*,” and “the output of this stage is a vector of probabilities over phones for each frame.” EX1010, 240. Jurafsky presents a representation of such a graph of its phone likelihood estimation based on the input acoustic waveform in Figure 7.2, excerpted below:



EX1010, 241, Fig. 7.2. Accordingly, a POSITA would understand Bazzi's generation of a scored phonetic graph to be consistent with the state of the art's "phone recognition stage" of identifying phone likelihoods in a received utterance. Jacobs, ¶67. This understanding by a POSITA is further confirmed by the '409 patent's specification, which describes that recognizing a stream of phonemes may involve generating preliminary interpretations representing a set of best guesses. *See* EX1001, 5:65-6:6 ("[S]peech engine 112 may generate one or more preliminary interpretations of the user verbalization. The preliminary interpretations may represent a set of best guesses as to the user verbalization arranged in any predetermined form or data structure, such as an array, a matrix, or other forms. In one implementation of the invention, **speech engine 112 may generate the preliminary interpretations by performing phonetic dictation to recognize a stream of phonemes**, instead of a stream of words.") (emphasis added).

Finally, Bazzi discloses receiving the user's utterance and processing that utterance on an *electronic device*. Bazzi describes implementing its speech recognition in a "client/server architecture," where the "two-stage recognition process could be configured to have the first stage run locally on small client devices (e.g., hand-held portables) and thus potentially require less bandwidth to communicate with remote servers for the second stage." EX1008, 1258(1:8-10). Bazzi further discloses experimental testing of its two-stage recognizer system,

which a POSITA would have understood to be performed using a computing device programmed for that purpose. Indeed, POSITAs understood that, generally, speech recognitions systems were computerized systems that used electronics devices to carry out each of their various steps based on voice inputs from users. *See* EX1009, 5 (depicting and describing the “Basic system architecture of a speech recognition system.”); Jacobs, ¶68.

Therefore, Bazzi’s generation of a scored phonetic graph constitutes recognizing on an electronic device a stream of phonemes contained in the received utterance.

d) Element [1.4]

[1.4] mapping the recognized stream of phonemes to an acoustic grammar that phonemically represents one or more syllables, the recognized stream of phonemes mapped to a series of one or more of the phonemically represented syllables; and

Bazzi teaches this limitation under the plain and ordinary meaning of “acoustic grammar”

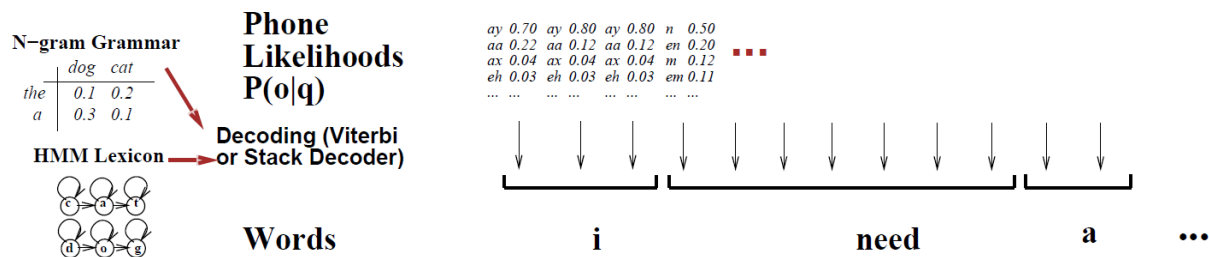
Bazzi discloses that the stream of phonemes of its scored phonetic graph, (denoted P in Bazzi’s Equation 5), is mapped to an acoustic grammar that phonemically represents one or more syllables, by composing the scored phonetic graph with a syllable lexicon and syllable grammar (denoted by L_s or G_s in Bazzi’s

Equation 5). Bazzi's syllable recognizer, represented as an FST in Equation (5), includes a first stage as shown below:

$$S = \underbrace{(P \circ L_s \circ G_s)}_{\text{First Stage}} \circ (L_w \circ G) \quad (5)$$

EX1008, 1258(2:38) (Equation (5), annotated). Bazzi's Equation 5 reflects that in Bazzi's first stage the phonemes of the phonetic graph P are mapped to the syllable lexicon and grammar in order to map the recognized phonemes to corresponding syllables. As Bazzi explains, "[i]n the first stage we compute a syllable graph by searching the composition of P with the precomposed FST $L_s \circ G_s$." *Id.*, 1258(2:35-36). " L_s and G_s are the **syllable lexicon and grammar**, respectively." *Id.*, 1258(2:26) (emphasis added). As explained for the State of the Art (section IV.B.3., *supra*), lexicons and pronunciation grammars, such as those utilized by Bazzi, were well-known and widely used data structures in the art to transform a probabilistic representation of one phonetic unit to another. Jurafsky, for example, describes a conventional speech recognition system in which, in the system's decoding stage, "we take a dictionary of word pronunciations and a language model (probabilistic grammar)." EX1010, 241. This process is depicted in Jurafsky Figure 7.2, excerpted

below, which shows that this dictionary of word pronunciations (a word-level lexicon in the form of an n-gram structure) and word-level grammar (in the form of a Hidden Markov Model structure) are applied to the probability graph of recognized likely phones to determine “the sequence of words which has the highest probability given the acoustic events.” *Id.*, 241, Fig. 7.2.



Id., Fig. 7.2 (excerpt); Jacobs, ¶69.

Bazzi’s syllable lexicon L_s is created from a word lexicon “through a **direct mapping from phonetic units to syllabic units**” by partitioning the phone sequence for each word in the lexicon “into syllables using an automatic syllabification procedure.” EX1008, 1258(2:27-28). (emphasis added). As explained in Section VII.A.1.c), *supra*, the relevant phonetic units of the syllable lexicon are phones, which a POSITA would have understood to meet the ’409 patent’s recitation of

“phonemes.”¹⁰ Accordingly, the syllable lexicon maps phonemes to corresponding syllables (representing those syllables phonemically), and the syllable lexicon thereby provides a model for mapping the recognized groups of phonemes in the scored phonetic graph to corresponding phonemically represented syllables. Jacobs, ¶70.

Bazzi’s grammar G_s constitutes “the syllable language model” that is built by starting with a “word-based training set” and “partition[ing] the words into syllables to obtain syllable sequences for training a syllable bigram or trigram.” EX1008, 1258(2:31-33). Bigrams and trigrams are data structures that were well-known in the art, belonging to a class called n-grams, which were commonly used to model probabilistic grammar. EX1015, 151; EX1010, 197-98. Specifically, bigrams and trigrams respectively model the probabilities of two or three units of speech (e.g., syllables in the case of syllable bigram/trigram) occurring in sequence. A POSITA would therefore have understood that the syllable grammar provides a language model reflecting the probability of a given syllable based on the sequence of preceding syllables. When composed with the scored phonetic graph and lexicon,

¹⁰ Or alternatively, as explained in Section VII.A.1.c) , *supra*, a POSITA would have found it obvious to implement phonemes as the phonetic units used in Bazzi’s system instead of phones.

this probabilistic grammar helps determine the best candidates for mapping probable phonemes detected in the received utterance to the correct corresponding syllables by favoring or disfavoring syllable interpretations based on the observed probabilities of syllable sequences in the training data used to build the n-gram model. Jacobs, ¶71.

The combination of the syllable lexicon and grammar thereby constitutes an “acoustic grammar that phonemically represents one or more syllables.” EX1001, 11:62-12:9 (claim 1). Specifically, the syllable lexicon L_s provides the mapping from the scored phonetic graph P to syllables and the syllable grammar G_s provides the language model that helps to weigh potential interpretations based on the recognized order of syllables. Jacobs, ¶72. As explained above for element [1.3] (*supra*, Section VII.A.1.c), P represents Bazzi’s recognized stream of phonemes. Bazzi maps this stream of phonemes to its acoustic grammar by “searching the composition of P with the precomposed FST $L_s \circ G_s$.” EX1008, 1258(2:35-36). In searching the composition of the scored phonetic graph with the syllable lexicon and syllable grammar, Bazzi maps the recognized phonemes to corresponding phonemically represented syllables contained in the acoustic grammar. Jacobs, ¶72.


Bazzi teaches this limitation under the Amazon Litigation construction of “acoustic grammar”

In the Amazon Litigation the parties agreed that the term “acoustic grammar” means “grammar of phonotactic rules of the English language that maps phonemes to syllables.” EX1012, 2. Phonotactic rules, or phonotactic constraints, were well known in the art and refer to constraints in a language related to what phonetic units tend to co-occur or follow one another. *See* EX1010, 113 (describing “phonotactic constraint on what segments can follow each other”), EX1009, 730 (“A statistical model of phoneme co-occurrence, or phonotactics, was constructed over the training set.”), EX1015, 151 (“The linguistic constraints employed by the second stage of this recognizer are based on the probabilities of groups of two or three words occurring in sequence.”). As explained above in this Section, Bazzi’s syllable grammar is based on the training of syllable bigrams or trigrams. Such syllable bigrams/trigrams constitute phonotactic rules because they provide an empirical model reflecting the phonotactics of the “word-based training set” (EX1008, 1258(2:31)) upon which the bigrams/trigrams are trained. When composed in Bazzi’s FST at recognition, the syllable grammar thereby constrains the determination of the best path through the composition based on the phonotactic rules reflected in the bigram/trigram model. Further, Bazzi explicitly contemplates its recognizer utilizing the English language. EX1008, 1257(1:33-38) (describing the problem in speech recognition caused by the constantly growing vocabulary of the English language). It would be obvious for a POSITA to implement Bazzi’s

system for English and implement its disclosed grammar for the English language. Therefore, Bazzi's acoustic grammar (i.e., the combination of its syllable lexicon and syllable grammar) constitutes a grammar of phonotactic rules that maps phonemes to syllables. Jacobs, ¶73.

Bazzi teaches this limitation under the Amazon IPR construction of “acoustic grammar”

In the Amazon IPR, the Board analyzed the specification of the '409 patent to construe the term “acoustic grammar” as it appeared in the claims of U.S. Patent No. 7,818,176 to mean “collection of the phonemes, or distinct units of sound of a spoken language, linked together to form syllables, which are linked together to form the words of the language.” EX1013, 11-12. As explained above with regard to Bazzi's Equation (5), Bazzi's syllable lexicon is an FST representing a series of phonemes linked together to form syllables and the syllable grammar reflects the order of syllables in the language, i.e. the language model. With regard to linking syllables together to form the words of a language, Bazzi discloses a second recognition step, depicted in Bazzi's Equation (5), as annotated below.

$$S = (P \circ L_s \circ G_s) \circ (L_w \circ G) \quad (5)$$


Second Stage

EX1008, 1258(2:38) (Equation (5), annotated). In this second stage, Bazzi applies a word lexicon and word grammar (“ L_w ” and “ G ”) to the syllable graph output of the first stage. EX1008, 1258(2:36-37). The word lexicon and grammar perform analogous functions to the syllable lexicon and grammar of the first step, but instead of transforming the phonetic output of the scored phonetic graph P to corresponding syllables, the word lexicon and grammar transforms the syllable units of the syllable graph to corresponding words. *Id.*, (2:28-30) (“Entries in the second-stage word lexicon, L_w are represented by sequences of syllable units.”). The word lexicon and word grammar therefore act to link the recognized syllables in the syllable graph to corresponding words to enable a best-path word determination. *Id.*, 1257(1:17-19) (“A finite-state transducer speech recognizer is utilized to configure the recognition as a two-stage process, where ... syllable graphs are computed in the first stage, and passed to the second stage to determine the most likely word hypotheses.”). Therefore, to the extent the Amazon IPR construction applies, the relevant “acoustic grammar” of Bazzi comprises the composition of Bazzi’s syllable-level lexicon and grammar and word-level lexicon and grammar, because these structures include a collection of phonemes linked to form syllables (in the first stage of Bazzi’s recognizer) and syllables linked together to form words (in the second stage of Bazzi’s recognizer). Jacobs, ¶74.

e) **Element [1.5]**

[1.5] generating at least one interpretation of the utterance, wherein the generated interpretation includes the series of syllables mapped to the recognized stream of phonemes.

Bazzi discloses this limitation because it generates one or more interpretations of the syllables contained in a received utterance in the form of a syllable graph. As explained in Sections VII.A.1.c)-d), *supra*, Bazzi determines a scored phonetic graph of the likely phonemes contained in a received utterance and, in the first stage of Bazzi's recognizer, transforms the phonetic units to corresponding syllables. The traversal of the syllable graph therefore represents determining the likely syllables contained in the utterance, thereby providing one or more syllable-level interpretations of the utterance. Jacobs, ¶75.

Bazzi also meets this limitation in a second way in that it provides a word-level interpretation of the received utterance. Again, Bazzi represents its two-stage, syllable-based recognizer using Equation (5):

$$S = (P \circ L_s \circ G_s) \circ \underbrace{(L_w \circ G)}_{\text{Second Stage}} \quad (5)$$

EX1008, 1258(2:38) (Equation (5), annotated). The Bazzi's "second-stage search composes this FST with the precomposed word FST $L_w \circ G$ " where L_w and G

constitute the respective word lexicon and grammar. *Id.*, 1258(2:36-37). Following creation of the syllable graph, “th[e] graph is then composed with the word FST to produce the best word hypothesis.” *Id.*, 1258(2:15). Bazzi thereby searches the syllable graph generated in the first stage against the pre-generated word lexicon and grammar to determine word interpretations corresponding to the series of syllables contained in the syllable graph. *Id.*, 1257(1:17-19) (“A finite-state transducer speech recognizer is utilized to configure the recognition as a two-stage process, where ... syllable graphs are computed in the first stage, and passed to the second stage **to determine the most likely word hypotheses.**”) (emphasis added); Jacobs, ¶76.

By outputting the word interpretation corresponding to the syllables contained in the syllable graph, which is generated by mapping Bazzi’s phonetic graph (i.e., the recognized stream of phonemes), Bazzi provides an interpretation that includes the series of syllables mapped to the recognized stream of phonemes. Jacobs, ¶77.

2. Claim 2

[2] The method of claim 1, the acoustic grammar phonemically representing the one or more syllables in accordance with acoustic elements of an acoustic speech model, wherein each syllable is represented by acoustic elements for an onset, a nucleus, and a coda.

Bazzi teaches this limitation. As explained in Section VII.A.1.d), *supra*, the acoustic grammar of Bazzi includes the syllable lexicon, L_s . And as explained in that section, the syllable lexicon includes acoustic elements of both phonemes and

syllables and phonemically represents one or more syllables by providing a mapping of syllables to corresponding phonemes contained in the syllables. This phonemic representation of syllables is in accordance with an acoustic speech model of a language as defined by the training set used for training Bazzi's acoustic grammar. *See* Bazzi, 1258(2:26-27) (explaining that the syllable lexicon is created from a word lexicon). For example, as explained in Section VII.A.1.d), *supra*, a POSITA would have found it obvious to implement the system of Bazzi for the English language and would have therefore selected an appropriate word lexicon and word-based training set to provide the acoustic speech model for the English language upon which to build the acoustic grammar. As such, Bazzi's acoustic grammar (the syllable lexicon and syllable grammar) includes acoustic elements (i.e. phonemes and syllables) as used in a language and that phonemically represents one or more syllables in accordance with the acoustic elements of an acoustic speech model of that language. Jacobs, ¶78.

With regard to the claim's recitation of "wherein each syllable is represented by acoustic elements for an onset, a nucleus, and a coda," a POSITA would have understood Bazzi's syllable lexicon to include such a representation. A POSITA was aware that the onset, nucleus, and coda are nothing more than the fundamental components of a syllable. *See, e.g.*, EX1001, 2:50-52 ("Portions of a word may be represented by a syllable, which may be further broken down into core components

of an onset, a nucleus, and a coda.”); EX1010, 102 (“A syllable is usually described as having an optional initial consonant or set of consonants called the onset, followed by a vowel or vowels, followed by a final consonant or sequence of consonants called the coda.”). Because Bazzi’s syllable lexicon represents syllables phonetically, i.e., it represents what constituent phonemes make up a given syllable, a POSITA would recognize that those constituent phonemes also reflect the core components—the onset, nucleus, and coda—of the syllable. Jacobs, ¶79. In summary, because Bazzi’s syllable lexicon maps syllables to their constituent phonemes, a POSITA would have understood that those phonemes comprise the onset, nucleus, and coda components of the syllable.

3. Claim 3

[3] The method of claim 2, the acoustic grammar including transitions between the acoustic elements, wherein the transitions are constrained according to phonotactic rules of the acoustic speech model.

Bazzi discloses this limitation. As explained in Section VII.A.1.d), *supra*, Bazzi’s acoustic grammar (syllable lexicon L_s and syllable grammar G_s) are trained using a word lexicon and word-based training set that model the rules the language. The acoustic grammar therefore reflects an acoustic speech model for a given language (i.e., English) upon which the acoustic grammar is trained. The acoustic grammar of Bazzi includes transitions between its acoustic elements; specifically transitions between recognized phonemes and corresponding syllables. *See* Section

VII.A.1.d), *supra* (explaining how Bazzi's acoustic grammar maps phonemes to syllables). This is evidenced by the fact that Bazzi's first stage takes as input the scored phonetic graph representing recognized likely phonemes and composes that graph to produce a syllable graph. *See* EX1008, 1258(2:22-35). Bazzi's acoustic grammar further contains phonotactic rules¹¹ that constrain such transitions. Bazzi's syllable lexicon, which maps phonemes to syllables, constrains such phoneme-to-syllable transition because the syllable lexicon dictates which syllables correspond to the stream of phonemes recognized from the received utterance. *See* EX1008, 1258(2:22-28). The syllable lexicon thereby constrains allowable syllables to specific combinations of phonemes. Jacobs, ¶80.

Bazzi's syllable grammar also includes phonotactic rules constraining syllable-to-syllable transitions. As explained in Section VII.A.1.d), *supra*, Bazzi's syllable grammar (i.e. "the syllable language model" (EX1008, 1258(2:31-34))) constitutes a trained model of syllable bigrams or trigrams to constrain the allowable sequence of syllables; syllable bigrams reflect the probability of two given syllables being in sequence, while syllable trigrams reflect the sequential probability for three

¹¹ As explained in Section VII.A.1.d), *supra*, phonotactic rules refer to constraints in a language governing the allowable sequences of phonetic elements (e.g., allowable syllable structures or phoneme combinations).

given syllables. These syllable bigrams/trigrams provide a model that specifies the relative probability of a given syllable in a language given the preceding syllable. Accordingly, the syllable bigrams/trigrams of the syllable grammar constitute a phonotactic rule constraining the transitions of one syllable to another by restricting allowable syllable sequences. Jacobs, ¶81.

4. **Claim 6**

a) **Element [6.1]**

[6.1] The method of claim 1, further comprising: generating a plurality of candidate interpretations of the utterance, wherein each candidate interpretation includes a series of words or phrases corresponding to the series of syllables mapped to the recognized stream of phonemes;

Bazzi discloses this limitation. As explained in Sections VII.A.1.c)-d), *supra*, the first stage of Bazzi's recognizer composes the scored phonetic graph (i.e., a graph structure of likely phonemes recognized in the received user utterance) with a syllable lexicon and syllable grammar to map the probable phoneme series of the scored phonetic graph to syllables, thereby generating a syllable graph. Bazzi's second recognition stage further composes this output with a word-level lexicon and grammar to recognize words corresponding to the series of syllables contained in the syllable graph. Thus, in the Bazzi recognizer, "syllable graphs are computed in the first stage, and passed to the second stage **to determine the most likely word hypotheses.**" EX1008, 1257(1:18-19) (emphasis added); Jacobs, ¶82.

Bazzi describes that in its FST-based framework, “recognition can be viewed as **finding the best path(s) in the composition.**” *Id.*, 1258(1:28-29) (emphasis added). Bazzi finds multiple potential best paths through its two-stage recognition process using the FST expressed in equation (5), thereby generating a plurality of word interpretation candidates for the user utterance. The FST composed of the phonetic graph, syllable lexicon and grammar, and word lexicon and grammar applies constraints at each level to the recognition process. This results in multiple candidate paths that are more or less probable based on the weights associated with each path and each complete path from start to end in the combined graph represents a sequence of phonemes, syllables, and words; i.e. a candidate interpretation. Jacobs, ¶83. Because the syllable lexicon provides a mapping of phonemes to corresponding syllables, every determined candidate interpretation includes “a series of words or phrases corresponding to the series of syllables mapped to the recognized stream of phonemes” (EX1001, claim 6).

Additionally, various algorithms were known in the art for finding such best paths and Bazzi identifies two such algorithms: “Typical recognizer configurations deploy a bigram language model in a forward Viterbi search, while a trigram (or higher-order) language model is used in a backward A^* search.” EX1008(1:25-27). These same algorithms were well known to a POSITA. Jurafsky, for example describes their use in conventional speech recognition systems: “Finally, in the

decoding stage, we take a dictionary of word pronunciations and a language model (probabilistic grammar) and use a Viterbi or A* decoder **to find the sequence of words which has the highest probability given the acoustic events.**” EX1010, 241 (emphasis added); *see also* EX1009, 592 (“Speech recognition search is usually done with the Viterbie or A* stack decoders.”). And using a forward Viterbi search algorithm together with a backward A* search to find best paths for word sequences, as Bazzi suggests was also well known in the art. *See* EX1009, 670-71 (describing the “Forward-Backward Search Algorithm” using a Viterbi search for forward searching and A* search for backward searching). Accordingly, it would have been obvious to a POSITA to utilize the disclosed search algorithms (i.e., a Viterbi forward search and A* backward search) for Bazzi’s search methodology. Jacobs, ¶¶84-85. As explained further below, a POSITA recognized that such a search methodology generates and scores a plurality of candidates for a given utterance.

b) Element [6.2]

[6.2] assigning a score to each of the plurality of candidate interpretations; and

As explained for Element [6.1], Bazzi discloses finding the best paths through its composition and a POSITA would implement the speech recognition using a forward-backward search methodology using a forward Viterbi search and backward A* search. Bazzi also discloses this limitation through its use of such a methodology. Using such a forward-backward search methodology, “[t]he idea is to first perform

a forward search, during which partial forward scores α for each state can be stored.” EX1009, 670. Subsequently, a backward A* search (also referred to in the art as stack decoding is performed) whereby “the first complete hypothesis found with a **cost below that of all the hypotheses in the stack is guaranteed to be the best word sequence.**” *Id.*, 671. Bazzi thereby provides each candidate interpretation with a score. Thereafter, subsequent complete hypotheses correspond sequentially to the n -best list, as they are **generated in increasing order of cost.**” *Id.*, 672 (emphasis added). Therefore, Bazzi’s disclosed forward-backward search methodology results in scoring the plurality of candidate interpretations. Use of such a backward A* search achieves Bazzi’s purpose of finding multiple best paths. *Id.*, EX1008, 1258(1:28-29) (“[R]ecognition can be viewed as finding the best path(s) in the composition.”). As Huang explains, “It is straightforward to extend stack decoding to produce the n -best hypotheses by continuing to extend the partial hypotheses according to the same A* criterion until n different hypotheses are found. These n different hypotheses are destined to be the n -best hypotheses.” EX1009, 671. Accordingly, a POSITA would implement Bazzi’s disclosed forward-backward search methodology in Bazzi’s speech recognition system to find the scored best paths for a given composition. Jacobs, ¶¶86-87.

c) Element [6.3]

[6.3] selecting a candidate interpretation having a highest assigned score as being a probable interpretation of the utterance.

As explained above for Element [6.2], a POSITA would have found it obvious to use Bazzi's forward-backward search method to find the best paths through the composition of Bazzi's recognition model and assign a score to each such path, thereby resulting in an *n*-best list of hypothesized word sequence candidates for a received utterance. Further, a POSITA would have understood that this methodology also results in selection of a candidate having a highest score as the best interpretation. Jacobs, ¶¶88. Regarding use of the A* backward search algorithm, Huang explains that “[t]he **first complete hypothesis generated by backward A* search** coincides with the best one found in the time-synchronous forward search and **is truly the best hypothesis**. Subsequent complete hypotheses correspond sequentially to the *n*-best list, as they are **generated in increasing order of cost**.” EX1009, 672 (emphasis added). Thereby, Bazzi's use of the backward A* algorithm results in an *n*-best list of word sequence hypotheses sorted by cost with the lowest cost hypotheses corresponding to the best probable interpretation of the utterance's word sequence. A POSITA would have understood a lowest cost for a word sequence interpretation (as returned by the backward A* algorithm) and a “highest assigned score” to be equivalent because a POSITA would recognize the path having

the lowest cost to have the highest probability of being the correct hypothesis. It would have been evident to a POSITA that a cost is mathematically interchangeable with inversely proportional score because a POSITA understood that a lower cost for a word sequence reflects a higher probability of that word sequence being a correct hypothesis. Jacobs, ¶88.

Additionally, it would have been an obvious design choice for a POSITA to implement Bazzi to identify the best word hypothesis based on a highest probability for path (i.e., a score) instead of as a lowest cost, because a POSITA would have recognized a score or a cost as mathematically interchangeable and constituting a limited number of choices for mathematically presenting the quality of a searched word hypothesis. In the field of speech recognition it was well known to represent a best interpretation hypothesis as either a highest probability (i.e., a probability score) or as a weight calculated as the negative log probability. *See e.g.*, EX1010, 187 (“As is commonly true with probabilistic algorithms, they actually use the negative log probability of the word ($-\log(P(w))$.”); EX1021, 7:57-67 (“‘Score’ is a numerical evaluation of how well a given hypothesis matches some set of observations. **Depending on the conventions in a particular implementation, better matches might be represented by higher scores (such as with probabilities or logarithms of probabilities) or by lower scores (such as with negative log probabilities or spectral distances**”) (emphasis added); EX1022, [0030] (same); EX1023, 3:52-60

(chart showing probability of phone realizations as both “probability” and “weight = -log prob.” and showing how the highest scored probability correlates to the lowest weight). The ’409 patent itself recognizes that a best word hypothesis can be interchangeably determined using a highest or lowest score. EX1001, 10:52-54 (“In one implementation of the invention, a candidate interpretation with a highest (or lowest) score may be designated as a probable interpretation.”). Thus, it would be an obvious design choice to a POSITA implementing the system of Bazzi to represent best word hypothesis paths using a score reflecting the probability that the path corresponds to the best interpretation, instead of reflecting the best path as the lowest cost path. Jacobs, ¶89.

VIII. GROUND 2: CLAIMS 2 & 3 ARE OBVIOUS OVER BAZZI IN FURTHER VIEW OF SABOURIN

To the extent Patent Owner argues that Bazzi alone does not render obvious claims 2-3, Sabourin provides additional details that complement Bazzi’s teachings. The combination of Bazzi and Sabourin further confirm the obviousness of claims 2 and 3.

A. Sabourin

Sabourin is U.S. Patent No. 6,108,627 titled “AUTOMATIC TRANSCRIPTION TOOL.” EX1024. Sabourin was filed on October 31, 1997, and

issued on August 22, 2000. *Id.* Sabourin is therefore prior art under at least 35 U.S.C. § 102(b).

Sabourin describes a method for phonemic transcription and generation of phonemic transcription dictionaries for use in speech recognition systems. Sabourin states the following:

A "phonemic transcription" encodes the sound patterns of a word using the phonemic alphabet. In addition to symbols from the phonemic alphabet, phonemic transcriptions may additionally include information relating to word stress and syllabification.... Phonemic transcription dictionaries are useful in a number of areas of speech processing, such as in speech recognition.

EX1024, 1:31-35, 41-43.

Sabourin discloses “[a]n automatic transcription tool ... us[ing] a variety of transcription methods to generate relatively accurate phonemic transcriptions.” *Id.*, 1:60-62. Sabourin performs automatic phoneme transcription by first generating a grapheme (i.e. letter) mapping from a training dictionary and then assigns a mapping value to each mapped grapheme-to-phoneme pair based on the relative frequency with which a particular phoneme string corresponds to its associated grapheme string. *Id.*, Fig. 5, 9:19-10:22. Then a phonemic transcription is created for each word in the training dictionary by decomposing each input orthography (i.e., word) into

possible component substrings, using the assigned grapheme-to-phoneme mapping values to generate a transcription score, and selecting the component substring decomposition with the highest score as the best transcription hypothesis. *Id.*, Fig. 6, 10:24-38.

Following its phonemic transcription process, Sabourin performs word transcription post-processing including syllabification, stress assignment, and phonotactic post-processing. *Id.*, 10:54-60. The first transcription post-processing step is “automatically partition[ing] a transcription into syllables.” *Id.*, 2:35-36.

Regarding this syllabification procedure, Sabourin describes the following:

For each input transcription to be syllabified, syllabification section 802 begins by assigning initial consonants to the onset of the first syllable (step 1001). Similarly, final consonants are assigned to the coda of the final symbol (step 1002). Vowels and diphthongs are then detected and labeled as nuclei (step 1003).

Id., 13:5-10. After adding syllabification information for transcribed phonemes, Sabourin discloses assigning stress information to syllables. *Id.*, 13:46-53. Lexical stress refers to the amount of energy expressed in a syllable, with stressed syllables being pronounced louder or longer. EX1010, 103. Finally, following syllable stress assignment, Sabourin discloses performing “phonotactic post-processing” on the

syllabified, stress assigned transcriptions generated according to Sabourin's method to verify and prune the generated transcriptions. EX1024, 15:8-9. Sabourin explains:

Phonotactic validation is the process of verifying the generated transcriptions. Preferably, for English transcriptions, the following phonotactics are checked: lax-tense vowel combinations, invalid consonant sequences, implausible vowel beginning or endings, implausible consonant beginnings or endings, double phonemes, and single syllable transcriptions whose only vowel is a schwa. ... If phonotactic irregularities are detected, the transcription is labeled as being phonotactically illegal and is aborted.

Id. 15:9-21; Jacobs, ¶¶91-96.

B. The Bazzi-Sabourin Combination

A POSITA would have been motivated to combine the teachings of Bazzi and Sabourin. Specifically, a POSITA would have been motivated to utilize Sabourin's automatic phonemic transcription method, including automatic syllabification, to generate the syllable lexicon of Bazzi. Jacobs, ¶97.

Bazzi discloses that for "each word in the [word] lexicon, we partition the phone sequence into syllables using an automatic syllabification procedure." EX1008, 1258(2:27-28). Bazzi does not provide details of how "syllabification" is accomplished and a skilled artisan would recognize that this process would be

critical to creating the syllable lexicon. Accordingly, a POSITA would look toward a suitable automatic syllabification procedure in order to partition phone sequences into syllables. Jacobs, ¶98. Sabourin discloses just such a technique in that it discloses “automatically partition[ing] a [phonemic] transcription into syllables.” EX1024, 2:36. A POSITA would implement Sabourin’s methodology because A POSITA would recognize the advantages of using Sabourin’s automatic syllabification technique, including labeling the phonemes corresponding to the onset, nucleus, and coda. Jacobs, ¶98. A POSITA would also recognize that Sabourin’s disclosure of assigning stress to syllables would improve word recognition performance, because “difference in lexical stress can affect the meaning of a word.” EX1010, 103. Therefore, Sabourin’s identification of lexical stress in the generated lexicon would help in correctly distinguishing words that may differ based on such stress. Finally, a POSITA would also implement Sabourin’s phonemic transcription methodology for the benefit of its disclosed phonotactic post-processing in order to provide improved verification of the legality of transcriptions in the lexicon. A POSITA would understand that by implementing such phonotactic post-processing, the lexicon would include only valid transcriptions, thereby reducing the potential search space and making the speech recognition system more efficient and accurate. Jacobs, ¶98.

A POSITA would have had a reasonable expectation of success in combining Bazzi and Sabourin because Bazzi explicitly calls for an automatic syllabification procedure for generating a “mapping from phonetic units to syllabic units” (EX1008, 1258(2:27) and Sabourin teaches just such a procedure. Moreover, pronunciation dictionaries (i.e., lexicons) such as those generated by Sabourin that “include[ed] syllabification and stress” were well known in the art (EX1010, 135), and a POSITA would therefore understand such a lexicon to be usable as the syllable lexicon as disclosed by Bazzi. Jacobs, ¶99.

1. Claim 1

The combination of Bazzi and Sabourin renders claim 1 of the '409 patent obvious for the same reasons as Bazzi alone, as explained in Section VII.A.1., *supra*.

2. Claim 2

[2] The method of claim 1, the acoustic grammar phonemically representing the one or more syllables in accordance with acoustic elements of an acoustic speech model, wherein each syllable is represented by acoustic elements for an onset, a nucleus, and a coda.

As explained in Section VII.A.1.c), *supra*, Bazzi teaches its acoustic grammar, includes its syllable lexicon and syllable grammar. And as explained in Section VIII.B., *supra*, it would have been obvious to a POSITA to utilize the phonemic transcription and syllabification methodology of Sabourin to generate the syllable lexicon of Bazzi. As Sabourin discloses, its phonemic transcription includes

syllabification and explicitly labels the onset, nucleus, and coda components of mapped syllables:

For each input transcription to be syllabified, syllabification section 802 begins by **assigning initial consonants to the onset** of the first syllable (step 1001). Similarly, **final consonants are assigned to the coda** of the final symbol (step 1002). **Vowels and diphthongs are then detected and labeled as nuclei** (step 1003).

Id., 13:5-10 (emphasis added). Accordingly, the syllable lexicon generated using Sabourin's phonemic transcription and syllabification includes each syllable represented by acoustic elements for an onset, a nucleus, and a coda. Those syllables are in accordance with acoustic elements (phonemes and syllables) of the acoustic speech model provided by Sabourin. That is, Sabourin's acoustic speech model is its methodology for phoneme transcription and post-transcription processing using a training dictionary for a language to be modeled. Jacobs, ¶¶101-104. Therefore, the acoustic grammar of Bazzi incorporating a syllable lexicon generated via the syllabification procedure of Sabourin renders claim 2 obvious. *Id.*

1. **Claim 3**

[3] The method of claim 2, the acoustic grammar including transitions between the acoustic elements, wherein the transitions are constrained according to phonotactic rules of the acoustic speech model.

As explained above in Section VIII.B., *supra*, it would have been obvious to a POSITA to utilize the phonemic transcription and syllabification methodology of Sabourin to generate the syllable lexicon of Bazzi. This methodology includes Sabourin's "phonotactic post-processing," which constrains transitions between acoustic elements, by checking for and discarding transcriptions having the following phonotactic irregularities: "lax-tense vowel combinations, invalid consonant sequences, implausible vowel beginning or endings, implausible consonant beginnings or endings, double phonemes, and single syllable transcriptions whose only vowel is a schwa." EX1024, 15:9-19. Accordingly, the transitions between acoustic elements (i.e. phonemes and syllables) in the syllable lexicon (and therefore the acoustic grammar comprising the syllable lexicon) generated using Sabourin's phonemic transcription and syllabification are constrained according to the phonotactic rules of the acoustic speech model provided by Sabourin (i.e. Sabourin's methodology for phoneme transcription and post-transcription processing). Jacobs, ¶¶105-106. Therefore, the acoustic grammar of Bazzi incorporating a syllable lexicon generated via the syllabification procedure of Sabourin renders claim 3 obvious.

**IX. GROUND 3: CLAIM 6 IS OBVIOUS
OVER BAZZI IN FURTHER VIEW OF EPSTEIN**

To the extent Patent Owner argues that Bazzi alone does not render obvious claim 6, Epstein provides additional details that complement Bazzi's teachings. The combination of Bazzi and Sabourin further confirm the obviousness of claim 6.

A. Epstein

Epstein is U.S. Patent Publication No. 2005/0055209A1 titled "Semantic language modeling and confidence measurement." EX1025. Epstein was filed on May 9, 2003, and published on March 10, 2005. *Id.* Epstein is therefore prior art under at least 35 U.S.C. §§ 102(a) and (b).

Epstein describes a "system and method for speech recognition [that] includes generating a set of likely hypotheses in recognizing speech, rescoring the likely hypotheses by using semantic content by employing semantic structured language models, and scoring parse trees to identify a best sentence according to the sentence's parse tree by employing the semantic structured language models to clarify the recognized speech." EX1025, Abstract. Epstein describes existing problems with speech recognition systems relying only on conventional n-gram language models: "Although n-gram language models achieve a certain level of performance, they are not optimal. N-grams do not model the long-range dependencies, semantic and syntactic structure of a sentence accurately." *Id.*, [0005]. Epstein addresses this

problem by employing a second stage after initial candidate recognition to re-score candidate sentence interpretations based on a model taking into account sentence semantics (a “semantic structured language model[]”). *Id.*, [0010]. Epstein’s semantic structure language model reflects a model of semantic information for a language, which “may include one or more of word choice, order of words, proximity to other related words, idiomatic expressions or any other information based word, tag, label, extension or token history.” *Id.* [0028].

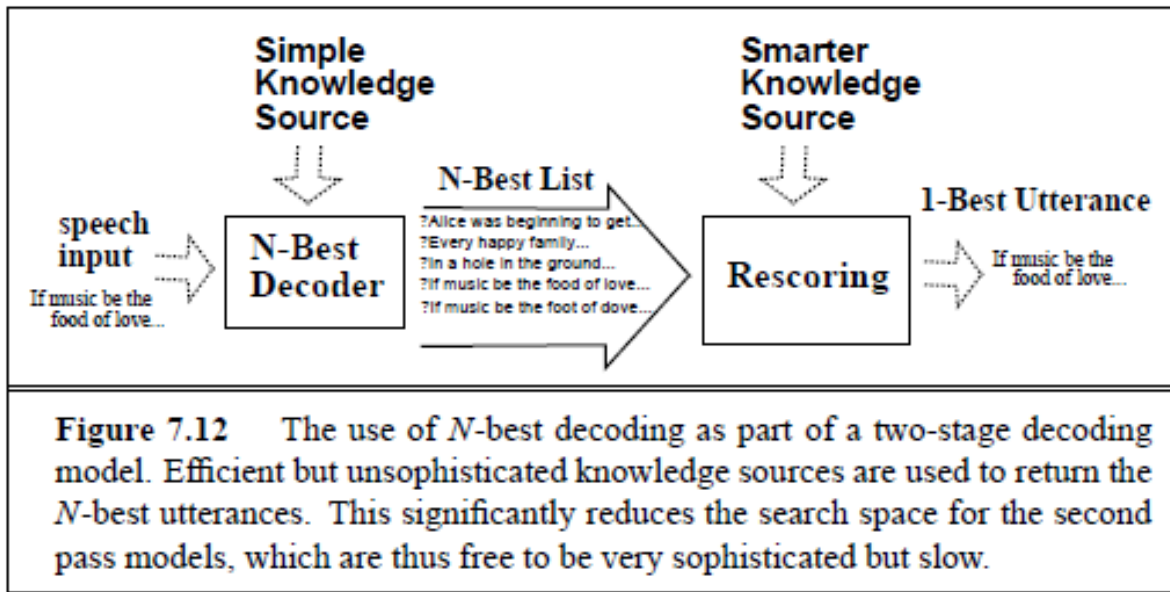
As Epstein explains, following receipt of a user speech input, the method then employs “one or more speech recognition methods ... to generate a set of likely hypotheses. The hypotheses are preferably in the form of an N-best list or lattice structure.” EX1025, [0068]. These likely hypotheses are then rescored in block 208 using semantic structured language models (SSLM) to “rescore the likely hypotheses based on the semantic content of the hypotheses.” *Id.*, [0069]. Finally, the best sentence interpretation is identified by scoring parse trees of the sentence hypotheses using the semantic structured language models. *Id.* A parse tree is a hierarchical representation of the structure of a sentence according to the rules of a grammar. *See* EX1009, 62; EX1025, [0023], [0045], Fig. 2; *see also* Jacobs, ¶¶107-109.

A. The Bazzi-Epstein Combination

A POSITA would have been motivated to combine the teachings of Bazzi and Epstein. Specifically, a POSITA would have been motivated to implement the

speech recognition methods of Bazzi to generate an initial set of likely sentence hypotheses and to then apply Epstein's teachings to re-score those hypotheses and select the best sentence interpretation using Epstein's semantic language models. Jacobs, ¶110.

A POSITA would have been motivated to make such combination due to the known limitations of n-gram language models, such as those implemented by Bazzi (*see* EX1008, 1258(2:31-34) (describing use of syllable bigrams and trigrams)), and the well-known addition of semantic models such as taught by Epstein to overcome those limitations. Epstein itself provides this motivation by explaining that n-gram language models are not optimal, in that they “do not model the long-range dependencies, semantic and syntactic structure of a sentence accurately.” EX1025, [0005]. And Epstein explains that its semantic language modeling techniques “improve speech recognition accuracy.” Indeed, implementing a system that iteratively determines an initial set of best hypotheses using a less sophisticated, more efficient knowledge source (e.g., a bigram language model such as used in Bazzi) and then rescores those hypotheses using a more sophisticated knowledge source (e.g., a semantic language model such as in Epstein) was well known in the art. Such a system is shown, for example in Figure 7.2 of Jurafsky, reproduced below.



EX1010, 253. Schwartz, et al., describe a similar implementation of applying a more efficient set of knowledge sources, including statistical grammar to generate a list of sentence candidates, and then re-ordering that list using additional knowledges sources, including semantics. EX1026, 81-82, Fig. 1; *see also* EX1027, 2:26-47 (describing a speech recognition paradigm of using lesser cost knowledge sources to output “a list of the most likely whole sentence hypotheses” and rescoring list using remaining knowledge sources to find the “highest overall scoring sentence” and stating that “[t]his approach has produced some impressive results”). Accordingly, a POSITA would have been motivated to combine the teachings of Bazzi and Epstein to achieve improved speech recognition accuracy (by rescoring initial hypotheses using more sophisticated knowledge sources) while not requiring

significantly more computational resources (by reducing the search space to the hypotheses returned by the initial search). Jacobs, ¶111.

A POSITA would have had a reasonable expectation of success in combining the teachings of Bazzi and Epstein. As explained in the preceding paragraph, two-stage searches performing rescoring of an initial search using a smarter knowledge source (including semantic knowledge) were well known in the art, and a POSITA would have therefore expected to be successful in using Bazzi’s speech recognition method as such an initial search in combination with Epstein’s teachings regarding reordering using a semantic language model. Moreover, Epstein does not specify a particular speech recognition method for generating initial hypotheses, only stating that “one or more speech recognition methods may be employed to generate a set of likely hypotheses.” EX1025, [0068]. Bazzi performs just such a speech recognition method of generating a set of likely hypotheses. EX1008, 1257(1:19) (determining “the most likely word hypotheses”), 1258(1:28-29) (“[R]ecognition can be viewed as finding the best path(s).”); Jacobs, ¶112.

1. **Claim 1**

The combination of Bazzi and Epstein renders claim 1 of the '409 patent obvious for the same reasons as Bazzi alone, as explained in Section VII.A.1, *supra*.

2. **Claim 6**

a) **Element [6.1]**

[6.1] The method of claim 1, further comprising: generating a plurality of candidate interpretations of the utterance, wherein each candidate interpretation includes a series of words or phrases corresponding to the series of syllables mapped to the recognized stream of phonemes;

The Bazzi-Epstein combination teaches this limitation for the same reasons as Bazzi alone, as explained in Section VII.A.4.a), *supra*. That is, Bazzi explicitly discloses generating a plurality of best path hypotheses for an utterance consisting of a series of words, where those words correspond to the syllables mapped to the recognized stream of phonemes through Bazzi's speech recognition method. Additionally, and in the alternative, as explained in that Section, a POSITA would have found it obvious to implement the search algorithms disclosed by Bazzi to output its word series hypotheses in the form of an n-best list. A POSITA would be further motivated to do so in combining Bazzi with Epstein to match Epstein's preferred format. EX1025, [0068] (“[O]ne or more speech recognition methods may be employed to generate a set of likely hypotheses. **The hypotheses are preferably in the form of an N-best list** or lattice structure.”) (emphasis added); Jacobs, ¶114.

b) Element [6.2]

[6.2] assigning a score to each of the plurality of candidate interpretations; and

The Bazzi-Epstein combination teaches this limitation. After performing an initial speech recognition to determine a set of likely sentence hypotheses (performed by the speech recognition method of Bazzi in the Bazzi-Epstein combination), Epstein explicitly discloses assigning scores to each such likely sentence hypothesis by rescoreing them using Epstein's semantic language models. EX1025, [0069] ("In block 208, semantic structured language models (SSLM) are employed to rescore the likely hypotheses based on the semantic content of the hypotheses . . . In block 210, parse trees are scored to identify a best sentence in accordance with its parse tree. This is performed by using SSLMs."). The Bazzi-Epstein combination therefore assigns a score to each of a plurality of candidate sentence interpretations; Jacobs, ¶115.

c) Element [6.3]

[6.3] selecting a candidate interpretation having a highest assigned score as being a probable interpretation of the utterance.

The Bazzi-Epstein combination teaches this limitation. Epstein explicitly states that a best sentence interpretation is selected based on the score for that sentence determined during its rescoreing procedure. EX1025, [0069] ("In block 210, parse trees are scored to identify a best sentence in accordance with its parse tree.").

While Epstein does not explicitly state that the score for the best sentence is the “highest assigned score,” it would be obvious for a POSITA to implement Epstein such that its best sentence hypothesis is determined using a highest assigned score. First, it would be common sense for a POSITA to score an ordered list of best interpretations from best to worst using a highest-to-lowest score. Further, it would have been an obvious design choice because it was well known in the art that best interpretations could be determined using a score in one of two ways, a highest or a lowest score, as explained in Section VII.A.4.c), *supra*. See EX1021, 7:57-67 (“‘Score’ is a numerical evaluation of how well a given hypothesis matches some set of observations. Depending on the conventions in a particular implementation, better matches might be represented by higher scores (such as with probabilities or logarithms of probabilities) or by lower scores (such as with negative log probabilities or spectral distances”). The Bazzi-Epstein combination therefore teaches selecting a sentence candidate interpretation having a highest assigned score as a probable interpretation of an input utterance. Jacobs, ¶116.

X. NO OBJECTIVE INDICIA OF NON-OBVIOUSNESS

Petitioner is not aware of any evidence of objective indicia of non-obviousness having a nexus to the challenged claims. *Novartis AG v. Torrent Pharms. Ltd.*, 853 F.3d 1316, 1331 (Fed. Cir. 2017) (affirming PTAB obviousness decision).

XI. CONCLUSION

Petitioner requests institution for claims 1, 2, 3, and 6, on Grounds 1-3 specified in this Petition.

Respectfully submitted,

Dated: July 21, 2025

By: /Andrew M. Mason/

Andrew M. Mason Reg. No. 64,034
andrew.mason@klarquist.com
KLARQUIST SPARKMAN, LLP
One World Trade Center, Suite 1600
121 S.W. Salmon Street
Portland, Oregon 97204
Tel: 503-595-5300
Fax: 503-595-5301

Counsel for Petitioner

**CERTIFICATE OF COMPLIANCE WITH
TYPE-VOLUME LIMITATION PURSUANT TO 37 C.F.R. § 42.24**

This brief complies with the type-volume limitation of 37 C.F.R. § 42.24(a)(1)(i).

The brief contains 13,859 words, excluding the parts of the brief exempted by 37 C.F.R. § 42.24(a).

The brief has been prepared in a proportionally spaced typeface using Microsoft Word for O365 in a 14-point Times New Roman font.

Dated: July 21, 2025

By: /Andrew M. Mason/

Andrew M. Mason Reg. No. 64,034
andrew.mason@klarquist.com
KLARQUIST SPARKMAN, LLP
One World Trade Center, Suite 1600
121 S.W. Salmon Street
Portland, Oregon 97204
Tel: 503-595-5300
Fax: 503-595-5301

Counsel for Petitioner

CERTIFICATE OF SERVICE
IN COMPLIANCE WITH 37 C.F.R. § 42.6(e)(4)

The undersigned certifies that the **PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 7,634,409 and Exhibits 1001 – 1031** were served on July 21, 2025, via **Express Mail** on the Patent Owner at the following address of record as listed on the USPTO Patent Center:

Dialect, LLC
c/o The Law Offices of David A. Gerasimow, P.C.
P.O. Box 10861
Chicago, IL 60610

By: /Andrew M. Mason/
Andrew M. Mason Reg. No. 64,034
andrew.mason@klarquist.com
KLARQUIST SPARKMAN, LLP
One World Trade Center, Suite 1600
121 S.W. Salmon Street
Portland, Oregon 97204
Tel: 503-595-5300
Fax: 503-595-5301

Counsel for Petitioner