

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

AMAZON.COM, INC.,
Petitioner,

v.

KAIFI LLC,
Patent Owner.

IPR2025-00625
U.S. Patent No. 8,930,196

**PETITION FOR *INTER PARTES* REVIEW
UNDER 35 U.S.C. § 312 AND 37 C.F.R. § 42.104**

TABLE OF CONTENTS

PETITIONER’S EXHIBIT LIST	5
LISTING OF CHALLENGED CLAIMS.....	8
I. INTRODUCTION	14
II. GROUNDS FOR STANDING.....	15
III. NOTE.....	15
IV. ’196 PATENT.....	15
A. Technological Background	15
B. Overview	15
V. PROSECUTION HISTORY	18
VI. LEVEL OF ORDINARY SKILL IN THE ART.....	19
VII. CLAIM CONSTRUCTION	19
VIII. RELIEF REQUESTED AND THE REASONS FOR THE REQUESTED RELIEF	20
IX. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE....	21
A. Challenged Claims	21
B. Statutory Grounds for Challenges	21
C. Ground 1: Claims 1-5, 8-9, 16-17, and 19-20 are obvious under §103 over Bou-Ghazale in view of Konopka.....	23
1. Bou-Ghazale.....	23
2. Konopka	26
3. Reasons to Combine Bou-Ghazale and Konopka.....	26

4.	Claim 1	29
5.	Claim 2	59
6.	Claim 3	64
7.	Claim 4	65
8.	Claim 5	68
9.	Claim 8	70
10.	Claim 9	74
11.	Claim 16	76
12.	Claim 17	80
13.	Claim 19	82
14.	Claim 20	83
D.	Ground 2: Claims 6-7 are obvious under §103 over Bou-Ghazale in view of Konopka and Lee.	83
1.	Lee.....	83
2.	Reasons to Combine Bou-Ghazale and Konopka with Lee.....	84
3.	Claim 6	86
4.	Claim 7	89
E.	Ground 3: Claims 10-13, and 18 are obvious under §103 over Bou-Ghazale in view of Konopka and Takiguchi.....	92
1.	Takiguchi	92
2.	Reasons to Combine Bou-Ghazale and Konopka with Takiguchi	93

3.	Claim 10	96
4.	Claims 11-12	98
5.	Claim 13	98
6.	Claim 18	98
F.	Ground 4: Claims 14-15 are obvious under §103 over Bou-Ghazale in view of Konopka, Takiguchi, and Lee.....	99
1.	Claims 14-15	99
X.	DISCRETIONARY DENIAL WOULD BE INAPPROPRIATE.....	99
A.	No basis for <i>Fintiv</i> denial	99
B.	No basis for §325(d) denial	101
XI.	CONCLUSION.....	101
XII.	MANDATORY NOTICES	102
A.	Real Party-in-Interest	102
B.	Related Matters.....	102
C.	Lead and Back-up Counsel and Service Information	102
	CERTIFICATE OF WORD COUNT	104
	CERTIFICATE OF SERVICE	105

PETITIONER’S EXHIBIT LIST¹

Ex.1001	U.S. Patent No. 8,930,196
Ex.1002	Prosecution History of U.S. Patent No. 8,930,196
Ex.1003	Declaration of Christopher Schmandt under 37 C.F.R. § 1.68
Ex.1004	<i>Curriculum Vitae</i> of Christopher Schmandt
Ex.1005	S. E. Bou-Ghazale et al., "Hands-free voice activation of personal communication devices," <i>2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings</i> (Cat. No.00CH37100), Istanbul, Turkey, 2000, pp. 1735-1738 vol.3, doi: 10.1109/ICASSP.2000.862087 (“Bou-Ghazale”)
Ex.1006	U.S. Patent Application Pub. 2008/0059188 to Konopka et al. (“Konopka”)
Ex.1007	Lee et al., “Competing models-based text-prompted speech independent verification algorithm,” <i>Speech communication</i> , 48(1), 28-44 (“Lee”)
Ex.1008	U.S. Patent Application Pub. 2006/0122832 to Takiguchi et al. (“Takiguchi”)
Ex.1009	Declaration of Gordon MacPherson
Ex.1010	Declaration of Caroline Champney
Ex.1011	U.S. Patent Application No. 09/692,846
Ex.1012	Definition of “MIPS (million instructions per second),” <i>Waterfields Guide to Computer Terms</i> , Waterfields Press, 1995, pp. 63-64

¹ Citations to non-patent document exhibits are to the original page numbering in the exhibits unless the Bates-stamped page numbers are specifically referenced in the format “Page X of X.”

Ex.1013	Definition of “Million Instructions Per Second (MIPS),” Specialist Computing’s Computer Jargon Dictionary and Thesaurus, 2 nd Edition, Specialist Computing Limited, 2004, p. 129
Ex.1014	Definition of “MIPS,” The Facts on File Dictionary of Computer Science, Revised Edition, Facts on File, 2006, p. 140
Ex.1015	U.K. Patent Application Pub. 2409750 to Hogenhout, et al.
Ex.1016	S.J. Young et al., “Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems,” Cambridge University Engineering Department, 1989
Ex.1017	C.H. Lee et al., “Acoustic modeling for large vocabulary speech recognition,” Computer and Speech Language, (1990) 4, 127-165
Ex.1018	C. Schmandt, Voice Communication with Computers, Van Nostrand Reinhold, 1994
Ex.1019	John Holmes and Wendy Holmes, Speech Synthesis and Recognition, 2 nd Edition, Taylor & Francis, 2001
Ex.1020	<i>Reserved</i>
Ex.1021	L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, Vol. 77, No. 2, February 1989 (“Rabiner 89”)
Ex.1022	L. Rabiner et al., Fundamentals of Speech Recognition, Prentice Hall PTR, 1993 (“Rabiner”)
Ex.1023	U.S. Patent No. 4,926,488 to Nadas et al. (“Nadas”)
Ex.1024	U.S. Patent Application Pub. 2003/0220796 to Aoyama et al. (“Aoyama”)
Ex.1025	U.S. Patent No. 6,092,045 to Stublely et al. (“Stublely”)
Ex.1026	U.S. Patent No. 6,151,574 to Lee et al. (“Lee ’574”)
Ex.1027	U.S. Patent No. 6,452,348 to Toyoda (“Toyoda”)

Ex.1028	Docket Control Order, <i>KAIFI LLC v. Amazon.com, Inc. et al</i> , No. 2:24-cv-00542 (EDTX)
Ex.1029	Preliminary Infringement Contentions, <i>KAIFI LLC v. Amazon.com, Inc. et al</i> , No. 2:24-cv-00542 (EDTX)
Ex.1030	EDTX Local Rules 2023
Ex.1031	Judge Gilstrap Docket for March 16, 2026
Ex.1032	Interim Procedure for Discretionary Denials in AIA Parallel District Court Litigation, June 21, 2022
Ex.1033	Statistics on District Court Timing (Dec. 30, 2024)
Ex.1034	S.E. Bou-Ghazale et al., “A robust endpoint detection of speech for noisy environments with application to automatic speech recognition,” <i>2002 IEEE International Conference on Acoustics, Speech, and Signal Processing</i> , vol.4, pp.IV-3808-IV-3811, 2002
Ex.1035	M. Kuhne et al., “Voice Activation Using Prosodic Features,” <i>2004 INTERSPEECH 2004 – ICSLP 8th International Conference on Spoken Language Processing ICC</i> , 2004
Ex.1036	W. Abdulla, “HMM-based techniques for speech segments extraction,” <i>Scientific Programming</i> , 2002
Ex.1037	E. Mainardi et al., “Controlling a prosthetic arm with a throat microphone,” <i>2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society</i> , Lyon, France, 2007, pp. 3035-3039, doi: 10.1109/IEMBS.2007.4352968
Ex.1038	K.C. Kwak, “Speaker identification and verification for intelligent service robots,” <i>Proceedings of the 2007 International Conference on Artificial Intelligence, ICAI 2007</i> .

LISTING OF CHALLENGED CLAIMS

Claim 1	
[1.0]	A method for recognizing a speech of a speaker by a device, the method comprising:
[1.1]	receiving, by a preset call command recognition unit, an input speech spoken by the speaker;
[1.2]	receiving, by the preset call command recognition unit, a subsequent input speech spoken, by the speaker, subsequently to the input speech;
[1.3]	comparing, by the preset call command recognition unit, the input speech with a preset call command;
[1.4]	determining, by the preset call command recognition unit, whether the input speech is recognized as the preset call command according to a result of the comparison; and
[1.5]	based on determining that the input speech is recognized as the preset call command, activating a continuous speech recognition unit configured to determine whether the subsequent input speech is recognized as an actual command, in order to reduce computational load on the device; and
[1.6]	wherein the preset call command recognition unit comprises a token passing unit configured to assume the input speech to be the preset call command according to the result of the comparison using a minimum recognition network composed of a silence interval accompanied by noise and the preset call command.
Claim 2	
[2.0]	The method of claim 1, wherein determining whether the input speech is recognized as the preset call command further comprises: computing, by a confidence detection unit, a confidence rate of the input speech; and
[2.1]	recognizing, by the confidence detection unit, the input speech as

	the preset call command based on the computed confidence rate.
Claim 3	
[3.0]	The method of claim 2, wherein recognizing the input speech comprises: recognizing the input speech as the preset call command if the computed confidence rate is higher than a threshold value.
Claim 4	
[4.0]	The method of claim 2, wherein recognizing the input speech comprises: determining whether or not the input speech is an authentic call command based on the computed confidence rate; and
[4.1]	recognizing the input speech as the preset call command if it is determined that the input speech is the authentic call command.
Claim 5	
[5.0]	The method of claim 1, wherein determining the input speech is recognized as the preset call command further comprises: computing a log likelihood ratio of the input speech; and
[5.1]	assuming the input speech to be the preset call command based on the computed log likelihood ratio.
Claim 6	
[6.0]	The method of claim 5, wherein the computed log likelihood ratio is a natural logarithm of a probability on the input speech for a phoneme model of the preset call command, and
[6.1]	the computed confidence rate is another log likelihood ratio.
Claim 7	
[7.0]	The method of claim 6, wherein said another log likelihood ratio is computed by subtracting, from the log likelihood ratio, a natural logarithm of a probability on the input speech for an anti-phoneme model of the preset call command.

Claim 8	
[8.0]	The method of claim 1, further comprising: determining, using a rule check unit, whether or not the recognized subsequent input speech corresponds to a target sentence.
Claim 9	
[9.0]	An apparatus for recognizing a speech of a speaker, the apparatus comprising:
[9.1]	a processor,
[9.2]	a preset call command recognition unit using the processor; and
[9.3]	a continuous speech recognition unit using the processor,
[9.4]	wherein the preset call command recognition unit is configured to receive an input speech spoken by the speaker;
[9.5]	receive a subsequent input speech spoken, by the speaker, subsequently to the input speech;
[9.6]	compare the input speech with a preset call command; and
[9.7]	recognize the input speech as the preset call command according to a result of the comparison,
[9.8]	wherein the continuous speech recognition unit is configured to recognize the subsequent input speech as an actual command if the input speech is recognized as the preset call command by the preset call command recognition unit, and
[9.9]	wherein the preset call command recognition unit comprises a token passing unit configured to assume the input speech to be the preset call command according to the result of the comparison using a minimum recognition network composed of a silence interval accompanied by noise and the preset call command.
Claim 10	

[10.0]	The apparatus of claim 9, wherein the preset call command recognition unit further comprises: an adaptive filter configured to receive the input speech and the subsequent input speech; and
[10.1]	a confidence detection unit configured to compute a confidence rate of the assumed input speech and recognize the assumed input speech as the preset call command based on the computed confidence rate.
Claim 11	
[11.0]	The apparatus of claim 10, wherein the confidence detection unit is further configured to recognize the assumed input speech as the preset call command if the computed confidence rate is higher than a threshold value.
Claim 12	
[12.0]	The apparatus of claim 10, wherein the confidence detection unit is further configured to determine whether or not the assumed input speech is an authentic call command based on the computed confidence rate and recognize the assumed input speech as the preset call command if it is determined that the assumed input speech is the authentic call command.
Claim 13	
[13.0]	The apparatus of claim 10, wherein the token passing unit is further configured to compute a log likelihood ratio of the input speech and assume the input speech to be the preset call command based on the computed log likelihood ratio.
Claim 14	
[14.0]	The apparatus of claim 13, wherein the computed log likelihood ratio is a natural logarithm of a probability on the input speech for a phoneme model of the preset call command, and
[14.1]	the computed confidence rate is another log likelihood ratio.
Claim 15	

[15.0]	The apparatus of claim 14, wherein said another log likelihood ratio is computed by subtracting, from the log likelihood ratio, a natural logarithm of a probability on the input speech for an anti-phoneme model of the preset call command.
Claim 16	
[16.0]	The apparatus of claim 9, wherein the continuous speech recognition unit comprises: a keyword spotting unit configured to recognize the subsequent input speech as the actual command if the input speech is recognized as the preset call command; and
[16.1]	a rule check unit configured to determine whether the recognized subsequent input speech corresponds to a target sentence.
Claim 17	
[17.0]	A method for recognizing a speech of a speaker by a single device, the method comprising:
[17.1]	receiving, by a preset call command recognition unit of the single device, an input speech spoken by the speaker;
[17.2]	receiving, by the preset call command recognition unit, a subsequent input speech spoken, by the speaker, subsequently to the input speech;
[17.3]	comparing, by the preset call command recognition unit, the input speech with a preset call command;
[17.4]	determining, by the preset call command recognition unit of the single device, whether the input speech is recognized as the preset call command according to a result of the comparison; and
[17.5]	only if the input speech is recognized as the preset call command, activating a continuous speech recognition unit of the single device that is configured to perform continuous speech recognition to determine whether the subsequent input speech is recognized as an actual command, in order to reduce overall computational load on the single device; and

[17.6]	wherein the preset call command recognition unit comprises a token passing unit configured to assume the input speech to be the preset call command according to the result of the comparison using a minimum recognition network composed of a silence interval accompanied by noise and the preset call command.
Claim 18	
[18.0]	The method of claim 17, wherein the single device is a portable device or an embedded system.
Claim 19	
[19.0]	The method of claim 17, wherein the single device,
[19.1]	wherein determining whether the input speech is recognized as the preset call command further comprises: computing, by a confidence detection unit, a confidence rate of the input speech; and
[19.2]	recognizing, by the confidence detection unit, the input speech as the preset call command based on the computed confidence rate.
Claim 20	
[20.0]	The method of claim 17, wherein determining whether the input speech is recognized as the preset call command further comprises: computing a log likelihood ratio of the input speech; and
[20.1]	assuming the input speech to be the preset call command based on the computed log likelihood ratio.

I. INTRODUCTION

U.S. Patent No. 8,930,196 (the “’196 patent,” Ex.1001) relates to a two-step “continuous speech recognition” method and system for noisy environments that continually listens for a “call command” (i.e., wake word). If the call command is recognized, the system then listens for a subsequent user command. As shown below, however, this concept was well known in the art. For example, eight years before the precursor to the ’196 patent was filed in Korea, engineers at Conexant Systems designed and built a speech recognition system that also uses a two-step speech recognition system that “continuously” listens for the call/activation command “wake-up” (while rejecting noise), and, if recognized, listens for a subsequent, actual command. *See* Ex.1005. The concept upon which allowance was based—token passing—was similarly a “well-known approach” to improving speech recognition systems. Ex.1006, [0052].

Accordingly, pursuant to §§311, 314(a), and 37 C.F.R. §42.100, Amazon.com, Inc. (“Petitioner”) respectfully requests that the Board review and cancel as unpatentable under (pre-AIA) §103(a) claims 1-20 (hereinafter, the “Challenged Claims”) of the ’196 patent.

II. GROUNDS FOR STANDING

Petitioner certifies that the '196 patent is eligible for IPR and that Petitioner is not barred or estopped from requesting IPR challenging the patent claims. 37 C.F.R. §42.104(a).

III. NOTE

Petitioner cites to exhibits' original page numbers unless the Bates-stamped page numbers are specifically referenced in the format "Page X of X." **Emphasis** in quoted material has been added. Claim terms are presented in *italics*.

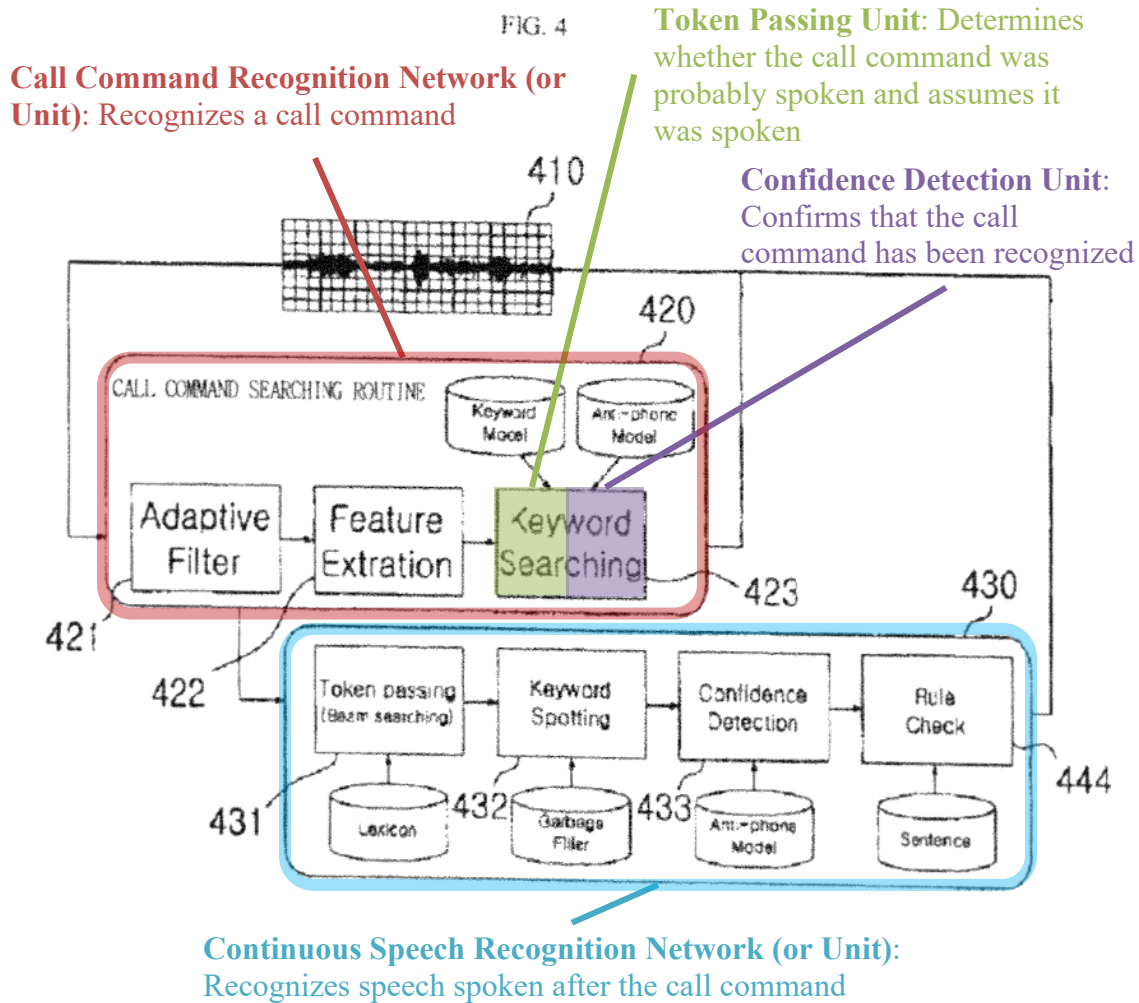
IV. '196 PATENT

A. Technological Background

The '196 patent relates to speech recognition concepts that were well-known as of the earliest possible priority date. Mr. Schmandt's declaration describes the relevant background technology predating the '196 patent, which is cited as appropriate below. Ex.1003, ¶¶27-56; Exs.1005-1008, 1012-1027.

B. Overview

The '196 patent relates to a two-step "continuous speech recognition" method and system that continually listens for a "call command" (i.e., wake word), and, if recognized, listens for a subsequent command (i.e., "actual command"). Ex.1001, 1:9-10, 6:24-28. The '196 patent's system is depicted in Fig. 4 below.



Ex.1001, Fig.4 (annotated); Ex.1003, ¶57.

“Call command recognition network 420” is responsible for “recognizing call commands” and “detecting the confidence thereof.” Ex.1001, Fig. 4, 5:7-12.

The “continuous speech recognition network 430” is responsible for “recognizing speech which is an actual command.” Ex.1001, Fig. 4, 6:24-28. If a call command is recognized, then the continuous speech recognition unit is “activat[ed].”

Ex.1001, claim 1.

The call command recognition network “includes...a keyword searching unit 423,” which in turn “includes a token passing unit...and a confidence detection unit.” Ex.1001, Fig. 4, 5:7-12, 5:40-52. To determine whether the call command was spoken, “every input speech frame is compared with a recognition network for only one word in real time.” Ex.1001, 6:52-7:3, 3:64-4:2, 7:34-39, claim 1. Specifically, the call command recognition network “us[es] a left-to-right model,” such as a “keyword model” as shown in Fig. 4. Ex.1001, 3:64-4:2, 6:52-7:3, Fig. 4. The call command recognition network also accounts for “noise” and “silence” in input speech frames: “the keyword searching unit 423 includes a token passing unit for configuring a speech frame, which includes a silence interval accompanied by noise and call commands from the speaker.” Ex.1001, 5:40-52, 5:13-19.

If the comparison determines, based on probability, that the call command was probably spoken, the token passing unit “assume[s] that an actual call command has been spoken.” Ex.1001, 6:62-7:3, 7:29-39, 7:18-24, claim 1. Then, the confidence detection unit confirms that the call command is recognized based on whether the confidence rate of the input speech (represented by a “log likelihood ratio” (LLR)) meets a threshold. Ex.1001, 8:28-42, 7:63-8:15. “When LLR is determined...and ‘LLR>threshold’ is satisfied, the continuous speech recognition network is executed” and “speech spoken subsequent to the call

command is transferred to the continuous speech recognition engine to perform continuous speech recognition.” Ex.1001, 7:63-8:15, 8:32-42, 6:24-28.

As shown below, each of the above speech recognition concepts were previously known in the art. Ex.1003, ¶¶57-62.

V. PROSECUTION HISTORY

The ’196 patent issued from U.S. Application 13/591,479 filed on August 22, 2012 as a continuation of U.S. Application 12/863,437, now U.S. 8,275,616 (the “’616 patent”). The ’616 patent is a U.S. national phase of PCT Application PCT/KR2009/002118 filed April 22, 2009, which claims priority to Korean Application 1020080049455 filed May 28, 2008. It is unnecessary to determine whether the ’196 patent is entitled to its earliest possible priority date because the prior art relied upon herein pre-dates the earliest possible priority date.

Throughout prosecution, the claims were variously rejected under §§101, 102, and 103. Ex.1002, Pages 90-95, 120-126, 148-154, 188-195 of 266. In a Notice of Allowance, the Examiner pointed to the “*activating a continuous speech recognition unit...*” and the “*token passing unit*” limitations as the alleged allowable subject matter. Ex.1002, Pages 227-228 of 266. As shown below, however, this subject matter was well known. *See, e.g.*, Ex.1006, [0052] (“token passing” is “a well-known approach to tracking the best word hypotheses”).

VI. LEVEL OF ORDINARY SKILL IN THE ART

A Person of Ordinary Skill in The Art (“POSITA”) in May 2008 (the earliest possible priority date) would have been someone knowledgeable and familiar with speech recognition arts that are pertinent to the ’196 patent. That person would have at least a bachelor’s degree in computer engineering, computer science, or equivalent training, and approximately two years of experience working in the field of speech recognition. Lack of work experience can be remedied by additional education, and vice versa. Ex.1003, ¶¶17-19.

VII. CLAIM CONSTRUCTION

No claim terms require express construction to resolve the issues presented in this Petition.² *Nidec Motor Corp. v. Zhongshan Broad Ocean Motor Co.*, 868 F.3d 1013, 1017 (Fed. Cir. 2017). Additionally, no claim includes the word “means,” and for the purposes of this Petition, Petitioner applies Patent Owner’s apparent position in its Infringement Contentions that no terms invoke §112, ¶6. *See* Ex.1029, Pages 13-347 of 347 (no terms identified as invoking §112, ¶6³); *see also* 37 C.F.R. §42.104(b)(3) (requiring the petition to set forth only “how the

² Petitioner does not concede that any term in the challenged claims satisfies §112 or recites patentable subject matter under §101.

³ *See* Ex.1030 (EDTX Local Rules), 55.

challenged claim is to be construed”); *Hospira, Inc. et al. v. Amgen Inc.*, IPR2021-00528, Paper 7 at 7 (Aug. 17, 2021) (“We determine that Petitioner’s willingness to adopt Patent Owner’s proposed constructions for purposes of this proceeding is sufficient to satisfy the requirements of §42.104(b)(3).”); *ZTE Corp. et al. v. WSOU Invests., LLC*, IPR2021-00698, Paper 10 at 8-10 (Oct. 18, 2021) (declining to apply §112 ¶6 where a claim term did not include the words “means” and Petitioner adopted Patent Owner’s plain and ordinary meaning construction for the purposes of the proceeding).

In the event the Board finds any of the “*unit*” terms in the Challenged Claims invoke §112, ¶6, Petitioner adopts a “combination of hardware, processing, and software functions” as the corresponding structure, which mimics the language Patent Owner uses to describe Petitioner’s allegedly infringing products in the Infringement Contentions. Ex.1029, Pages 13-347 of 347.

VIII. RELIEF REQUESTED AND THE REASONS FOR THE REQUESTED RELIEF

Petitioner asks that the Board institute a trial for *inter partes* review and cancel the Challenged Claims in view of the analysis below.

IX. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE

A. Challenged Claims

Petitioner challenges claims 1-20, which correspond to the claims asserted in the plaintiff's Infringement Contentions in the co-pending litigation. Ex.1029.

B. Statutory Grounds for Challenges

Grounds	Claim(s)	Basis (pre-AIA)
#1	1-5, 8-9, 16-17, 19-20	§103 Bou-Ghazale and Konopka
#2	6-7	§103 Bou-Ghazale , Konopka , and Lee
#3	10-13, 18	§103 Bou-Ghazale , Konopka , and Takiguchi
#4	14-15	§103 Bou-Ghazale , Konopka , Takiguchi , and Lee

S. E. Bou-Ghazale et al., “Hands-free voice activation of personal communication devices,” *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Cat. No.00CH37100), Istanbul, Turkey, 2000, pp. 1735-1738 vol.3, doi: 10.1109/ICASSP.2000.862087 (“**Bou-Ghazale**,” Ex.1005) is an article that was publicly available by at least August 6, 2002, and as early as June 9, 2000. As indicated by the declaration of the custodian of records for IEEE (Ex.1009), the article was made publicly available as of the last day of the IEEE International Conference on Acoustics, Speech, and Signal Processing, which occurred between June 5-9, 2000, and it was published and

made available for public download in the IEEE digital library, IEEE Xplore on August 6, 2022. Ex.1009, ¶¶10, 11. As further evidence of its accessibility, Bou-Ghazale was routinely cited by POSITAs in papers before the priority date of the '196 patent. Ex.1034-1036. Bou-Ghazale is thus prior art under at least §§102(a)-(b).

U.S. Patent Application Pub. 2008/0059188 to Konopka et al. (“**Konopka**,” Ex.1006) published on March 6, 2008, and was filed as U.S. Patent Application 11/932,771 (the “’771 application”) on October 31, 2007. The ’771 application was filed as a continuation of U.S. Application 09/692,846 (the “’846 application,” Ex.1011), which was filed on October 19, 2000. The ’846 application claims priority to U.S. Provisional Application 60/160,281 filed on October 19, 1999. Konopka is thus prior art under at least §§102(a) and (e).

Lee et al., “Competing models-based text-prompted speech independent verification algorithm,” *Speech communication*, 48(1), 28-44. (“**Lee**,” Ex.1007) is an article that was publicly available by at least July 5, 2005. As indicated by the declaration of the custodian of records for Elsevier (Ex.1010), the article was published and made available online for public download in Volume 48, Issue 1 of the *Speech Communication* journal, available online through Elsevier’s Science Direct digital library, on July 5, 2005. Ex.1010, ¶¶10, 11. As further evidence of its accessibility, Lee was routinely cited by POSITAs in papers before the priority

date of the '196 patent. Ex.1037-1038. Lee is thus prior art under at least §§102(a)-(b).

U.S. Patent Application Pub. 2006/0122832 to Takiguchi et al.
 (“**Takiguchi**,” Ex.1008) published on June 8, 2006. Takiguchi is thus prior art under at least §§102(a)-(b).

C. Ground 1: Claims 1-5, 8-9, 16-17, and 19-20 are obvious under §103 over Bou-Ghazale in view of Konopka.

1. Bou-Ghazale

Like the '196 patent, Bou-Ghazale describes a two-step continuous speech recognition system that “continuously” listens for a “pre-defined activation command” while rejecting extraneous speech and noise, and, if recognized, listens for a subsequent, actual command. Ex.1005, Abstract, Sec.1, 1735. Bou-Ghazale describes a way to use one’s voice to activate and control a device, even in noisy “busy-hands busy-eyes environments.” *Id.*; Ex.1003, ¶68.

With reference to Figure 2 below, Bou-Ghazale’s voice activation system includes a hidden Markov model (HMM)-based speech recognition component that continually “processes input speech on a frame-by-frame basis and makes a decision at every frame whether the activation word has been spoken.” Ex.1005, Sec.2, 1735, 1736. If the speech recognition component determines that it is likely that the activation word has been spoken, then a post-processing component next

“validates the speech recognition result” using a set of criteria and confidence measurements. Ex.1005, Sec.2, 2.3, 3.1-3.2, 1735-1737. Once the post-processing component confirms that the activation word has been recognized, then the main speech recognition system is activated for recognizing “further commands.” *Id.*; Ex.1003, ¶69.

2. Konopka

Like the '196 patent and Bou-Ghazale, Konopka relates to a two-step HMM-based continuous speech recognition system. Ex.1005, Abstract; Ex.1006, Abstract, [0002], [0039], Fig. 4; Ex.1003, ¶70. Konopka explains that “token passing” was a “well-known approach” to improve HMM-based systems. Ex.1006, [0052]. Token passing provides a memory efficient solution to the known problem in HMM-based systems: “storage of back-pointers that allows for finding the best word sequence hypothesis once all speech data has been processed.” *Id.*; Ex.1003, ¶71.

3. Reasons to Combine Bou-Ghazale and Konopka

As described below, a POSITA would have been motivated to combine the teachings of Bou-Ghazale and Konopka. Ex.1003, ¶¶72-79. It would have been obvious, beneficial, and predictable to utilize Konopka’s token passing technique in combination with Bou-Ghazale’s voice activation system—for example, to provide a more memory efficient way of matching incoming speech to known words. Ex.1003, ¶72.

As an initial matter, a POSITA, when considering the teachings of Bou-Ghazale would have also considered the teachings of Konopka. Ex.1003, ¶73. Both Bou-Ghazale and Konopka relate to HMM-based continuous speech recognition and are therefore analogous art to the '196 patent. Ex.1005, Abstract; Ex.1006,

[0002], [0039]; Ex.1001, Abstract, Fig. 6; Ex.1003, ¶73.

Bou-Ghazale teaches a voice activation system that includes an HMM-based continuous speech recognition component that “**performs a trace-back at every frame, and passes the result string with the highest likelihood score** to the post-processing component.” Ex.1005, Sec.2.2, 1736. However, Bou-Ghazale does not explicitly describe how it performs back-tracing to determine the result string with the highest likelihood score, including whether it uses the traditional approach or token passing. Konopka explains that a well-known problem with the traditional back-tracing approach is that it requires “each state contain a pointer back to the previously best state for each frame” which can lead to “high memory requirements” for storing the back-pointers. Ex.1006, [0052]. Konopka provides a “well-known” and predictable solution to this problem “based on the idea of passing ‘tokens’ around to the various states instead of allocating memory on a per-state basis.” *Id.* Using Konopka’s well-known token passing technique, token pointers corresponding to the sequence (or string) of states that are less likely to correspond to the unknown input speech are discarded at each state during back-tracing. Ex.1006, [0052]; Ex.1003, ¶76. As a result, “a highly memory efficient solution to an otherwise well-known problem in HMM-based speech recognition systems; the storage of back-pointers that allows for finding the best word sequence hypothesis once all speech data has been processed” is provided.

Ex.1006, [0052]; Ex.1003, ¶¶74-76.

A POSITA would have been motivated to implement Konopka's token passing technique into Bou-Ghazale's HMM-based speech recognition component because doing so would "minimize the memory requirements associated with storing" back-pointers and make Bou-Ghazale's system more efficient. Ex.1006, [0052]; Ex.1003, ¶77; *see also KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398, 417 (2007) (finding obviousness when a known technique "would improve similar devices in the same way"). Specifically, the benefits of Konopka's technique would have motivated a POSITA to use token passing to discard the strings that are less likely to match the unknown speech during Bou-Ghazale's speech recognition process (as modified by Konopka) to determine the result string that most likely matches the unknown speech ("the result string with the highest likelihood score") and pass it to the post-processing component. Ex.1003, ¶77.

A POSITA would have had a reasonable expectation of success applying Konopka's technique to Bou-Ghazale. Ex.1003, ¶78. For example, the combination does not change the intended functionality of either Bou-Ghazale's voice activation system or Konopka's token passing technique. Ex.1003, ¶78. Konopka's token passing was specifically intended to improve HMM-based speech recognition systems, such as Bou-Ghazale's voice activation system. Ex.1006, [0052]. Moreover, POSITAs routinely utilized token passing in the context of

HMM-based systems. Ex.1003, ¶78 (citing Ex.1015-1016). Thus, it would have been within the background knowledge and standard capabilities of a POSITA to modify Bou-Ghazale’s speech recognition algorithm to use token passing as its back-tracing technique. Ex.1003, ¶78.

Accordingly, a POSITA would have found it obvious to apply Konopka’s teachings to Bou-Ghazale’s voice activation system and methods because the combination merely amounts to applying a known token passing technique to a system ready for improvement. Ex.1003, ¶¶72-79.

4. Claim 1

[1.0]

First, Bou-Ghazale describes a method for recognizing a pre-defined activation command (“*speech*”) spoken by a user (“*speaker*”) using a voice activation system. Bou-Ghazale’s voice activation system “continuously listen[s] to incoming unconstrained **speech**” in order “to **recognize a pre-defined activation command.**”⁴ Ex.1005, Abstract, 1735; *see also* Ex.1005, Sec.1, 1735. Specifically, recognizing the activation command involves “a speech recognition

⁴ Bou-Ghazale uses the terms “pre-defined activation command,” “pre-selected activation utterance,” “activation word,” “desired word,” and “keyword” interchangeably.

component process[ing] the input speech on a frame-by-frame basis and mak[ing] a decision at every frame whether the **activation word** has been spoken[.]” according to a “recognition algorithm,” as illustrated by Block R in Figure 2 below. Ex.1005, Sec.2, 2.2, 1735-1736. Then, a “postprocessing component validates the **speech recognition result**” according to the steps of the “post-processing **algorithm**” in Figure 2. Ex.1005, Sec.2, 2.3-3.2, 1735-37; Ex.1003, ¶80.

Second, Bou-Ghazale’s method is performed by “*a device*” because Bou-Ghazale teaches that its voice activation system “can be implemented in a **personal communication device.**” Ex.1005, Sec.5, 1738, 1735; Ex.1003, ¶81.

Thus, Bou-Ghazale’s teaching of a method for recognizing a pre-defined activation command spoken by a user using a voice activation system implemented on a personal communication device renders obvious this limitation. Ex.1003, ¶¶80-82.

[1.1]

First, Bou-Ghazale teaches that “[t]he voice activation system continuously **listens to and processes incoming speech[,]**” including speech spoken by “the **user.**” Ex.1005, Sec.1, 1735, Abstract. Bou-Ghazale further explains that when there is speech activity, “[t]he **input speech is sampled** at 8kHz, then filtered by a band-pass filter[,]” and then “[t]he filtered speech is **processed.**” Ex.1005, Sec.2.2, 1736. Thus, Bou-Ghazale’s teachings of (i) listening to incoming speech spoken by a user, (ii) sampling input speech, and (iii) processing speech renders obvious “*receiving...an input speech spoken by the speaker.*” Ex.1003, ¶83.

Second, Bou-Ghazale’s discloses a *preset call command recognition unit* because it teaches a “voice activation system” that “continuously listens to and processes incoming speech” to “**recognize a pre-defined activation command.**” Ex.1005, Abstract. As explained below for [1.3], Bou-Ghazale’s teaching of a

“pre-defined activation command” renders obvious a “*preset call command*.”

Ex.1003, ¶¶84, 95-96. Bou-Ghazale explains that the “voice activation system” is a component of a two-component system that includes the “voice activation system” and the “main speech recognition system.” Ex.1005, 1735 (“Upon recognizing that the activation command has been spoken, **the voice activation system** turns the device on or switches to the **main speech recognition system** to receive further commands from the user.”). A component is a “*unit*” according to the ’196 patent. Ex.1005, Sec.2-2.3, 1735-36; Ex.1001, 5:59-6:3 (“token passing **unit 431 is a component**”).

Thus, Bou-Ghazale’s voice activation system, which is a component for recognizing a pre-defined activation command, that continuously listens to, samples, and processes incoming speech spoken by a user, renders obvious this limitation. Ex.1003, ¶¶83-85.

In the event this limitation (or any other “*unit*” limitation) is construed as a means-plus-function limitation, a combination of hardware and software (Bou-Ghazale’s personal communication device with processor executing speech recognition software) carries out the functions recited in association with the “*preset call command recognition unit*” in [1.1]-[1.4]. Ex.1003, ¶86. This same alternative means-plus-function mapping applies to every “*unit*” term in Challenged Claims 1-20. Ex.1003, ¶¶86, 120, 138, 143, 167, 186.

[1.2]

As explained for [1.1], Bou-Ghazale’s voice activation system renders obvious “*a preset call command recognition unit.*” Ex.1003, ¶87. The voice activation system also receives “*subsequent input speech*” because it “**continuously**” listens to and processes incoming speech. Ex.1005, Sec.1, 1735; Ex.1003, ¶87.

Bou-Ghazale teaches that **all** incoming speech is *received* by the voice activation system, including speech spoken after the activation word (“*subsequent input speech*”) because it teaches that “[t]he voice activation system **continuously listens to and processes incoming speech.**” Ex.1005, Sec.1, 1735, Abstract (“a system designed to continuously **listen to incoming unconstrained speech**”); *see also* Ex.1005, Sec.2.2, 3.1, 3.2, 3.5, Figure 2, 1735-1738; Ex.1003, ¶¶88-93.

Thus, because Bou-Ghazale’s voice activation system “continuously” listens to and processes **all** incoming speech spoken by a user including speech spoken subsequent the activation command, Bou-Ghazale renders obvious this limitation. Ex.1003, ¶¶87-94.

[1.3]

First, regarding the “*preset call command*,” Bou-Ghazale describes a “**pre-defined activation command**” to be recognized by the voice activation system. Ex.1005, Abstract. Bou-Ghazale describes that it can be activated “simply by

speaking a **pre-selected utterance.**” Ex.1005, Sec.1, 1735; Ex.1003, ¶96. As one example, Bou-Ghazale uses the activation word “wake-up.” Ex.1005, Sec.4.1, 4.3-4.4, 1737-38. Thus, Bou-Ghazale’s teaching of a “pre-defined activation command”—also called the “pre-selected activation utterance,” “activation word,” “desired word,” and “keyword”⁵—renders obvious “*a preset call command*” because it is a command that is defined before speech is spoken and is recognized to activate a device or subsequent main speech recognition system. Ex.1003, ¶¶95-96.

Second, to determine whether the activation word has been spoken, Bou-Ghazale teaches that its voice activation system (“*preset call command recognition unit*”) compares incoming speech to “a keyword model for the **desired activation word**” (“*preset call command*”). Ex.1005, Sec.1, 1735; Ex.1003, ¶¶97-99. During the comparison, the speech recognition component (of Bou-Ghazale’s voice activation system) “**processes the input speech on a frame-by-frame basis** and makes a decision at every frame **whether the activation word has been spoken.**”⁶ Ex.1005, Sec.2, 2.2, 1735-1736.

⁵ See *supra* note 4.

⁶ Bou-Ghazale uses the terms “real-time speech recognition system” and “real-time speech recognition component” interchangeably.

Further, Bou-Ghazale explains that its speech recognition component uses a “finite state network” based on “on continuous density hidden Markov models (CD-HMM)” as shown in Figure 1 below. Ex.1005, Sec.2.2, 1736.

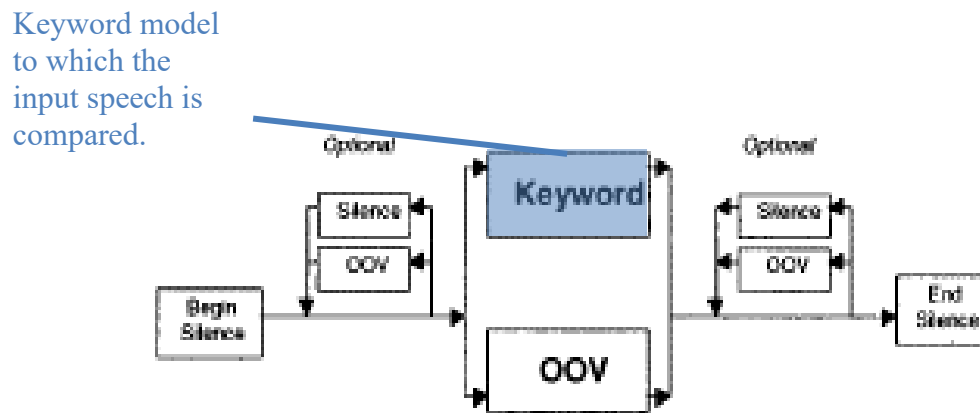


Figure 1. A finite state network representing the allowable word sequences.

Ex.1005, Figure 1 (annotated); Ex.1003, ¶¶99.

Using the finite state network, the speech recognition component compares the acoustic “parameters” of the input speech to identify similarities to the (i) activation word, (ii) silence, or (iii) an OOV word. Ex.1005, Sec.2.2, 1736; Ex.1003, ¶¶97-98. As a result of this comparison, the speech recognition component conveys the input speech frames as a “sequence” or “string” of keyword, silence, or OOV frames, and ultimately determines “the result string with the highest likelihood score”—the string that most likely matches the input speech. Ex.1005, Sec.2.2, 1736; Ex.1003, ¶¶98-99.

Accordingly, during the frame-by-frame processing illustrated in Figure 2

below, Bou-Ghazale's HMM-based speech recognition component *compares* the input speech to the "keyword model for the **desired activation word**" in Figure 1 above. Ex.1005, Sec.1, 1735; Ex.1003, ¶¶97-100.

Compares the input speech to the keyword model for the activation word

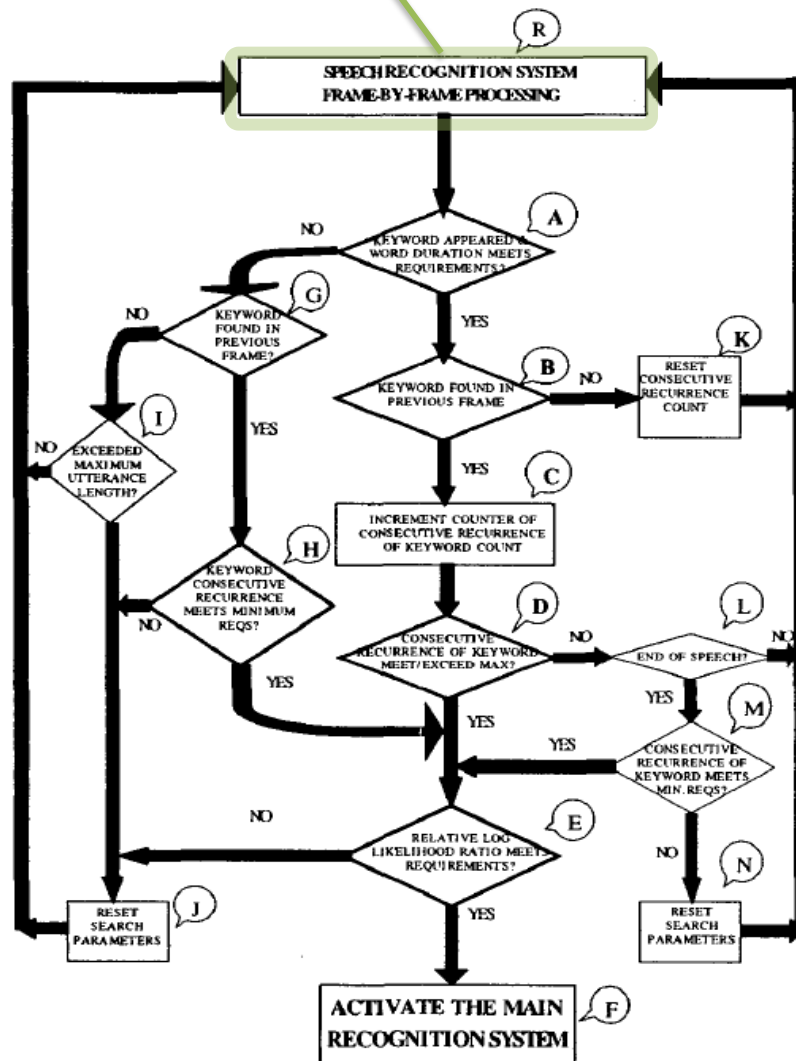


Figure 2. Flow graph of the post-processing algorithm

Ex.1005, Figure 2 (annotated); Ex.1003, ¶101.

Konopka confirms that Bou-Ghazale's speech recognition component *compares*

the input speech to the activation word's keyword model: "the basic principle behind HMM-based speech recognition systems is that **an utterance is compared with a number of speech models**...in order to find the model that best matches the utterance." Ex.1006, [0043]; Ex.1003, ¶101.

Thus, Bou-Ghazale's teaching of an HMM-based speech recognition component performing frame-by-frame processing in which it compares incoming speech to a keyword model for an activation word renders obvious this limitation. Ex.1003, ¶¶95-102.

[1.4]

As explained in association with [1.3], Bou-Ghazale's speech recognition component compares input speech to an activation word's keyword model. If the comparison indicates it is likely that the activation word has been spoken ("*a result of the comparison*"), the result string (which is a representation of the input speech) is passed to a post-processing component to confirm the activation word has been recognized, as shown in Figure 2 below.⁷ Ex.1005, Sec.2, 1735, Sec.2.3, 1736; Ex.1003, ¶103.

⁷ This is consistent with the '196 patent which recognizes the call command in two steps utilizing its token passing unit and its confidence detection unit. *See infra* notes 8, 9.

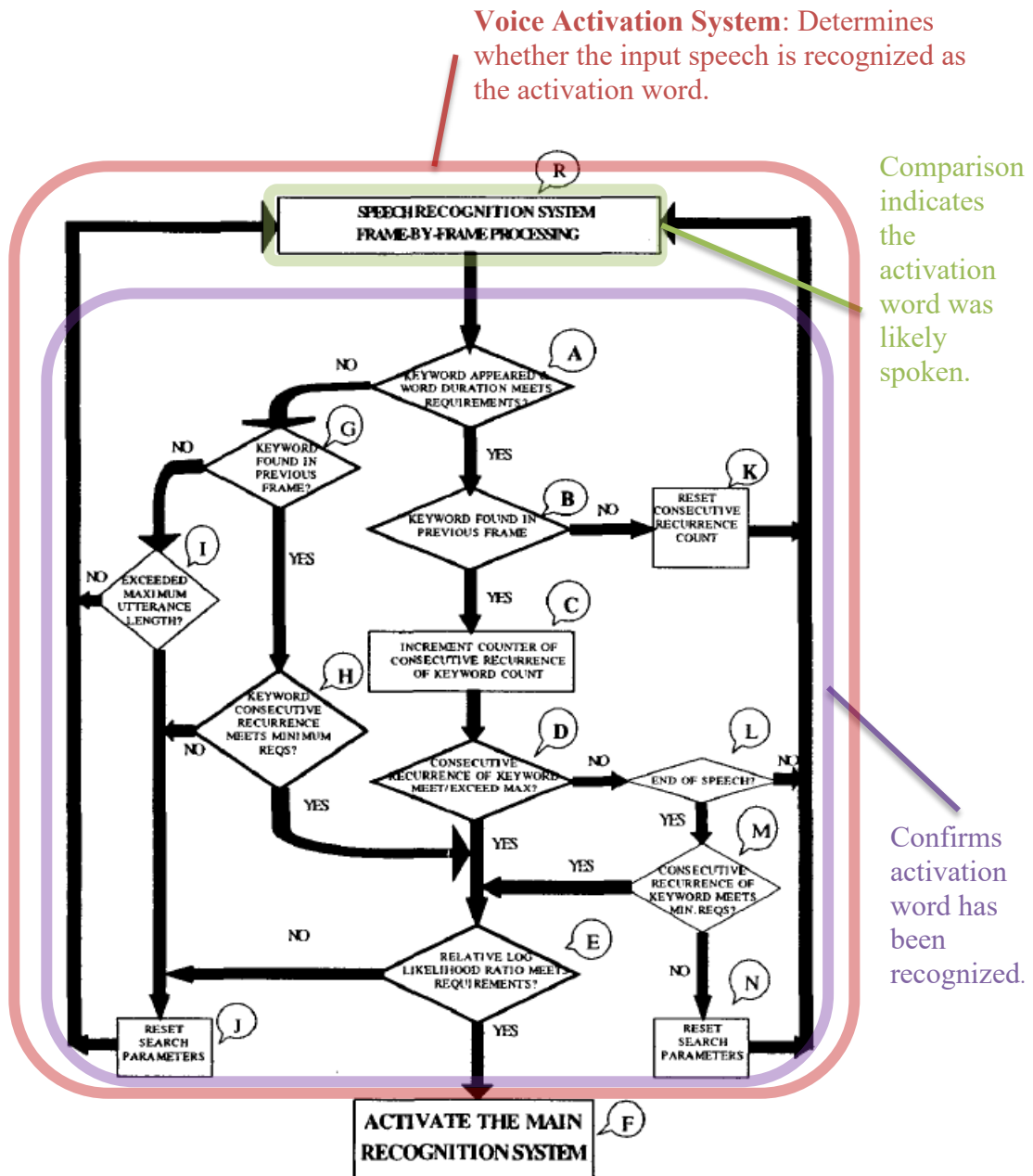


Figure 2. Flow graph of the post-processing algorithm

Ex.1005, Figure 2 (annotated); Ex.1003, ¶103

First, as indicated by the green annotations in Figure 2 above, “the **speech recognition component** processes the input speech on a frame-by-frame basis and **makes a decision** at every frame **whether the activation word has been spoken.**” Ex.1005, Sec.2, 1735. Specifically, Bou-Ghazale teaches that its speech recognition component performs frame-by-frame processing during which it compares incoming speech to a keyword model for an activation word, as explained for [1.3]. Ex.1005, Sec.1-2, 2.2, 1735-36; Ex.1003, ¶104. As a “*result of the comparison*,” Bou-Ghazale’s speech recognition component determines “the **result string** with the highest likelihood score” and “passes [it] to the post-processing component.” Ex.1005, Sec.2.2, 1736; Ex.1003, ¶104.⁸ The “result string with the highest likelihood score” reflects the string that most likely matches the unknown input speech—it indicates whether and which speech frames likely include the activation word. *See* [1.3]; Ex.1006, [0052]; Ex.1003, ¶104.

Moreover, as indicated by the purple annotations in Figure 2 above, Bou-

⁸ This is consistent with the ’196 patent which explains that “every input speech frame is compared with a recognition network for only one word in time” and based on the tokens with the highest accumulated probability the token passing unit “assume[s] that an actual call command has been spoken.” Ex.1001, 6:55-57, 6:62-7:3, 7:34-39; Ex.1003, ¶104.

Ghazale teaches that the post-processing component next verifies whether the input speech matches the activation word, utilizing the result from the speech

recognition comparison: “the **post-processing component** decodes the **result string** on a frame-synchronous basis and applies a set of criteria and confidence measures **to decide whether the activation word was a part of the utterance**”.

Ex.1005, Sec.2.3, 1736, Abstract (“**The post-processing component validates the recognition results** by applying a set of criteria and confidence measures.”). If the input speech—represented by the result string—meets all the criteria associated with the pre-defined activation word, then “**the activation word is declared to be found**” (“*recognized*”) in the input speech by the post-processing component.

Ex.1005, Sec.2.3, 1736 (“**The activation word is declared to be found only after it meets a set of statistics** derived from training data such as the overall duration of the keyword, the number of times the keyword appears in the result string of consecutive frames, and the confidence level associated with the keyword.”).⁹ With reference to Figure 2 below, Bou-Ghazale explains that “[i]n the path traversing A, B, C, D, E, F, the keyword has met all the necessary requirements, and hence **the**

⁹ This is consistent with the ’196 patent which describes a confidence detection unit that determines “[w]hen the confidence of each of the portions is greater than 0...the call command ‘Nepiya’ is recognized.” Ex.1001, 8:28-38; Ex.1003, ¶106.

post-processor declares that the activation word has been found” and “[i]n the path A, B, C, D, L, M, E, F...the post-processor declares that the activation word has been found.” Ex.1005, Sec.3.2, 1737.

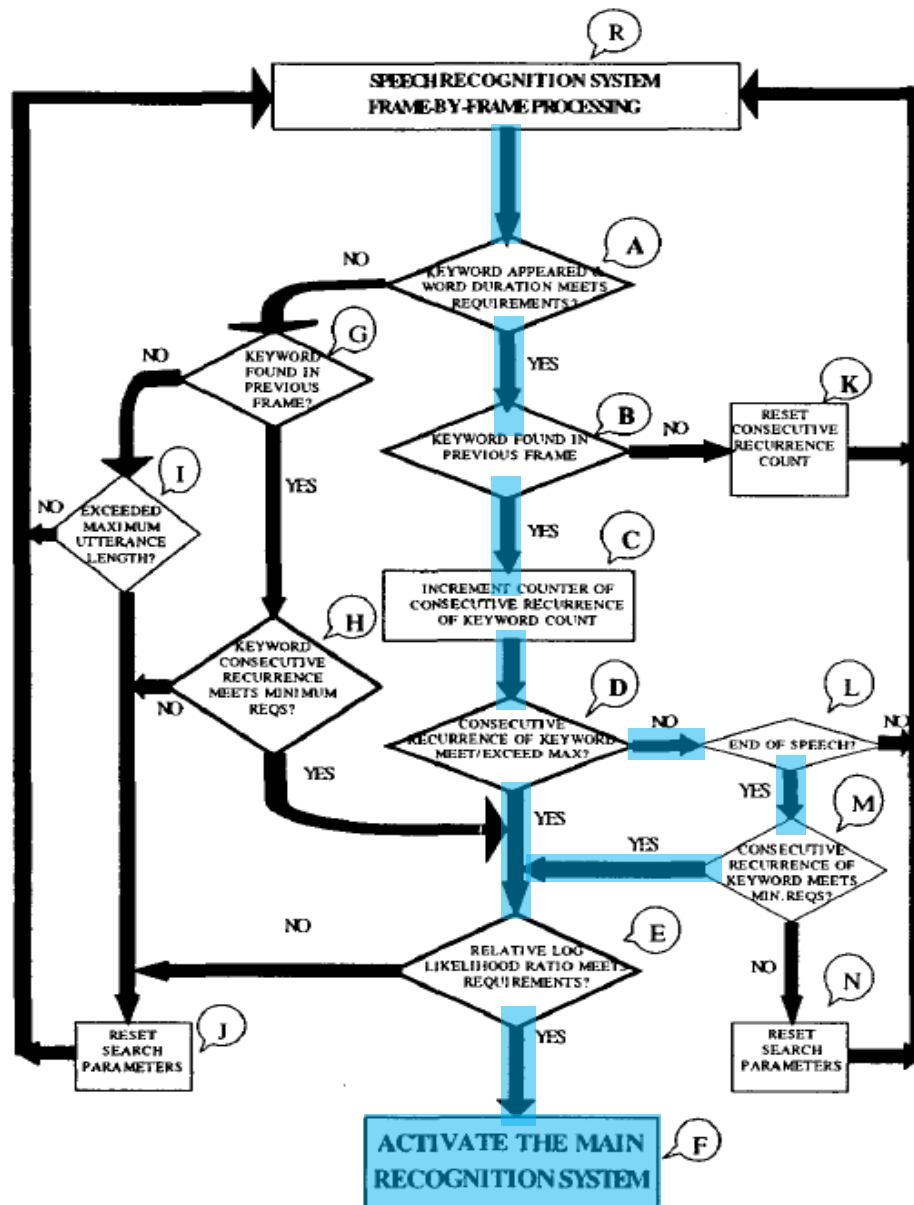


Figure 2. Flow graph of the post-processing algorithm
Ex.1005, Figure 2 (annotated); Ex.1003, ¶106.

Thus, Bou-Ghazale’s voice activation system determining whether the

activation word has been recognized in the input speech based on (i) the result string that represents the most likely keyword match to the unknown input speech as determined during the comparison step of [1.3] (“*result of the comparison*”) and (ii) whether the input speech meets the criteria and confidence measurements associated with the activation word renders obvious this limitation. Ex.1003, ¶¶103-107.

[1.5]

First, Bou-Ghazale teaches that once the activation word is declared as found in the input speech (“*based on determining that the input speech is recognized as the preset call command*”), a main speech recognition system is activated (“*activating a continuous speech recognition unit*”).

As explained in [1.4], Bou-Ghazale’s post-processing component determines whether the activation word has been found in the input speech based on whether the input speech (represented by the result string) meets the criteria and confidence measurements associated with the activation word. Ex.1005, Sec.2.3, 1736; Ex.1003, ¶109. Bou-Ghazale explains that “only after it meets [the] set of statistics, such as...the confidence level associated with the keyword” the activation word is declared to be found (i.e., recognized). *Id.*

Specifically, in Figure 2 below, Bou-Ghazale shows that “[t]here are three paths...that meet all the requirements and hence result in **activating the main**

speech recognition system.” Ex.1005, Sec.3.2, 1737. As one example, “[i]n the path traversing A, B, C, D, E, F, the keyword has met all the necessary requirements, and hence **the post-processor declares that the activation word has been found**” and the main speech recognition system is activated at Block F, as shown in Figure 2 below. Ex.1005, Sec.3.2, Figure 2, 1737; Ex.1003, ¶108.

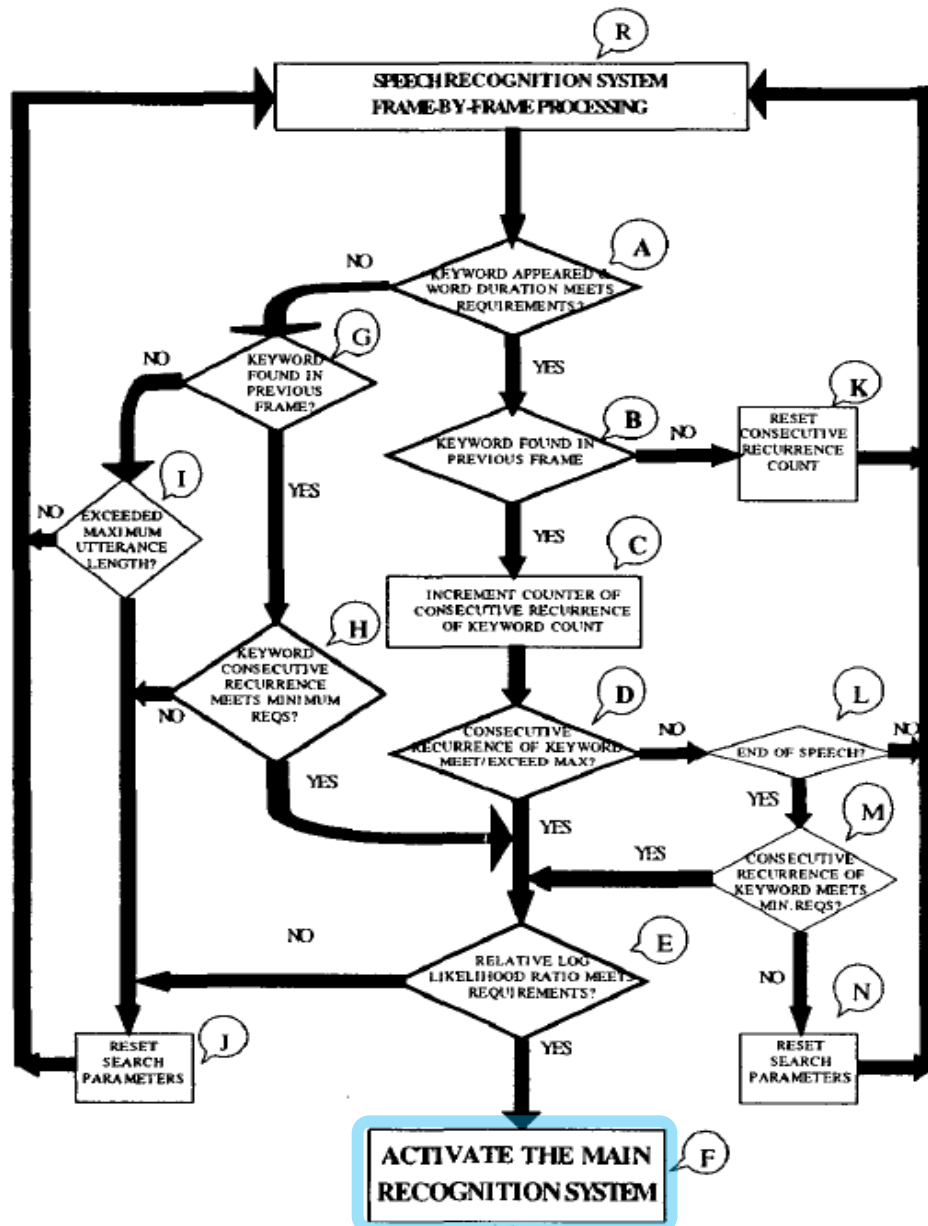


Figure 2. Flow graph of the post-processing algorithm

After the activation word is declared found, the main speech recognition system is activated.

Ex.1005, Figure 2 (annotated); Ex.1003, ¶108.

Bou-Ghazale further explains that “once the activation word is recognized, the device turns on or switches to the main recognition system.” Ex.1005, Sec.5,

1738, Sec.1, 1735.

Thus, Bou-Ghazale’s teaching of activating a main speech recognition system, based on the determination that the activation word has been found in the input speech when it meets the statistical criteria for the activation word, renders obvious “*based on determining that the input speech is recognized as the preset call command, activating a continuous speech recognition unit.*” Ex.1003, ¶¶108-111.

Second, Bou-Ghazale’s main speech recognition system is the claimed *continuous speech recognition unit* because, when activated by the activation word, it continuously “waits” to receive valid command words spoken *subsequent* the activation word for a period of time before deactivating. Ex.1005, Sec.5, 1738 (“once the activation word is recognized, the device turns on or **switches to the main recognition system and waits to be prompted by a word from the main menu of the speech recognition system**” and “**if no valid word is spoken in the following 10-15 seconds**, then the system switches back to the stand-by mode and waits to be activated again”); *see also* [1.2] regarding “*subsequent input speech.*” Bou-Ghazale’s main speech recognition system is “*continuous*” in the same manner as the ’196 patent’s “*continuous speech recognition system*”—both continuously recognize user commands only when activated. Ex.1001, 8:32-41 (explaining that after the “call command” is recognized, continuous speech

recognition is performed and when “it is determined that the call command is not spoken...the process returns to the initial state.”); Ex.1003, ¶112.

Further, Bou-Ghazale explains that the “main speech recognition system” is a component of a two-component system that includes the “voice activation system” and the “main speech recognition system.” Ex.1005, 1735 (“Upon recognizing that the activation command has been spoken, **the voice activation system** turns the device on or switches to the **main speech recognition system** to receive further commands from the user.”). A component is a “*unit*” according to the ’196 patent. Ex.1005, Sec.2-2.3, 1735-36; Ex.1001, 5:59-6:3 (“token passing **unit 431 is a component**”); Ex.1003, ¶113.

Third, Bou-Ghazale’s main speech recognition system is configured to determine whether speech received after the activation word is recognized as a word from the main menu of the speech recognition system (“*actual command*”). Bou-Ghazale explains that after the activation command is recognized, the device switches to “the main speech recognition system to receive **further commands** from the user”—i.e., speech spoken *subsequent* the activation word. Ex.1005, Sec.1, 1735; *see also* [1.2] regarding “*subsequent input speech*”; Ex.1003 ¶¶88-93 (explaining how the voice activation system would receive “further commands” because it “continuously” receives all input speech). Once the device has switched to (i.e., activated) the main speech recognition system, it “**waits to be prompted**

by a word from the main menu of the speech recognition system.” Ex.1005, Sec.5, 1738. Bou-Ghazale’s main speech recognition system either determines that the further command includes one of the words from its menu, or “if no valid word is spoken in the following 10-15 seconds, then the system switches back to the stand-by mode and waits to be activated again.” Ex.1005, Sec.5, 1738; Ex.1003, ¶114. Thus, Bou-Ghazale’s teaching of a main speech recognition system, configured to determine whether further commands (“*subsequent input speech*”) include a word from its menu (“*actual command*”) renders obvious “*a continuous speech recognition unit configured to determine whether the subsequent input speech is recognized as an actual command.*” Ex.1003, ¶114.

Fourth, Bou-Ghazale teaches that the main speech recognition system only activates when the activation word is found “*in order to reduce the computational load on the device.*” As discussed above, Bou-Ghazale explains that “the activation word is declared to be found **only after** it meets a set of statistics.” Ex.1005, Sec.2.3, 1736. For example, as shown by the highlighted paths in Figure 2 below, if the input speech does not meet some of the “statistics,” then the voice activation system will “terminate the search and start again” and the main speech recognition system will not be activated. Ex.1005, Sec.3.2, Figure 2, 1736-37. Further, Bou-Ghazale’s main speech recognition system has a “menu” of words to which it compares speech to (Ex.1005, Sec.5, 1738), while its voice activation system only

compares speech to one “keyword model” (Ex.1005, Sec.1, 2.2, Figure 1, 1735-1736). Accordingly, a POSITA would understand that Bou-Ghazale’s system “*reduce[s] the computational load on the device*” because it uses a smaller vocabulary voice activation system requiring less computational load, and only activates its larger vocabulary main speech recognition system when the activation word is found. Ex.1003, ¶¶115-117.

If the statistical criteria is not met, then the voice activation system will not activate the main speech recognition system.

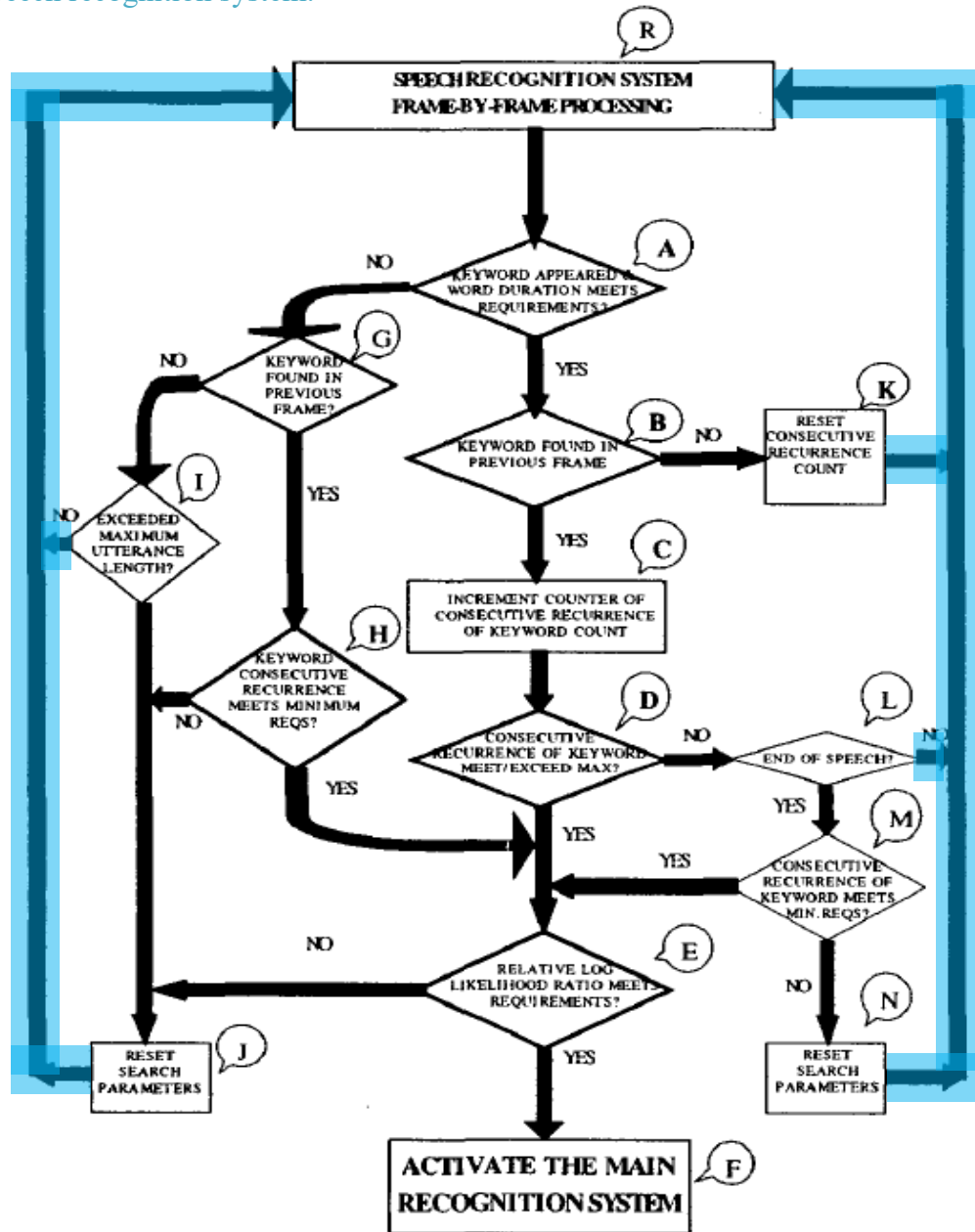


Figure 2. Flow graph of the post-processing algorithm

Ex.1005, Figure 2 (annotated); Ex.1003, ¶117.

Moreover, Bou-Ghazale explains that its voice activation system “**requires relatively low power** and can be implemented in a personal communication device

due to its **low memory and MIPS requirements (< 1 MIPS).**” Ex.1005, Sec.5, 1738. It was well-known at the time of the ’196 patent that “MIPS” is an acronym for “Million Instructions Per Second” as a measurement of a processor’s speed. Ex.1003, ¶118 (citing Exs.1012-14). Accordingly, Bou-Ghazale’s voice activation system that operates at a lower processing speed, renders obvious that it is designed to have a lower computational load to process. Ex.1003, ¶118.

Thus, because Bou-Ghazale teaches that (i) once the activation word is determined to have been found in the input speech, a main speech recognition system is activated for determining whether subsequent speech matches a word in the main speech recognition system, and (ii) Bou-Ghazale’s system requires less power and processing speed because the larger vocabulary main speech recognition system only activates when the statistical criteria is met and the activation word is recognized, Bou-Ghazale renders obvious this limitation. Ex.1003, ¶¶108-120.

[1.6]

First, regarding “*token passing unit*,” Bou-Ghazale teaches that its voice activation system comprises a speech recognition component.¹⁰ Ex.1005, Sec.2, 1735 (also explaining that “**the voice activation system is comprised of... a**

¹⁰*See supra* note 6.

speech recognition algorithm”). Bou-Ghazale’s “speech recognition component” is a “*unit*” as claimed. Ex.1003, ¶122; Ex.1001, 5:59-6:3 (“token passing **unit 431 is a component**”).

Bou-Ghazale in view of Konopka renders obvious that Bou-Ghazale’s speech recognition component is the “*token passing unit*.” Ex.1003, ¶¶123-127. Bou-Ghazale explains that its speech recognition component “is based on continuous density hidden Markov models (CD-HMM).” Ex.1005, Sec.2.2, 1736, Abstract. Further, the speech recognition component “performs a trace-back at every frame, and **passes the result string with the highest likelihood score** to the post-processing component.” Ex.1005, Sec.2.2, 1736; Ex.1003, ¶¶123-124.

A POSITA would have found it obvious for Bou-Ghazale’s HMM-based speech recognition component to use “token passing,” as taught by Konopka. Ex.1003, ¶125. Konopka teaches that “**token passing is a well-known approach to tracking the best word hypotheses through an HMM.**” Ex.1006, [0052]. Konopka explains that this “**token passing technique provides a highly memory efficient solution to an otherwise well-known problem in HMM-based speech recognition systems: the storage of back-pointers.**” Ex.1006, [0052].

For at least the reasons explained above in §IX.C.3., a POSITA would have been motivated to implement Konopka’s token passing technique into Bou-Ghazale’s HMM-based speech recognition component. Ex.1003, ¶126.

Specifically, a POSITA would have been motivated to discard the strings that are less likely to match the unknown speech during the recognition process in Bou-Ghazale, to more efficiently and effectively determine the result string that most likely matches the unknown speech (“result string with the highest likelihood score”) and pass it to the post-processing component. Ex.1003, ¶126.

Thus, Bou-Ghazale’s teaching of a voice activation system comprising an HMM-based speech recognition component in view of Konopka’s teaching of a well-known token passing technique for use with HMM-based speech recognition systems renders obvious “*wherein the preset call command recognition unit comprises a token passing unit.*” Ex.1003, ¶¶122-127.

Second, Bou-Ghazale teaches that its speech recognition component assumes the activation word is in the input speech according to the result string with the highest likelihood score (“*configured to assume the input speech to be the preset call command according to the result of the comparison*”). As explained for [1.3], Bou-Ghazale teaches that its speech recognition component compares incoming speech to “a keyword model for the **desired activation word**” to determine whether the activation word has been spoken. Ex.1005, Sec.1, 1735, Sec.2.2, 1736. As a “*result of the comparison*,” Bou-Ghazale’s speech recognition component determines “the **result string with the highest likelihood score**” and “passes [it] to the post-processing component”. Ex.1005, Sec.2.2, 1736; Ex.1003,

¶128.

According to the token passing technique taught by Konopka, as explained above, Bou-Ghazale’s “result string with the **highest likelihood score**” corresponds to the “**best** word sequence” that is determined to be closest to the unknown speech. Ex.1006, [0052], [0043]; Ex.1003, ¶129. Therefore, if the likelihood score associated with the result string for the keyword model is higher, then the input speech is most similar to the activation word and the speech recognition component will “assume” that the input speech is the activation word. Ex.1005, Sec.2.2, Figure 1, 1736. Accordingly, the “result string with the highest likelihood score” passed to the post-processing component would also convey that the input speech is “*assume[d]*” to include the activation word, as shown by the green annotations in Figure 2 below. Ex.1003, ¶¶130-131.

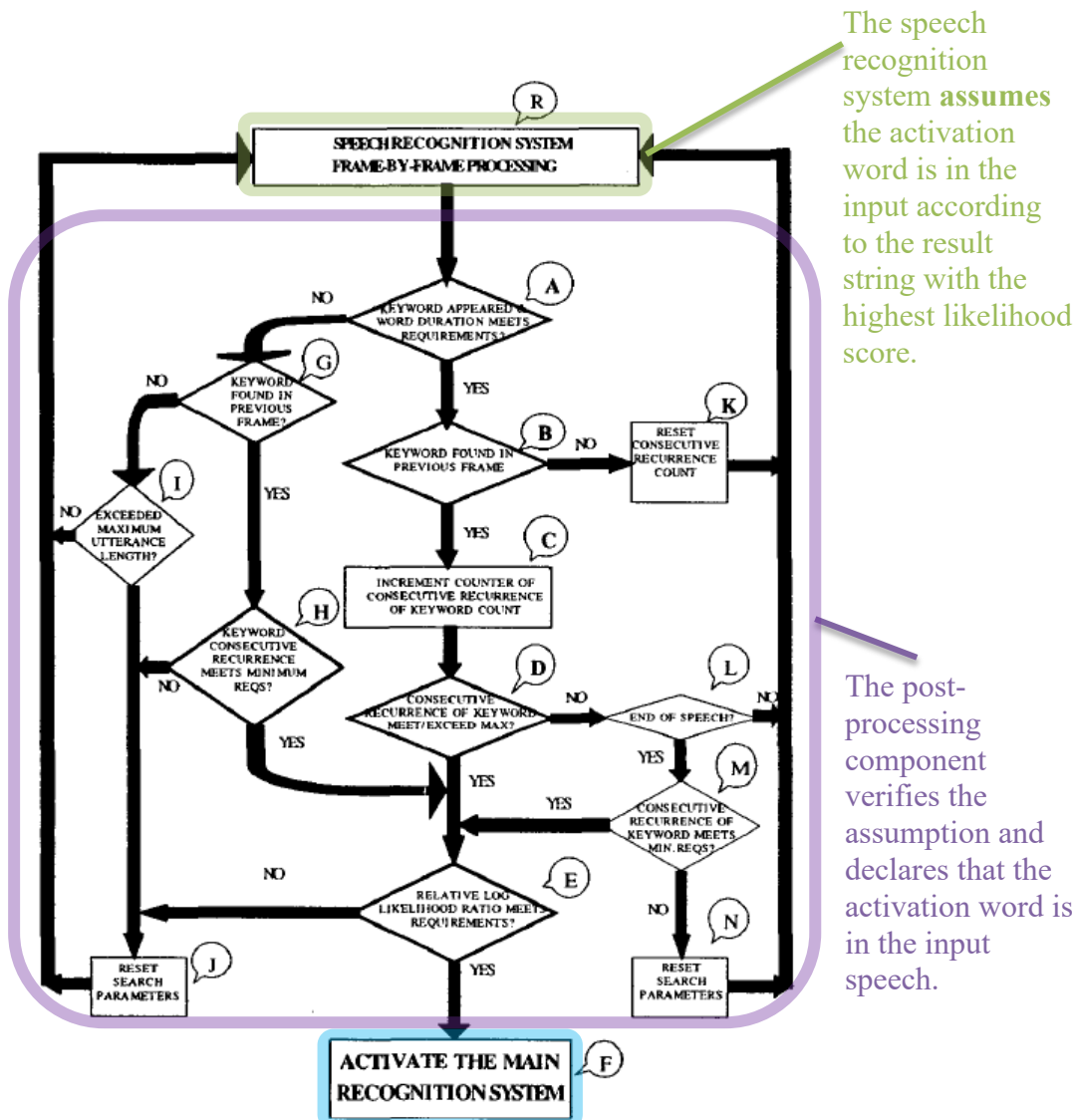


Figure 2. Flow graph of the post-processing algorithm

Once the activation word is declared found, the main recognition system is activated

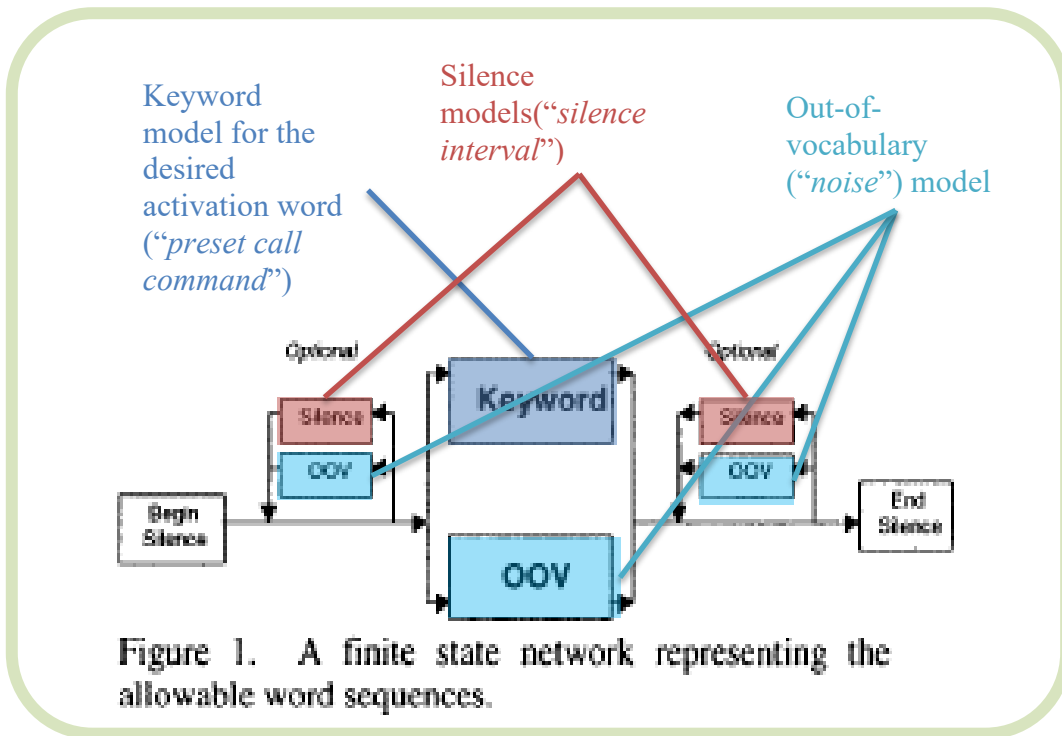
Ex.1005, Figure 2 (annotated); Ex.1003, ¶131.

Thus, Bou-Ghazale's teaching of a speech recognition component that assumes the activation word is in the input speech according to the result string

with the highest likelihood score, as determined using the token passing technique of Konopka, renders obvious a “*token passing unit configured to assume the input speech to be the preset call command according to the result of the comparison.*” Ex.1003, ¶¶128-132.

Third, Bou-Ghazale teaches that its speech recognition component uses “*a minimum recognition network composed of a silence interval accompanied by noise and the preset call command.*” Specifically, the speech recognition system uses a “finite state network” that includes a keyword model for the desired activation word (“*preset call command*”), a model for out-of-vocabulary words (“*noise*”), and a model for silence (“*silence interval*”): “The **finite state network** consists of optional **silence and OOV models** at the beginning and end, and a parallel network of the **activation keyword** and one **OOV model**,” as shown in Figure 1 below. Ex.1005, Sec.2.2, 1736, (“the recognition system uses a finite state network, as shown in Figure 1, to represent the allowable sequences of **silence, keyword, and out-of-vocabulary (OOV) words**”); Ex.1003, ¶133.

Finite state network (“*minimum recognition network*”)



Ex.1005, Figure 1 (annotated); Ex.1003, ¶133.

In Figure 1 above, Bou-Ghazale shows only one keyword model in its finite state network. *See also* Ex.1005, Sec.2.2, 1736 (“the recognizer can detect **the word** whether it is spoken in isolation or embedded in extraneous speech”). Bou-Ghazale’s teachings are consistent with the example of a “*minimum recognition network*” provided in the ’196 patent which recognizes only one word, or a minimum number of tokens, in recognizing a call command. Ex.1001, 6:60-63 (“Since the recognition network is a recognition network for only one word, a minimum number of (20 to 30) tokens are required.”); Ex.1003, ¶134.

Accordingly, a POSITA would recognize that Bou-Ghazale’s finite state network

renders obvious a “*minimum recognition network*” because it is a network for recognizing one keyword, rather than a sequence of words. Ex.1003, ¶134.

A POSITA would also recognize that the out-of-vocabulary (OOV) models in Bou-Ghazale’s finite state network renders obvious “*noise*.” Ex.1003, ¶¶135-136. Bou-Ghazale explains that its voice activation system “can be realized by training a keyword model for the desired activation word, **a garbage model representing** both extraneous speech and **background noises**, and a model for background silence.” Ex.1005, Sec.1, 1735. Bou-Ghazale also explains that the OOV models “can be repeated many times to represent all **extraneous...speech** before or after the keyword, or the **utterances that do not include the keyword**.” Ex.1005, Sec.2.2, 1736. Accordingly, Bou-Ghazale’s teaching of “background noises,” “extraneous speech,” and “utterances that do not include the keyword” discloses “*noise*.” This is also consistent with the ’196 patent which discloses that noise may be “spoken voices.” Ex.1001, 5:30-34 (“...surrounding noise from **noise-containing input speech (all spoken voices including call commands)** and eliminating the surrounding noise...”).

In summary, (i) Bou-Ghazale’s teaching of a voice activation system comprising an HMM-based speech recognition component in view of Konopka’s teaching of a token passing technique to determine the result string with the highest likelihood, and (ii) Bou-Ghazale’s teaching that the speech recognition component

“assumes” that the activation word is included in the input speech according to the result string with the highest likelihood score and using a finite state network for recognizing one word that consists of models for the keyword, for out-of-vocabulary words (e.g., extraneous speech and background noises), and silence, renders obvious this limitation. Ex.1003, ¶¶121-138.

5. Claim 2

[2.0]

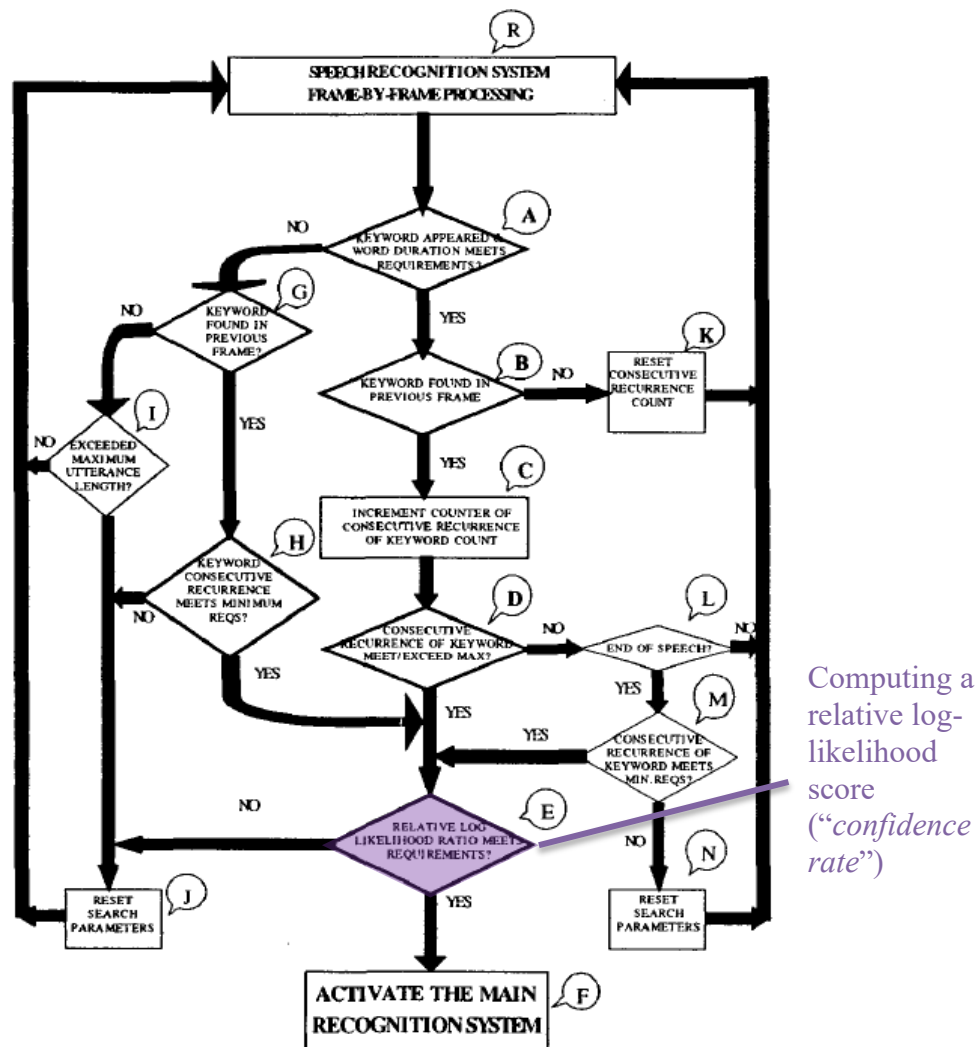
First, regarding “*a confidence detection unit*,” Bou-Ghazale teaches a post-processing component that verifies whether the activation word has been found in the input speech by determining a confidence rate for the input speech, called the “relative log-likelihood score” as discussed below. Ex.1005, Sec.2, Abstract, 1735, Sec.2.3-3.1, 1736; Ex.1003, ¶139. Further, Bou-Ghazale’s “post-processing component” is a “*unit*” as claimed. Ex.1003, ¶139; Ex.1001, 5:65-67 (“the confidence detection **unit** 433 is a **component**”).

Second, Bou-Ghazale discloses “*computing...a confidence rate for the input speech*” because it teaches that the post-processing component determines a “relative log-likelihood score” for the input speech to determine whether the input speech is recognized as the activation word. Ex.1005, Sec.2.3-3.1, 1736; Ex.1003, ¶¶140-141. As explained for [1.4], Bou-Ghazale’s post-processing component verifies whether the activation word has been found in the input speech

(“determining whether the input speech is recognized as the preset call command”) based on whether the input speech meets the criteria and confidence measurements associated with the activation word. Ex.1005, Sec.2.3, 1736 (“the **post-processing component** decodes the **result string** on a frame-synchronous basis and **applies a set of criteria and confidence measures to decide whether the activation word was a part of the utterance**”). For example, for one such criteria illustrated by Block E in Figure 2 below, Bou-Ghazale teaches that its post-processing component determines a “relative log-likelihood score” (also referred to as a “relative log likelihood ratio”¹¹) for the input speech and compares it to the “confidence level associated with the keyword,” where the confidence level is represented as a threshold value, η . Ex.1005, Sec.2.3, Sec.3.1 (“Criterion 6: The **relative log-likelihood score should be equal to or greater than a threshold, η** , given by: $\eta = (\text{Score}_{\text{keyword}} - \text{Score}_{\text{OOV}})/|\text{Score}_{\text{keyword}}|$, where the accumulated log-likelihood score for the keyword model is denoted by $\text{Score}_{\text{keyword}}$, and the log-likelihood score for the OOV model is denoted by

¹¹ Bou-Ghazale uses the terms “log-likelihood score” and “log likelihood ratio” interchangeably. Ex.1005, Sec.3.1, 1736 (“relative log-likelihood **score**” in Criterion 6), Sec.3.2, Figure 2 (Block E uses “relative log likelihood **ratio**” with reference to Criterion 6); Ex.1003, ¶140.

Score_{OOV}.”), Sec.3.2, 1736. A POSITA would have understood that Bou-Ghazale’s “relative log-likelihood score” as defined in Criterion 6 is a “*confidence rate*” as claimed because it is compared to another confidence—the “confidence level associated with the keyword” represented by the threshold, η . This is consistent with the ’196 patent which explains that a “confidence rate” is a “log-likelihood ratio,” having an equation similar to that of Bou-Ghazale, that is compared to a “threshold” value. Ex.1005, Sec.2.3-3.1, 1736; Ex.1001, 7:63-8:13; Ex.1003, ¶141 (citing Ex.1007).



Ex.1005, Figure 2 (annotated); Ex.1003, ¶141.

Thus, Bou-Ghazale’s teaching of determining whether the activation word has been found in the input speech by determining, using the post-processing component, a relative log-likelihood score that is compared to a threshold value, η renders obvious this limitation. Ex.1003, ¶¶139-143.

[2.1]

Bou-Ghazale explains that “**the activation word is declared to be found only after it meets** a set of statistics derived from training data such as...**the**

confidence level associated with the keyword.” Ex.1005, Sec.2.3, 1736. More specifically, Bou-Ghazale explains that the “relative log-likelihood score **should be equal to or greater than a threshold, η ,**” where the threshold, η is the confidence level for the activation word and the relative log-likelihood score is the “*computed confidence rate*”, as explained for [2.0]. Ex.1005, Sec.2.3-3.1, 1736; Ex.1003, ¶144. As shown in Figure 2 below, if the “**relative log likelihood ratio** meets requirements,” then the post-processing component “**declares that the activation word has been found**” and activates the main speech recognition system. Ex.1005, Sec.3.2, 1736-37; *see also* Ex.1001, 7:63-8:13; Ex.1003, ¶145.

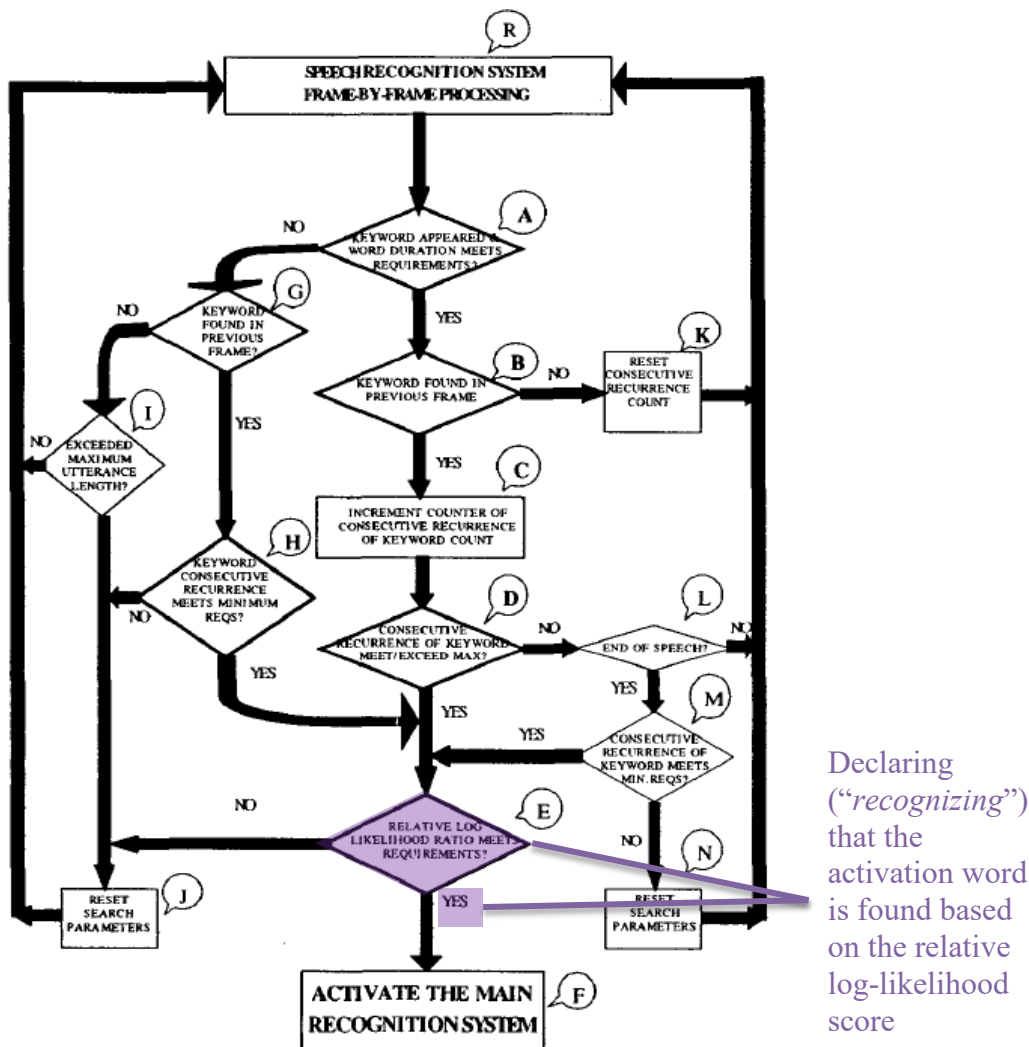


Figure 2. Flow graph of the post-processing algorithm

Ex.1005, Figure 2 (annotated); Ex.1003, ¶145.

Thus, Bou-Ghazale's teaching of the post-processing component declaring that the activation word has been found ("recogniz[ed]") in the input speech based on the relative log-likelihood score ("computed confidence rate") renders obvious this limitation. Ex.1003, ¶¶144-146.

6. Claim 3

As explained for [2.0]-[2.1], Bou-Ghazale's post-processing component

“declares that the activation word has been found” in the input speech if the relative log-likelihood score (“*computed confidence rate*”) is “equal to or **greater than a threshold, η** .” Ex.1005, Sec.2.3-3.2, 5, Figure 2, 1736-1738; Ex.1003, ¶¶147-148.

7. Claim 4

[4.0]

As explained for [2.0]-[2.1], Bou-Ghazale’s post-processing component “declares that the activation word has been found” (i.e., recognized) based on whether the relative log-likelihood score of the input speech (“*computed confidence rate*”) is “equal to or greater than a threshold, η .” Ex.1005, Sec.2.3-3.2, 1736-37. Bou-Ghazale explains that the threshold, η may be adjusted to achieve lower “false alarm rates”—i.e., a lower number of commands that are falsely recognized as the activation word. Ex.1005, Sec.4.4, 1738 (“**As the threshold value is increased**, the false rejection rate increases and **the false alarm rate decreases.... For achieving very low false alarm rates** of 0.26% for isolated utterances and 0% for embedded utterances, **η can be set to 0.007.**”); Ex.1003, ¶149. Because Bou-Ghazale describes false alarm rates as a “function of η ,” a POSITA would have understood that when the post-processing component compares the relative log-likelihood score of the input speech to the threshold, η , it is determining whether the input speech is a “false alarm” (false command).

Ex.1005, Tables 2-3; Ex.1003, ¶¶149-150 (citing Ex.1007). If the relative log-likelihood score is equal to or greater than the threshold, η , then the post-processing component determines that the input speech is an “*authentic command*” and not a false alarm, and “declares that the activation word has been found.”

Ex.1005, Sec.3.1-3.2, Figure 2, 1736-37; Ex.1003, ¶150.

Thus, Bou-Ghazale’s teaching of determining whether the activation word is a false alarm and is found based on whether the relative log-likelihood score of the input speech meets a threshold, η , renders obvious this limitation. Ex.1003, ¶¶149-151.

[4.1]

As explained for [4.0], Bou-Ghazale’s post-processing component determines whether the activation word is an “*authentic command*” or a false alarm based on whether the relative log-likelihood score of the input speech meets the threshold requirement, η , as shown by the inquiry in Block E of Figure 2 below. Ex.1005, Sec.3.1-3.2, 1736-37. If the relative log-likelihood score is “equal to or greater than a threshold, η ” then the post-processing component determines that the input speech is an “*authentic command*” and not a false alarm (“yes” path following Block E), and “**declares that the activation word has been found**” (“*recogniz[ed]*”). Ex.1005, Sec.3.1-3.2, 1736-37; Ex.1003, ¶152.

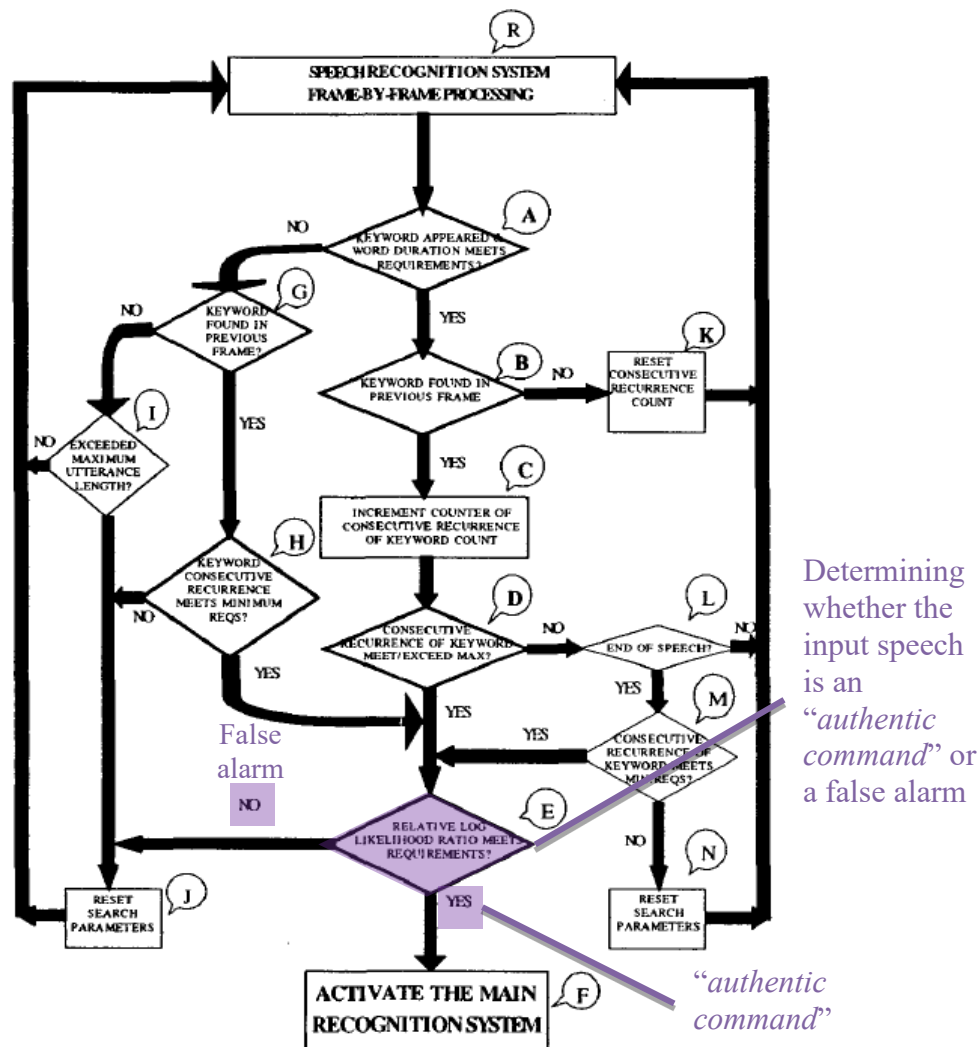


Figure 2. Flow graph of the post-processing algorithm

Ex.1005, Figure 2 (annotated); Ex.1003, ¶152.

Thus, Bou-Ghazale’s teaching of the post-processing component declaring that the activation word has been found if it is determined that the input speech is not a false alarm because the relative log-likelihood score is greater than the threshold, η renders obvious this limitation. Ex.1003, ¶¶152-153.

8. Claim 5

[5.0]

First, as explained for [1.4], Bou-Ghazale’s voice activation system determines whether the activation word has been found (“*determining the input speech is recognized as the preset call command*”) in the input speech based on “the result string with the **highest likelihood score**” as determined by the speech recognition component during the comparison step [1.3]. Ex.1005, Sec.2.2, 1736; Ex.1003, ¶154.

Second, regarding “*log*,” to the extent Bou-Ghazale doesn’t explicitly state that its “likelihood score” determined by the speech recognition component is a “logarithm”, it would have been obvious to a POSITA because it was well-known to compute a logarithm of a likelihood score for a keyword model to determine the best strings that most likely match input speech. Ex.1003, ¶¶155-157.

For example, Konopka teaches an equation for determining the best text strings to be passed to a natural language interface module for post-processing: “[**Ln(IV score)**–Ln(OOV score)]/number of frames,” where Ln(IV score) is a natural logarithm of the “**in-vocabulary (IV) Viterbi (likelihood) score**” and Ln(OOV score) is a natural logarithm of the “**out-of-vocabulary (OOV) Viterbi score.**” Ex.1006, [0080], Figure 1. While this equation shows a logarithm of an in-vocabulary score and an out-of-vocabulary score, Konopka explains that it was

well-known to rely on only the in-vocabulary score to determine the best text strings (“state sequence”). Ex.1006, [0043] (explaining that “if a **Viterbi score passes a threshold, the utterance is determined to be an in-vocabulary word**....Such a Viterbi score is generated using the Viterbi algorithm. This algorithm calculates **a single best state sequence through an HMM and its corresponding probability**, given an observation sequence” where “Viterbi score” here refers to an in-vocabulary score); Ex.1003, ¶¶155-156. Accordingly, a POSITA would have understood that Bou-Ghazale’s “highest likelihood score” as determined by the speech recognition component may be a “*log*” likelihood score for its keyword model. Ex.1003, ¶¶155-156, 157 (also explaining that it was common practice to compute a “*log*” to reduce computational requirements).

Third, regarding “*ratio*,” regardless of the terminology used, Bou-Ghazale (as modified by Konopka) teaches that it computes a logarithm function for a single model—a log likelihood score for its keyword model—in a first recognition stage as explained above, and thus matches the example of the first “*log likelihood ratio*” given in the ’196 patent which is a logarithm function for a single model:

“ **$LLR_k(0, \lambda_k) = \log p(0 | \lambda_k)$ (1) where $LL[R]$ is the log likelihood ratio, λ_k is a k-th model parameter, and $p(0 | \lambda_k)$ is the probability of the λ_k model of the input speech frame.**” Ex.1005, Sec.2.2, 1736; Ex.1006, [0080], [0043]; Ex.1001, 7:5-24; Ex.1005, Sec.3.1-3.2, Figure 2, 1736 (using the terms “log-likelihood score” and

“log likelihood ratio” interchangeably); Ex.1003, ¶158.

Thus, Bou-Ghazale’s teaching of determining whether the activation word has been found in the input speech by determining “the result string with the **highest likelihood score**,” in light of Konopka’s teachings of computing a logarithm of an in-vocabulary Viterbi (likelihood) score to determine the best text strings, renders obvious this limitation. Ex.1003, ¶¶154-159.

[5.1]

Bou-Ghazale’s speech recognition component assumes the input speech to be the activation word according to “the result string with the highest likelihood score” that is passed to the post-processing component, as explained for [1.6]. Ex.1003, ¶160. Bou-Ghazale’s “highest likelihood score” renders obvious the “*computed log likelihood ratio*” considering Konopka’s teachings of computing a logarithm of an in-vocabulary likelihood score to determine the best text strings, as explained for [5.0]. Ex.1003, ¶160.

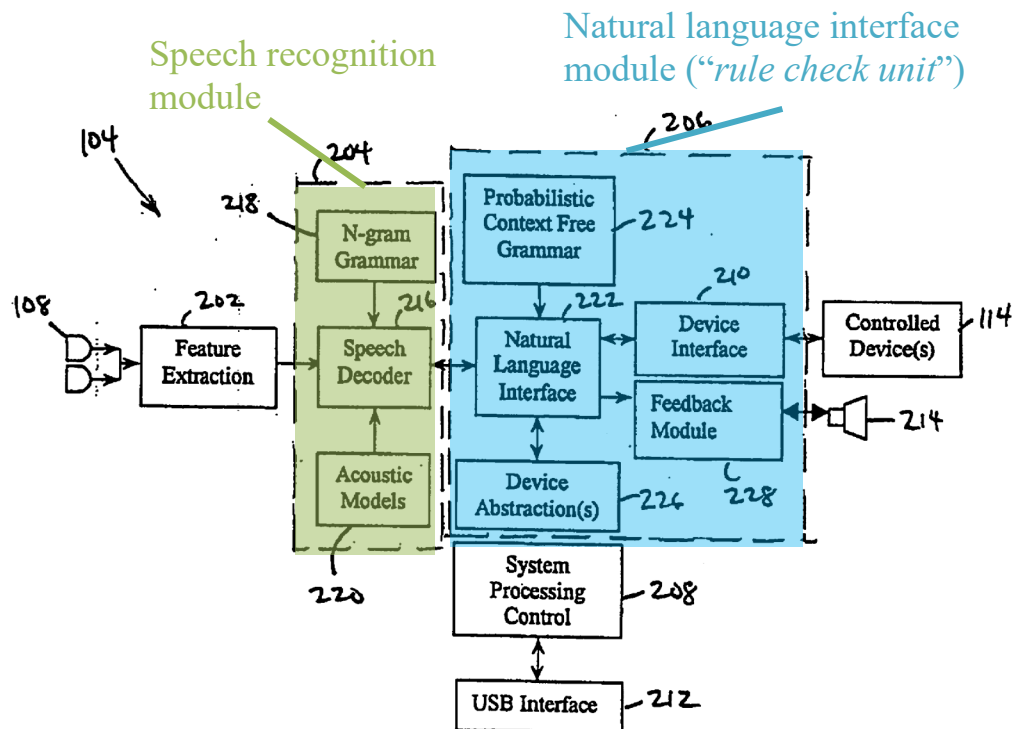
Thus, Bou-Ghazale’s teaching of assuming the input speech to be the activation word based on the result string with the highest likelihood score renders obvious this limitation. Ex.1003, ¶¶160-161.

9. Claim 8

First, as discussed for [1.5], once the activation word is recognized, Bou-Ghazale’s main speech recognition system receives “further commands”

subsequent the activation word and determines whether they include a word from the main speech recognition system's menu. Ex.1005, Sec.1, 1735, Sec.5, 1738. A POSITA would have understood that the main speech recognition system determines whether speech recognized after the activation word corresponds to a sequence of valid words from the "menu" ("*a target sentence*"). Ex.1003, ¶163.

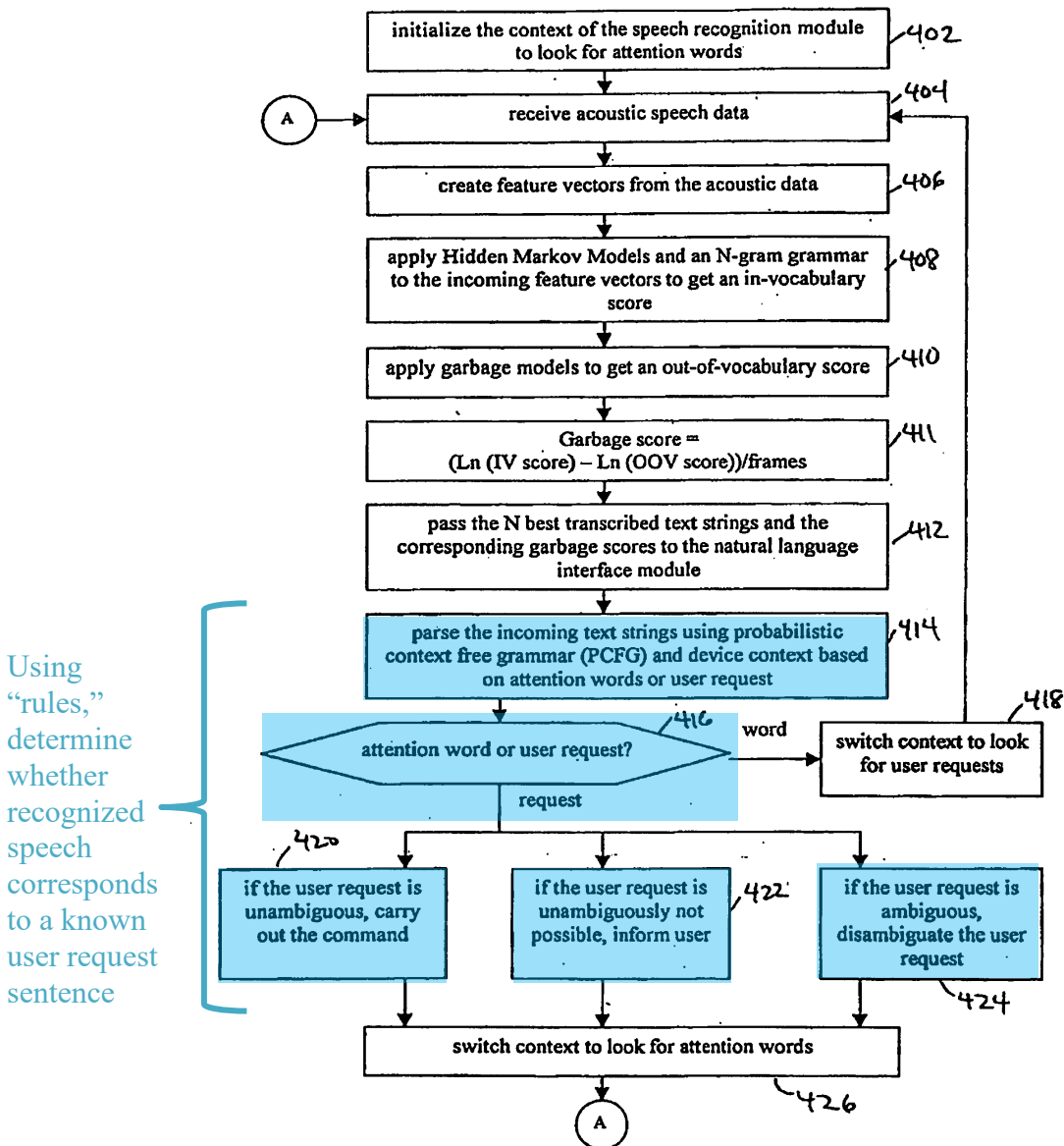
Second, to the extent Bou-Ghazale doesn't explicitly teach a "*target sentence*" and using a "*rule check unit*," Konopka teaches that it was well-known to determine whether input speech corresponds to a known command sentence using a separate "rules"-based module. Ex.1006, [0059], [0020], [0022], Figs. 2, 4; Ex.1003, ¶164. Konopka describes a "natural language interface control system" that includes a "speech recognition module" and a "**natural language interface control module**," shown in Fig. 2 below. Ex.1006, [0020], [0002]; Ex.1003, ¶164.



Ex.1006, Fig. 2 (annotated); Ex.1003, ¶164.

As shown in Fig. 4 below, after an “attention word” is recognized (step 416), the speech recognition module recognizes speech as “user requests” (steps 418, 408-412). Ex.1006, [0081], [0040]; Ex.1003, ¶165. Then, in steps 414-416, and 420-424, the natural language interface module determines whether the recognized speech from the speech recognition module is a known (“unambiguous”) user request sentence using “rule[s]”: **“Using the PCFG rules on the incoming text strings, the natural language interface module 222 reaches one of three conclusions: (1) that it unambiguously understands and can comply with the user request, in which case it carries out the command...”** Ex.1006, [0059], [0020], [0083]. For example, “if the user states ‘hey, lets watch TV’, ‘I wanna

watch TV’, ‘turn on the TV’, ‘whattya say we watch a little television’, the system will use its natural language interface module to probabilistically **determine that the user is requesting to watch the television**” (“*target sentence*”). Ex.1006, [0022].



Ex.1006, Fig. 4 (annotated); Ex.1003, ¶166.

Accordingly, in view of the well-known speech recognition techniques described by Konopka, a POSITA would have found it obvious that Bou-Ghazale's main speech recognition system ("*continuous speech recognition unit*") would have a separate rules-based module for determining whether speech recognized after the activation word corresponds to a known command sentence. Ex.1003, ¶¶162-167.

10. Claim 9

[9.0]

Bou-Ghazale teaches a voice activation system that "can be implemented in a **personal communication device**" ("*apparatus*") for recognizing a pre-defined activation command ("*speech*") spoken by a user ("*speaker*"). Ex.1005, Sec.5, 1738. Bou-Ghazale's voice activation system is "designed to continuously listen to incoming unconstrained **speech** and to **recognize a pre-defined activation command** while rejecting any other extraneous speech or noise." Ex.1005, Abstract, 1735. "Upon **recognizing that the activation command has been spoken**, the voice activation system turns the **device** on or switches to the main speech recognition system to receive further commands from the **user**." Ex.1005, Sec.1, 1735, Abstract.¹²

Thus, Bou-Ghazale's teaching of a personal communication device on which

¹² See *supra* note 4.

a voice activation system can be implemented for recognizing a pre-defined activation command spoken by a user renders obvious this limitation. Ex.1003, ¶¶168-170.

[9.1]

Bou-Ghazale teaches that its voice activation system requires a low-speed processor (“*processor*”). Bou-Ghazale explains that its voice activation system “can be implemented in a personal communication device **due to its low memory and MIPS requirements (<1MIPS).**” Ex.1005, Sec.5, 1738. It was well-known at the time of the ’196 patent that “MIPS” is an acronym for “Million Instructions Per Second” and it refers to a measurement of a processor’s speed. Ex.1003, ¶171 (citing Ex.1012-14). Thus, a POSITA would have understood that Bou-Ghazale’s voice activation system utilizes a low-speed (e.g., <1 MIPS) processor. Ex.1005, Sec.5, 1738; Ex.1003, ¶171.

[9.2]

As discussed for [1.1], Bou-Ghazale’s “voice activation system” renders obvious “*a preset call command recognition unit.*” Ex.1003, ¶172. The voice activation system is “*using the processor*” because the system “continuously **processes** incoming speech.” Ex.1005, Sec.1, 1735; Ex.1003, ¶172.

[9.3]

As discussed for [1.5], Bou-Ghazale’s “main speech recognition system”

renders obvious “*a continuous speech recognition unit.*” Ex.1003, ¶173. The main speech recognition system is also “*using the processor*” of the personal communication device to recognize further commands. As explained for [9.1]-[9.2], the voice activation system uses the device’s processor to process incoming speech. Ex.1005, Sec.1, 1735, Sec.5, 1738; Ex.1003, ¶173. Bou-Ghazale explains that “once the activation word is recognized” in incoming speech, “**the device turns on or switches to the main speech recognition system and waits to be prompted by a word from the main menu** of the speech recognition system.” Ex.1005, Sec.5, 1738. A POSITA would have understood that when the device “switches” between the two components, the main speech recognition system uses the same processor to “receive further commands” and recognize them as a word in its menu. Ex.1005, Sec.5, 1738, Sec.1, 1735; Ex.1003, ¶¶173-174.

[9.4]

See [1.1]. Ex.1003, ¶175.

[9.5]-[9.9]

See [1.2]-[1.6]. Ex.1003, ¶¶176-180.

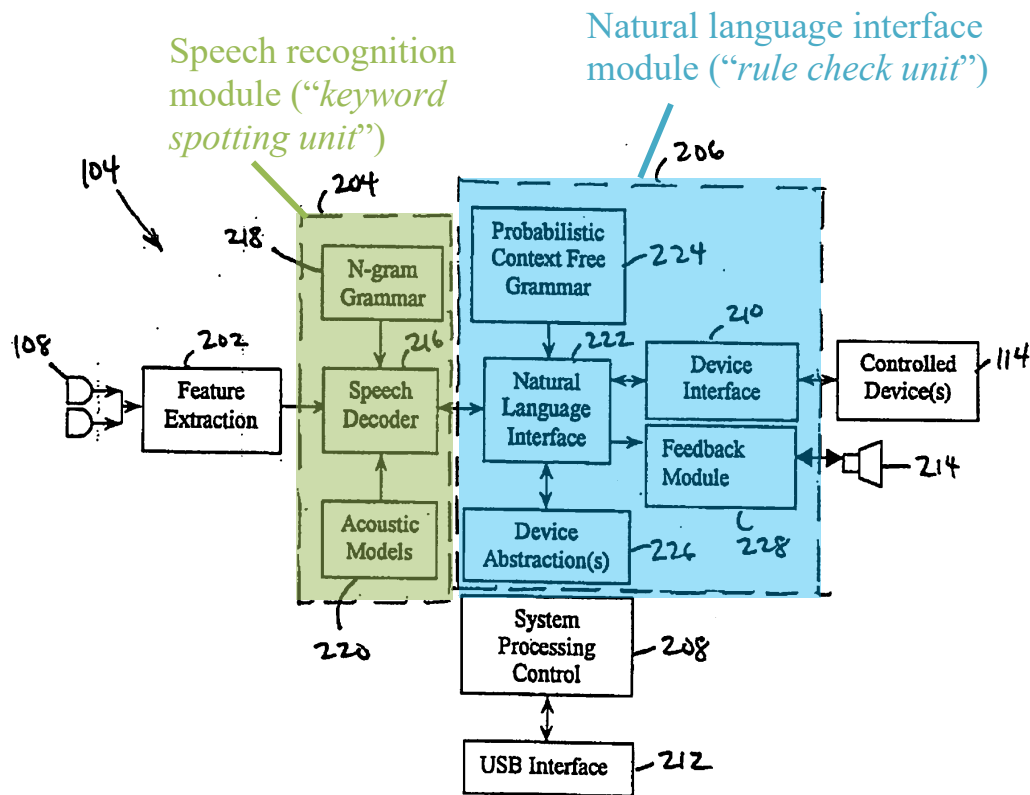
11. Claim 16

[16.0]

First, as discussed for [1.5] and [9.8], Bou-Ghazale’s main speech recognition system is configured to determine whether speech received after the

activation word is recognized as a word from the main menu of the speech recognition system (“*configured to recognize the subsequent input speech as the actual command*”) once the activation word is declared as found in the input speech (“*if the input speech is recognized as the preset call command*”). Ex.1005, Sec.1, 1735, Sec.5, 1738. Bou-Ghazale’s main speech recognition system spots words from its “menu.” Ex.1003, ¶182.

Second, to the extent Bou-Ghazale doesn’t explicitly teach that it uses a separate “*keyword spotting unit*,” Konopka teaches that it was well-known to use a separate module, such as a “*keyword spotting unit*,” to recognize user commands after an activation command has been recognized. Ex.1003, ¶183. Konopka describes a “natural language interface control system” that includes a “**speech recognition module**” and a “natural language interface control module”, shown in Fig. 2 below. Ex.1006, [0020], [0002]; Ex.1003, ¶183.



Ex.1006, Fig. 2 (annotated); Ex.1003, ¶183.

After an “attention word is recognized,” the speech recognition module recognizes “user requests” by “finding the **most likely match of the corresponding utterance with a set of speech models**, while rejecting non-speech events, using an approach based on Hidden Markov Models (HMMs),” as shown by steps 416, 418, and 408-412 of Fig. 4 below. Ex.1006, [0040], [0020] (“**a speech recognition module** utilizing hidden Markov models (HMMs), as known and understood in the art, **to detect the speech.**”), [0080]-[0082]; Ex.1003, ¶184. To recognize user requests, the speech recognition module spots “words” from a set of “lexicons.” Ex.1006, [0028], [0045]; Ex.1003, ¶184.

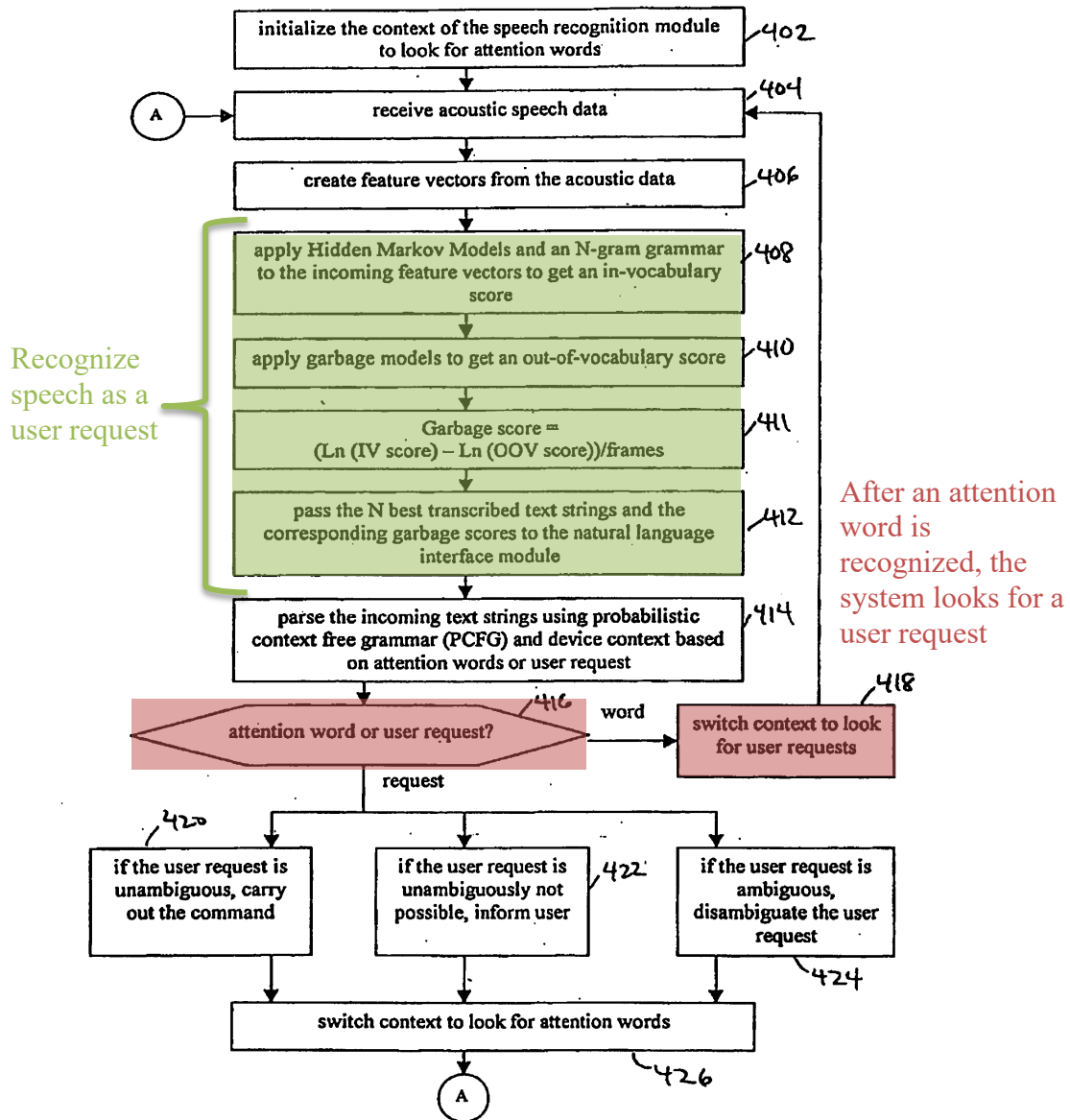


FIG. 4

Ex.1006, Fig. 4 (annotated); Ex.1003, ¶184.

Accordingly, in view of the well-known speech recognition techniques described by Konopka, a POSITA would have found it obvious that Bou-Ghazale's main speech recognition system ("*continuous speech recognition unit*") would

have a separate module that looks for particular words in speech and recognizes that speech as user commands after an activation command is recognized. Ex.1003, ¶¶181-186.

[16.1]

See claim 8. Ex.1003, ¶187.

12. Claim 17

[17.0]-[17.1]

[17.0]-[17.1] are substantially similar to [1.0]-[1.1], except with reference to a “*single device*” instead of a “*device*.” As explained in [1.0], Bou-Ghazale’s voice activation system is “implemented in a **personal communication device**” (“*single device*”). Ex.1005, Sec.5, 1738, Sec.1, 1735, Abstract; Ex.1003, ¶¶188-192.

[17.2]-[17.3]

See [1.2]-[1.3]. Ex.1003, ¶¶193-194.

[17.4]

[17.4] is substantially similar to [1.4], except it specifies that the “*preset call command recognition unit*” is “*of the single device*.” As explained in [17.0], Bou-Ghazale discloses a “*single device*”. Ex.1003, ¶195.

[17.5]

[17.5] is substantially similar to [1.5] except for trivial wording differences. Ex.1003, ¶196.

First, [17.5] recites “***only if the input speech is recognized as the preset call command***” then the “*activating*” step is performed, instead of “***based on determining that the input speech is recognized as the preset call command***,” as recited in [1.5]. As discussed for [1.5], Bou-Ghazale explains that “the **activation word is declared to be found only after** [the input speech] meets a set of statistics”. Ex.1005, Sec.2.3, 1736. Bou-Ghazale further explains that “**once the activation word is recognized, the device turns on or switches to the main recognition system.**” Ex.1005, Sec.5, 1738, Sec.1, 1735, Figure 2. Thus, Bou-Ghazale’s teaching that the main speech recognition system is activated “only after” the input speech meets the statistical criteria and is recognized as the activation word renders obvious “*only if the input speech is recognized as the preset call command, activating a continuous speech recognition unit.*” Ex.1003, ¶197.

Second, [17.5] specifies that the “*continuous speech recognition unit*” is “*of the single device.*” As explained in [1.5], Bou-Ghazale’s main speech recognition system renders obvious “*a continuous speech recognition unit.*” Ex.1003, ¶198. Moreover, Bou-Ghazale teaches that its main speech recognition system is implemented on the same “personal communication device” (“*single device*”) as the voice activation system discussed in [17.0]. Ex.1005, Sec.5, 1738, Sec.1, 1735 (“the voice activation system turns **the device** on or **switches to the main speech**

recognition system”). Thus, Bou-Ghazale renders obvious “*a continuous speech recognition unit of the single device.*” Ex.1003, ¶198.

Third, [17.5] specifies that the “*continuous speech recognition unit*” is “*configured to perform continuous speech recognition.*” For the same reasons as [1.5], Bou-Ghazale renders obvious that the main speech recognition system performs continuous speech recognition “*to determine whether the subsequent input speech is recognized as an actual command.*” Ex.1003, ¶199.

Fourth, regarding “*in order to reduce overall computational load on the single device,*” where [17.5] recites “*single device*” instead of “*device,*” Bou-Ghazale renders this obvious for the same reasons as [1.5] and [17.0]. Ex.1003, ¶¶200-201.

[17.6]

See [1.6]. Ex.1003, ¶202.

13. Claim 19

[19.0]

[19.0] refers to method claim 17 and recites “*wherein the single device*” without limiting this feature further. [19.0] is obvious for the same reasons as claim 17. Ex.1003, ¶203.

[19.1]-[19.2]

See [2.0]-[2.1]. Ex.1003, ¶¶204-205.

14. Claim 20

[20.0]

[20.0] is identical to [5.0], except it includes the word “*whether*” in reciting the step from [17.4]. [20.0] is obvious for the same reasons as [5.0]. Ex.1003, ¶206.

[20.1]

See [5.1]. Ex.1003, ¶207.

D. Ground 2: Claims 6-7 are obvious under §103 over Bou-Ghazale in view of Konopka and Lee.

1. Lee

Lee, like Bou-Ghazale and Konopka relates to HMM-based continuous speech recognition. Ex.1007, 29. Lee teaches a technique for recognizing and verifying a voice code while “reject[ing] non-speech noises, out-of-vocabulary words, and potential recognition mistakes” for systems of “limited memory size.” Ex.1007, 29. Lee’s technique uses a phoneme model that represents characteristics of the voice code and an anti-phoneme model that is “based on phoneme and syllable units against the claimed voice code using the statistical distance of phonemes.” Ex.1007, 30. Because Lee’s anti-phoneme model is constructed by reusing a general acoustic model for the phoneme units that are “statistically far from the phoneme group,” an extra trained garbage model is not required and therefore Lee’s technique is advantageous for systems with limited memory size.

Ex.1007, 33, 29; Ex.1003, ¶208.

2. Reasons to Combine Bou-Ghazale and Konopka with Lee

For the below reasons, a POSITA would have been motivated to combine the teachings of Lee with the teachings of Bou-Ghazale and Konopka. Ex.1003, ¶¶209-213. It would have been obvious, beneficial, and predictable to apply Lee's teachings of using phoneme and anti-phoneme models for recognizing and verifying commands to Bou-Ghazale's voice activation system (as modified by Konopka)—for example, to further Bou-Ghazale's and Lee's goals of accurately recognizing commands while rejecting noise or extraneous speech, and to further reduce the memory requirements of the system. Ex.1003, ¶209.

As an initial matter, a POSITA, when considering the teachings of Bou-Ghazale and Konopka would have also considered the teachings of Lee. Ex.1003, ¶210. Lee relates to HMM-based continuous speech recognition and is therefore analogous art to the '196 patent. Ex.1007, 29; Ex.1001, Abstract, Fig. 6; Ex.1003, ¶210.

A POSITA considering Bou-Ghazale would understand that among Bou-Ghazale's goals is accurately recognizing commands while rejecting noise or extraneous speech and reducing the system's memory requirements. Ex.1005, Abstract (explaining that its system is designed “to recognize a pre-defined activation command while **rejecting any other extraneous speech or noise.**”),

Sec.1, 1735, Sec.5, 1738 (explaining that its system has **low memory and MIPS requirements**”). At the time of the ’196 patent, a well-known way to “reject non-speech noises, out-of-vocabulary words, and potential recognition mistakes” while reducing memory requirements was to use phoneme and anti-phoneme models, where the phoneme model represents characteristics of a voice code (i.e., activation word) and the anti-phoneme model is “based on phoneme and syllable units against the claimed voice code using the statistical distance of phonemes,” as taught by Lee. Ex.1007, 29-30. Lee’s technique does not require an extra trained garbage model and therefore is advantageous for systems with limited memory requirements. Ex.1007, 29 (“Most algorithms require the extra model trained for a garbage model or anti-model. However, the limited memory size of our proposed embedded system prevents the algorithm from using and storing the extra alternative hypothesis model.”). Because Bou-Ghazale’s system is “implemented in a personal communication device due to its low memory...requirements” a POSITA would have been motivated to use Lee’s phoneme models and anti-phoneme models as its keyword model and out-of-vocabulary models, respectively, to recognize commands while rejecting noise or other extraneous speech. Ex.1005, Sec.5, 1738; Ex.1003, ¶211. Doing so would eliminate the need to train and store an extra garbage model for out-of-vocabulary words in Bou-Ghazale and Konopka’s system, thereby saving memory space. Ex.1003, ¶211; *see*

also KSR, 550 U.S. at 417.

A POSITA would have had a reasonable expectation of success applying Lee’s technique to Bou-Ghazale and Konopka. Ex.1003, ¶212. For example, Lee’s technique was specifically intended to be implemented in an HMM-based speech recognition system, such as Bou-Ghazale’s voice activation system. *See* Ex.1007, 29. Moreover, the combination would not change the intended functionality of Bou-Ghazale’s voice activation system, it simply would improve it: Bou-Ghazale’s system and method would still operate to reject noise or other extraneous speech, but do so without having to train and store an extra garbage model, as taught by Lee. Ex.1003, ¶212.

Accordingly, a POSITA would have found it obvious to apply Lee’s teachings to the system and methods taught by Bou-Ghazale and Konopka because the combination merely amounts to applying a known technique to a system and methods ready for improvement. Ex.1003, ¶¶209-213.

3. Claim 6

[6.0]

As explained for [5.0], Bou-Ghazale teaches that its speech recognition component determines “the result string with the **highest likelihood score**” and teaches that a likelihood score may be a “**log-likelihood score for the keyword model**,” while Konopka teaches computing a natural logarithm of an “in-

vocabulary (IV) Viterbi (likelihood) score” to determine the best text string (e.g., Bou-Ghazale’s result string with the highest likelihood score). Ex.1005, Sec.2.2-3.1, 1736; Ex.1006, [0080], Fig. 4. Konopka explains that the Viterbi score represents “a single best state sequence through an **HMM** and its corresponding **probability**.” Ex.1006, [0043]. Accordingly, a POSITA would understand that Konopka’s in-vocabulary likelihood score is a “*probability*” on the input speech for a hidden Markov model (HMM) for an in-vocabulary word. Ex.1003, ¶217 (citing Ex.1018-1019). Therefore, based on the teachings of Konopka, Bou-Ghazale’s “likelihood score,” as computed by the speech recognition component is a “probability” on the input speech for its keyword model, of which a natural logarithm can be computed (“*natural logarithm of a probability on the input speech for a phoneme model of the preset call command*”). Ex.1003, ¶¶214-217.

A POSITA would have recognized that states in HMMs for speech recognition typically model phonemes, but to the extent Bou-Ghazale and Konopka do not explicitly recite that their “keyword model” and “in-vocabulary model” are a “*phoneme model of the preset call command*,” Lee renders it obvious. Ex.1003, ¶218. Lee teaches a “voice code verification” system and process, that uses “**a phoneme model**...modeled using HMM having states” and an “anti-phoneme model.” Ex.1007, 29, 33, 38. In particular, Lee describes a two-step process, where in the first step “recognition is performed via a conventional Viterbi

beam search algorithm that segments the test utterance into the ***N*-best strings of a phoneme hypotheses.**” Ex.1007, 30. As was well-known in the art, the Viterbi algorithm uses models (e.g., acoustic models) to analyze the input speech and determine “scores” for the strings in the models—acoustic models “**represent general characteristics of speech**” and “can be modeled as a...**phoneme unit.**” Ex.1007, 29; *see* Ex.1006, [0041] (“**As is well understood in the art, the received utterances are analyzed with the models, e.g., using a Viterbi algorithm, and scores are assigned representing how well the utterance fits the given models.**”). Accordingly, a POSITA would have understood that Lee’s “phoneme hypotheses” refers to “phoneme models” representing characteristics of the voice code. Ex.1003, ¶220. Thus, Lee’s use of a phoneme model representing a voice code word to determine the “best strings” with the highest likelihood score renders obvious “*a probability on the input speech for a phoneme model of the preset call command.*” Ex.1003, ¶¶218-220.

Thus, Bou-Ghazale’s teaching of a log-likelihood score for a keyword model in view of Konopka’s teachings of computing a natural logarithm of a probability on the input speech for an in-vocabulary word model, and Lee’s teachings of a phoneme model renders obvious this limitation. Ex.1003, ¶¶214-221.

[6.1]

As explained for [2.0], Bou-Ghazale teaches that the post-processing

component determines a “relative log-likelihood score” for the input speech (“*computed confidence rate*”). Ex.1005, Sec.2.3-3.1, 1736. While Bou-Ghazale’s “likelihood score” determined by the speech recognition component may be a log-likelihood score for the keyword model as explained for [5.0] and [6.0], the “relative log-likelihood score” determined by the post-processing component is “*another log likelihood ratio*” because it uses an “accumulated log-likelihood score for the keyword model” **and** a “log-likelihood score for the [out-of-vocabulary (OOV)] model.” Ex.1005, Sec.3.1, 1736 (“The **relative log-likelihood score** should be equal to or greater than a threshold, η , given by: $\eta = (\text{Score}_{\text{keyword}} - \text{Score}_{\text{OOV}})/|\text{Score}_{\text{keyword}}|$, where the **accumulated log-likelihood score for the keyword model** is denoted by $\text{Score}_{\text{keyword}}$, and the **log-likelihood score for the OOV model** is denoted by $\text{Score}_{\text{OOV}}$.”); *see also* Ex.1005, Sec.3.1-3.2, Figure 2, 1736-37 (using “relative log-likelihood score” and “relative log likelihood **ratio**” interchangeably).

Thus, Bou-Ghazale’s teaching of a relative log-likelihood score that is different from the likelihood score determined by the speech recognition component renders obvious this limitation. Ex.1003, ¶¶222-223.

4. Claim 7

Bou-Ghazale teaches that its “relative log-likelihood score” (“*another log likelihood ratio*”) is computed by subtracting a “log-likelihood score for the OOV

model” from an “accumulated log-likelihood score for the keyword model.”

Ex.1005, Sec.3.1, 1736 (“The **relative log-likelihood score should be equal to or greater than a threshold, η** , given by: $\eta = (\text{Score}_{\text{keyword}} - \text{Score}_{\text{OOV}}) / |\text{Score}_{\text{keyword}}|$, where the **accumulated log-likelihood score for the keyword model** is denoted by $\text{Score}_{\text{keyword}}$, and the **log-likelihood score for the OOV model** is denoted by $\text{Score}_{\text{OOV}}$.”).

First, as explained for [5.0] and [6.0], Bou-Ghazale’s “log-likelihood score for the keyword model” in the above equation corresponds to “*the log likelihood ratio*” that is “*a natural logarithm of a probability on the input speech for a phoneme model of the preset call command.*” Ex.1003, ¶226.

Second, Bou-Ghazale’s “log-likelihood score for the [out-of-vocabulary] OOV model” in the above equation corresponds to “*a natural logarithm of a probability on the input speech for an anti-phoneme model of the preset call command.*” Ex.1003, ¶¶227-228. Konopka teaches a similar equation to that of Bou-Ghazale for determining a log-likelihood “score” using a log-likelihood score for an in-vocabulary model ($\text{Ln}(\text{IV score})$) and a log-likelihood score for an out-of-vocabulary model ($\text{Ln}(\text{OOV score})$): “ **$\text{Ln}(\text{IV score}) - \text{Ln}(\text{OOV score})$ /number of frames.**” Ex.1006, [0080], Figure 1. For similar reasons as discussed for an in-vocabulary likelihood score in [6.0], a POSITA would understand that Konopka’s

out-of-vocabulary likelihood score is a “**probability**” on the input speech for a hidden Markov model (HMM) for out-of-vocabulary words. Ex.1006, [0043] (explaining that the out-of-vocabulary score is “**another Viterbi score** that is obtained by **reprocessing the utterance through an alternative HMM** whose task is to represent all out-of-vocabulary events or filler sounds”). Therefore, Bou-Ghazale’s “log-likelihood score for the OOV model,” is a natural logarithm probability on the input speech for out-of-vocabulary model. Ex.1003, ¶¶227-228.

Third, to the extent Bou-Ghazale and Konopka do not explicitly recite that an “out-of-vocabulary model” is an “*anti-phoneme model of the preset call command*,” Lee renders it obvious because Lee teaches using an “**anti-phoneme model**” in a second step to “compute[] a **confidence measure** that determines whether or not to reject the recognized voice code” from the first step. Ex.1007, 38, 30. Lee explains that its anti-phoneme models are specific to the voice code (“*of the preset call command*”) because they are “based on **phoneme and syllable units against the claimed voice code** using the statistical distance of phonemes.” Ex.1007, 30, 33; Ex.1003, ¶229. A POSITA would have been motivated to use Lee’s anti-phoneme model as Bou-Ghazale’s out-of-vocabulary model because Lee’s technique may be applied to “**reject non-speech noises, out-of-vocabulary words**, and potential recognition mistakes” and save memory space by eliminating the need to train and store a garbage model for out-of-vocabulary words. Ex.1007,

29; §IX.D.2.; Ex.1003, ¶230.

Thus, Bou-Ghazale’s teaching of computing a relative log-likelihood score by subtracting a log-likelihood score for the OOV model from a log-likelihood score for the keyword model, in view of Konopka’s teachings of a similar equation that computes a natural logarithm of a probability on the input speech for an out-of-vocabulary model, and Lee’s teachings of an anti-phoneme model for determining a confidence measurement to reject out-of-vocabulary words renders obvious this limitation. Ex.1003, ¶¶224-231.

E. Ground 3: Claims 10-13, and 18 are obvious under §103 over Bou-Ghazale in view of Konopka and Takiguchi.

1. Takiguchi

Takiguchi relates to “a speech enhancement technique which is effective for an extemporaneous noise without a noise interval and also for unknown extemporaneous noises” and which “can be applied to, for example, a speech recognition device and method.” Ex.1008, [0026], [0063]. More specifically, Takiguchi’s technique utilizes an “adaptive filter” with a corresponding filter coefficient that is applied to incoming speech to “reduce the noise signal.” Ex.1008, [0027]. In Takiguchi’s technique, “the filter coefficient is updated for each utterance [or for each frame or plurality of frames], and the target speech signal is enhanced using the updated filter coefficient.” Ex.1008, [0101].

Accordingly, Takiguchi's adaptive filter adapts to and effectively removes different types of noises, thereby improving speech recognition accuracy in noisy environments. Ex.1008, [0026], [0022]; Ex.1003, ¶232.

2. Reasons to Combine Bou-Ghazale and Konopka with Takiguchi

For the below reasons, a POSITA would have been motivated to combine the teachings of Takiguchi with the teachings of Bou-Ghazale and Konopka. Ex.1003, ¶¶233-240. It would have been obvious, beneficial, and predictable to utilize Takiguchi's teachings of an adaptive filtering technique in combination with Bou-Ghazale's voice activation system—for example, to further Bou-Ghazale's and Takiguchi's goals of effectively reducing extemporaneous noises (or extraneous speech), thereby improving speech recognition. Ex.1003, ¶233.

As an initial matter, a POSITA, when considering the teachings of Bou-Ghazale and Konopka would have also considered the teachings of Takiguchi. Ex.1003, ¶234. Takiguchi relates to HMM-based speech recognition and is therefore analogous art to the '196 patent. Ex.1008, [0026]-[0027], [0057], [0059]; Ex.1001, Abstract, Fig. 6; Ex.1003, ¶234.

A POSITA considering Bou-Ghazale would understand that one of Bou-Ghazale's primary concerns is accurately recognizing commands while rejecting noise or other extraneous speech. Ex.1005, Abstract (explaining that its system is

designed “to recognize a pre-defined activation command while **rejecting any other extraneous speech or noise.**”), Sec.1, 1735. Bou-Ghazale teaches a voice activation system that includes a “band-pass filter” for filtering speech within a set range, thereby rejecting noise outside of that range. Ex.1005, Sec.2.2, 1736.

However, Takiguchi explains that it is difficult to estimate a set filtering coefficient for effectively removing different types of extemporaneous noises, and therefore speech recognition accuracy deteriorates. Ex.1008, [0022] (“it is difficult to estimate the filter coefficient $w(\omega)$ for removing extemporaneous noise which is completely superimposed on the target speech signal, which exists only in the speech occurrence interval, and which continues for a short time. Accordingly, in...speech recognition in a car, or the like, extemporaneous noise...is one cause of deteriorating recognition accuracy.”), [0023]. Takiguchi provides a well-known and predictable solution to this problem using an “adaptive filter” where its filter coefficient is “updated” for each portion of speech “to reduce the noise signal.” Ex.1008, [0027], [0101]. As a result, an adaptive filtering (“speech enhancement”) technique “which is effective for an extemporaneous noise without a noise interval and also for unknown extemporaneous noises” is provided. Ex.1008, [0026]; Ex.1003, ¶¶235-237.

A POSITA would have been motivated to implement Takiguchi’s adaptive filtering technique into Bou-Ghazale’s voice activation system to more effectively

reduce extemporaneous noises and improve speech recognition. Ex.1003, ¶238. Specifically, the benefits of Takiguchi's technique would have motivated a POSITA to modify Bou-Ghazale's band-pass filter to be adaptive, so that it updates its filter coefficient for each portion of speech. Ex.1003, ¶238. Doing so would reduce extemporaneous noise (or extraneous speech) in the filtered speech and improve recognition accuracy of the activation word. Ex.1003, ¶238; *see also KSR*, 550 U.S. at 417.

A POSITA would have had a reasonable expectation of success applying Takiguchi's technique to Bou-Ghazale and Konopka. Ex.1003, ¶239. For example, Takiguchi's technique was specifically intended to be implemented in an HMM-based speech recognition system, such as Bou-Ghazale's voice activation system. Ex.1008, [0057], [0059]. Moreover, the combination would not change the intended functionality of Bou-Ghazale's voice activation system, it would simply improve its current range for rejecting extemporaneous noises, as taught by Takiguchi. Ex.1003, ¶239.

Accordingly, a POSITA would have found it obvious to apply Takiguchi's teachings to those of Bou-Ghazale and Konopka because the combination merely amounts to applying a known technique to a system and methods ready for improvement. Ex.1003, ¶¶233-240.

3. Claim 10

[10.0]

Bou-Ghazale teaches that, in its voice activation system (“*preset call command recognition unit*”), “[t]he **input speech** is sampled at 8kHz, then **filtered by a band-pass filter** with a pass-band from 300 to 3300 Hz.” Ex.1005, Sec.2.2, 1736. The band-pass filter “*receive[s] the input speech and the subsequent input speech*” because Bou-Ghazale’s voice activation system “**continuously**” listens to and processes incoming speech, but before the speech is processed it is “filtered.” Ex.1005, Sec.1, 1735, Sec.2, 1736 (“[t]he **filtered speech** is processed”); Ex.1003, ¶¶241-242.

While Bou-Ghazale’s band-pass filter uses a fixed frequency range, Takiguchi teaches that an “adaptive filter” can be used to adjust the frequency range to more effectively reduce extemporaneous noises and improve speech recognition. Ex.1008, [0026]-[0027] (“it is an aspect of the present invention to provide a speech enhancement technique which is effective for an extemporaneous noise without a noise interval and also for unknown extemporaneous noises”). Takiguchi describes a “signal enhancement device designed to enhance a target signal...by controlling a filter coefficient of **an adaptive filter to be applied to the reference signal** to reduce the noise signal.” Ex.1008, [0027]. Takiguchi explains that the “reference signal” to which the “adaptive filter” is applied to is

obtained from “input signals” including speech utterances. Ex.1008, [0086]
(“signal synthesis unit 31 also **adds the input signal m2 to the input signal m1**
after inverting the input signal m2... and **outputs** the resultant signal as the
reference signal...The adaptive filter unit 33 **adjusts the amplitude of the**
reference signal $r(\omega, n)$ in accordance with the filter coefficient $w(\omega)$.”), [0094]
(“input signals $m1(t)$ and $m2(t)$ were created using **utterance** data”). Accordingly,
Takiguchi’s “adaptive filter” receives “*input speech*” in the form of reference
signals. Ex.1003, ¶243.

Thus, Bou-Ghazale’s teaching of a band-pass filter that filters all incoming
speech, in light of Takiguchi’s teaching of an adaptive filter for more effectively
filtering noise, renders obvious this limitation. Ex.1003, ¶¶241-244.

[10.1]

[10.1] is substantially similar to [2.0]-[2.1], except it recites “*assumed input speech*” instead of “*input speech*.” As explained for [1.6] and [9.9], Bou-Ghazale teaches that the speech recognition component “assumes” the input speech to be the activation word according to “the **result string** with the highest likelihood score” that is passed to the post-processing component (Ex.1005, Sec.2.2, 1736) and further teaches that the post-processing component verifies the assumption using criteria and confidence measurements applied to the result string (Ex.1005, Sec.2.3, 1736). Accordingly, the input speech (represented by the result string) that

is passed from the speech recognition component to the post-processing component renders obvious the “*assumed input speech*.” Ex.1003, ¶245.

4. Claims 11-12

Claims 11-12 are substantially similar to claims 3-4 except they recite the “*assumed input speech*.” As explained in [10.1], Bou-Ghazale discloses the “*assumed input speech*.” Ex.1003, ¶¶246-247.

5. Claim 13

See [1.6], [5.0]-[5.1]. Ex.1003, ¶248.

6. Claim 18

First, as explained for [17.0], Bou-Ghazale’s “personal communication device” renders obvious the “*single device*.” Bou-Ghazale teaches that a personal communication device is, for example, a “cellular phone” which is a “*portable device*” as claimed. Ex.1005, Sec.1, 1735; Ex.1003, ¶¶249-250.

Second, regarding “*an embedded system*,” because Bou-Ghazale’s system is “**implemented** in a personal communication device” which is a larger device or system, a POSITA would have understood that it is an “*embedded system*.” Ex.1005, Sec.5, 1738; Ex.1003, ¶251. Further, Takiguchi teaches that it was well-known to provide speech recognition components “**embedded** in a computer program product,” for loading into the computer of a “car,” or other device. Ex.1008, [0103], [0091]-[0092]; Ex.1003, ¶251.

Thus, Bou-Ghazale in view of Takiguchi render obvious this limitation.

Ex.1003, ¶¶249-252.

F. Ground 4: Claims 14-15 are obvious under §103 over Bou-Ghazale in view of Konopka, Takiguchi, and Lee.

1. Claims 14-15

See claims 6-7. Ex.1003, ¶¶253-254.

X. DISCRETIONARY DENIAL WOULD BE INAPPROPRIATE

A. No basis for *Fintiv* denial

The *Fintiv* factors favor institution. *Apple Inc. v. Fintiv, Inc.*, IPR2020-00019, Paper 11 (Mar. 20, 2020) (precedential).

Factor 1: No motion to stay has been filed, and thus this factor is neutral.

Factor 2: The Board will likely issue a Final Written Decision in July/August of 2026, and the district court trial date is currently set for March 16, 2026. Ex.1028. The estimated trial date based on time-to-trial statistics is June 2026. Ex.1033, Page 5 of 11 (23 months). This factor is neutral because the trial date is speculative and subject to change. *See* Ex.1032, 8 (Interim Procedure For Discretionary Denials) (“scheduled trial dates are unreliable and often change”). For example, Judge Gilstrap has *eleven* trials set for March 16, 2026. Ex.1031. Further, a Motion to Dismiss for Lack of Venue is currently pending before the Court.

Factor 3: The co-pending litigation is in its early stages, and the investment in it has been minimal. The claim construction hearing is not until September 16, 2025, fact discovery is not set to close until October 17, 2025, and expert discovery is not set to close until December 4, 2025. Ex.1028, Pages 3-4 of 7. Moreover, Petitioner only learned which claims were being asserted on November 14, 2024, and has worked expeditiously to file this petition within three months. Ex.1029 (infringement contentions). This factor favors institution.

Factor 4: Final invalidity theories will not be submitted until March 11, 2026, so the extent of overlap is currently speculative. If Patent Owner raises a *Fintiv* challenge in its preliminary response, Petitioner will evaluate whether a stipulation is appropriate based on the state of the district court proceeding at that time. This factor is neutral.

Factor 5: Petitioner is a defendant in the co-pending litigation, and thus this factor is neutral.

Factor 6: This Petition presents compelling evidence of unpatentability, and thus denial is inappropriate. Ex.1032, 3-5. The Petition—supported by Mr. Schmandt’s expert testimony—plainly shows that the ’196 patent claims no more than what was already known or obvious.

In sum, every *Fintiv* factor is either neutral or weighs against denial.

B. No basis for §325(d) denial

None of references Bou-Ghazale, Konopka, Lee, or Takiguchi were previously considered by the Office.

XI. CONCLUSION

Petitioner has established a reasonable likelihood that the Challenged Claims are unpatentable.

Respectfully submitted,

Dated: February 25, 2025
HAYNES AND BOONE, LLP
2801 N. Harwood St., Suite 2300
Dallas, Texas 75201
Customer No. 27683

/Scott T. Jarratt/
Scott T. Jarratt
Lead Counsel for Petitioner
Registration No. 70,297

XII. MANDATORY NOTICES

A. Real Party-in-Interest

Pursuant to 37 C.F.R. §42.8(b)(1), Petitioner certifies that the real parties-in-interest are Amazon.com, Inc, Amazon.com Services LLC, and Amazon Web Services, Inc.

B. Related Matters

Pursuant to 37 C.F.R. §42.8(b)(2), to the best knowledge of the Petitioner, the '196 patent is or was involved in the following case:

Case Heading	Number	Venue	Filed
<i>KAIFI LLC v. Amazon.com, Inc. et al</i>	2:24-cv-00542	EDTX	Jul. 17, 2024

C. Lead and Back-up Counsel and Service Information

Lead Counsel

Scott T. Jarratt
HAYNES AND BOONE, LLP
2801 N. Harwood St., Suite 2300
Dallas, TX 75201

Phone: (972) 739-8663
Fax: (214) 200-0853
scott.jarratt.ipr@haynesboone.com
USPTO Reg. No. 70,297

Back-up Counsel

Brian Kwok
HAYNES AND BOONE, LLP
2801 N. Harwood St., Suite 2300
Dallas, TX 75201

Phone: (415) 293-8916
Fax: (214) 200-0853
brian.kwok.ipr@haynesboone.com
USPTO Reg. No. 58,828

Destiny L. Stephenson
HAYNES AND BOONE, LLP
2801 N. Harwood St., Suite 2300
Dallas, TX 75201

Phone: (214) 651-5792
Fax: (214) 200-0853
lyric.stephenson.ipr@haynesboone.com
USPTO Reg. No. 80,404

Please address all correspondence in this proceeding to lead and back-up counsel. Petitioner consents to service in this proceeding by email at the addresses above.

CERTIFICATE OF WORD COUNT

Pursuant to 37 C.F.R. §42.24(d), Petitioner hereby certifies, in accordance with and reliance on the word count provided by the word-processing system used to prepare this Petition, that the number of words in this paper is 13,952. Pursuant to 37 C.F.R. §42.24(d), this word count excludes the table of contents, table of authorities, mandatory notices under §42.8, certificate of service, certificate of word count, appendix of exhibits, and any claim listing.

Dated: February 25, 2025

/Scott T. Jarratt/
Scott T. Jarratt
Lead Counsel for Petitioner
Registration No. 70,297

CERTIFICATE OF SERVICE

The undersigned certifies that, in accordance with 37 C.F.R. § 42.6(e) and 37 C.F.R. § 42.105, service was made on Patent Owner as detailed below.

Date of service February 25, 2025

Manner of service FEDERAL EXPRESS

Documents served Petition for *Inter Partes* Review Under 35 U.S.C. § 312 and 37 C.F.R. § 42.104 of U.S. 8,930,196; Petitioner's Power of Attorney; Petitioner's Exhibit List; Exhibits Ex.1001 – Ex.1019 and Ex.1021 – Ex.1038.

Persons served Cantor Colburn LLP
20 Church Street, 22nd Floor
Hartford, CT 06103

/Scott T. Jarratt/
Scott T. Jarratt
Lead Counsel for Petitioner
Registration No. 70,297