

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

CELLULAR SOUTH, INC.,
Patent Owner.

Case IPR2025-00875
Patent 9,940,972 B2
Issue Date: April 10, 2018

Title: VIDEO TO DATA

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 9,940,972 B2**

Table of Contents

	Page
I. MANDATORY NOTICES UNDER §42.8(A)(1)	1
A. Real Party-In-Interest under §42.8.(b)(1).....	1
B. Related Matters under §42.8(b)(2)	1
C. Lead and Back-Up Counsel under §42.8(b)(3)	1
D. Service Information	2
II. FEE PAYMENT	2
III. REQUIREMENTS UNDER §§ 42.104 AND 42.108 AND CONSIDERATIONS UNDER §§ 314(A) AND 325(D).....	2
A. Grounds for Standing	2
B. Identification of Challenge and Statement of Precise Relief Requested	3
C. Considerations Under §§ 314(a) and 325(d).....	3
IV. OVERVIEW OF THE PATENT	5
A. Level of Ordinary Skill	5
B. Specification Overview	5
V. CLAIM CONSTRUCTION	9
VI. THE CHALLENGED CLAIMS ARE UNPATENTABLE.....	9
A. Overview of Grounds	9
B. Prior Art Status of Relied-Upon References	10
C. Ground 1: Claims 1-20 Are Obvious Over Fontana in view of Lau	10
1. Independent Claim 1: “A method to generate video data from a video comprising:” (Claim 1[pre]).....	10
(a) “generating audio files and image files from the video;” (Claim 1[a]).....	11

Table of Contents
(continued)

	Page
(b) “distributing the image files across a plurality of processors and processing the image files in parallel, wherein processing the image files comprises extracting one or more objects and identifying the one or more objects;” (Claim 1[b]).....	13
(c) “processing the audio files;” (Claim 1[c]).....	18
(d) “converting audio files associated with the video to text;” (Claim 1[d])	18
(e) “converting the image files associated with the video to video data;” (Claim 1[e]).....	19
(f) “generating a topical meta-data that describes content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio files;” (Claim 1[f])	20
(g) “adding the topical meta-data to the video; and” (Claim 1[g])	23
(h) “cross-referencing the text and the video data based on the generated topical meta-data to determine topics;” (Claim 1[h]).....	26
(i) “generating video text based on the cross-referencing, wherein the video text describes content of the video;” (Claim 1[i])	32
(j) “generating a text, image, or animation based on the video text; and” (Claim 1[j])	35
(k) “placing the text, image, or animation in the video.” (Claim 1[k])	36
2. Claim 2: “The method according to claim 1, further comprising: generating a content-rich video based on the video, the text, and the video data.”	36

Table of Contents
(continued)

Page

3.	Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”	37
4.	Claim 4: “The method according to claim 2, further comprising: applying natural language processing to the text to extract the topical meta-data.”	38
5.	Claim 5: “The method according to claim 1, further comprising: processing the image files to extract additional text.”	39
6.	Claim 6: “The method according to claim 5, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”	40
7.	Claim 7: “The method according to claim 1, further comprising: determining a motion associated with the one or more objects.”	41
8.	Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”	41
9.	Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum thresholds.”	42
10.	Claim 10: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”	45
11.	Claim 11: “The method according to claim 10, further comprising: placing the advertisement in the video at a preferred time.”	45
12.	Claim 12: “The method according to claim 6 wherein the additional text includes information regarding context associated with the video.”	46
13.	Claim 13: “The method according to claim 6, wherein the additional text relates to a symbol appearing in the video.”	46

Table of Contents
(continued)

	Page
14. Claim 14: “The method according to claim 13, wherein the symbol is a brand logo, and wherein the additional text includes information regarding placement and time of appearance of the brand logo.”	46
15. Claim 15: “The method according to claim 1, wherein the one or more objects are letters appearing in the video.”	48
16. Claim 16: “The method according to claim 6, wherein the additional text relates to faces appearing in the video.”	49
17. Independent Claim 17: “A system for extracting data from a video, comprising:” (Claim 17[pre]).....	49
(a) “a computer processor having a plurality of processors for parallel processing; and” (Claim 17[a]).....	50
(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 17[b]).....	50
(c) “converting audio associated with the video to text; ” (Claim 17[c])	51
(d) “converting images associated with the video to video data; ” (Claim 17[d]).....	51
(e) “generating the video data by segmenting image files of the video before processing the image files in parallel; ” (Claim 17[e])	51
(f) “identifying one or more objects in the image files; ” (Claim 17[f]).....	52
(g) “generating data topics, from the text and the video data, that describe content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio; ” (Claim 17[g]).....	52

Table of Contents
(continued)

	Page
(h) “adding the data topics to the video as meta-data; ” (Claim 17[h])	52
(i) “cross-referencing the text, the video data, and the topics with the video based on the generated data topics; ” (Claim 17[i]).....	53
(j) “generating a text, image, or animation based on the data topics; and” (Claim 17[j])	54
(k) “placing the text, image, or animation in the video.” (Claim 17[k])	54
18. Claim 18: “The system according to claim 17, wherein converting the audio comprises natural language processing.”	54
19. Claim 19: “The system according to claim 17, the computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel.”	55
20. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”	57
D. Ground 2: Claims 8-9 and 20 Are Obvious Over Fontana in view of Lau and Arakawa	57
1. Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”	57
2. Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum thresholds.”	63
3. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”	65
VII. CONCLUSION.....	65

Table of Contents
(continued)

Page

CERTIFICATE OF SERVICE67

List of Exhibits

Exhibit No.	Description of Document
1001	U.S. Patent No. 9,940,972 B2 to Naeem Lakhani et al. (filed February 7, 2014, issued April 10, 2018) (“’972” or “’972 patent”)
1002	Declaration of Henry Houh, Ph.D. (“ Houh ”)
1003	U.S. Patent Application Publication No. 2012/0078712 A1 to James A. Fontana et al. (filed September 27, 2010, issued March 29, 2012) (“ Fontana ”)
1004	U.S. Patent Application Publication No. 2007/0112630 A1 to Wai Kit Lau et al. (filed November 7, 2006, issued May 17, 2007) (“ Lau ”)
1005	U.S. Patent Application Publication No. 2012/0239401 A1 to Takayuki Arakawa (filed November 26, 2010, issued September 20, 2012) (“ Arakawa ”)
1006	U.S. Patent No. 9,697,230 B2 to Henry Houh et al. (filed March 31, 2006, issued July 4, 2017)
1007	’972 Patent File History
1008	Adobe Flash Video File Format Specification, Version 10.1, https://rtmp.veriskope.com/pdf/video_file_format_spec_v10_1.pdf
1009	Excerpts from Scot Hacker, MP3 The Definitive Guide (1st ed. 2000)
1010	U.S. Patent Application Publication No. 2011/0305394 A1 to David W. Singer et al. (filed June 15, 2010, issued December 15, 2011) (“ Singer ”)
1011	U.S. Patent No. 9,407,942 B2 to Andrew Brenneman (filed November 7, 2012, issued August 2, 2016) (“ Brenneman ”)
1012	U.S. Patent Application Publication No. 2004/0125877 A1 to Shin-Fu Chang et al. (filed April 9, 2001, published July 1, 2004) (“ Chang ”)
1013	Proof of Service of Complaint

I. MANDATORY NOTICES UNDER §42.8(A)(1)

A. Real Party-In-Interest under §42.8.(b)(1)

Google LLC (“Petitioner”) is the real party-in-interest to this IPR petition.¹

B. Related Matters under §42.8(b)(2)

The ’972 patent was the subject of pending litigation involving Petitioner: *Cellular South, Inc. v. Google LLC*, Case No. 4:25-cv-01487-YGR (N.D. Cal.). Petitioner was served on May 10, 2024. (EX1013, p.001.) The case was originally filed in the Western District of Texas and was transferred to the Northern District of California on February 12, 2025.

C. Lead and Back-Up Counsel under §42.8(b)(3)

Petitioner provides the following designation of counsel.

LEAD COUNSEL	BACK-UP COUNSEL
Heidi L. Keefe (Reg. No. 40,673) hkeefe@cooley.com COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Ave. NW, Suite 700 Washington, DC 20004 Tel: (650) 843-5001 Fax: (650) 849-7400	Andrew C. Mace (Reg. No. 63,342) amace@cooley.com Mark R. Weinstein (Admission <i>pro hac vice</i> to be requested) mweinstein@cooley.com

¹ Google LLC is a subsidiary of XXVI Holdings Inc., which is a subsidiary of Alphabet Inc. XXVI Holdings Inc. and Alphabet Inc. are not real parties in interest to this proceeding.

LEAD COUNSEL	BACK-UP COUNSEL
	Reuben Chen (Admission <i>pro hac vice</i> to be requested) rchen@cooley.com Alexandra D. Leeper (Admission <i>pro hac vice</i> to be requested) aleeper@cooley.com COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Ave. NW, Suite 700 Washington D.C. 20004

D. Service Information

This Petition is being served by Federal Express to the attorney of record for the '972 patent, 27890 - STEPTOE LLP/DC, 1330 CONNECTICUT AVENUE, N.W., WASHINGTON, DC 20036. Petitioner consents to electronic service at the addresses provided above for lead and back-up counsel.

II. FEE PAYMENT

Petitioner requests review of 20 claims, with a \$51,875 payment.

III. REQUIREMENTS UNDER §§ 42.104 AND 42.108 AND CONSIDERATIONS UNDER §§ 314(A) AND 325(D)

A. Grounds for Standing

Petitioner certifies that the '972 patent is available for IPR and that Petitioner is not barred or otherwise estopped.

B. Identification of Challenge and Statement of Precise Relief Requested

Petitioner requests IPR institution based on:

Ground	Claims	Basis for Challenge under §103
1	1-20	Fontana in view of Lau
2	8-9, 20	Ground 1 prior art in further view of Arakawa

Submitted with this Petition is the Declaration of Henry Houh, Ph.D. (EX1002) (“Houh”), a qualified technical expert. (EX1002, ¶¶1-15, App’x A.)

C. Considerations Under §§ 314(a) and 325(d)

Petitioner hereby stipulates that if this IPR is instituted, then Petitioner will not pursue in the related pending litigation the specific grounds of invalidity that were raised or that reasonably could have been raised under 35 U.S.C. §§ 102 or 103 on the basis of prior art patents or printed publications in this IPR. To avoid any doubt, if the PTAB declines institution or rescinds institution of IPR, then Petitioner reserves the right to pursue any grounds of invalidity, including but not limited to the grounds raised or that reasonably could have been raised in this IPR, in the related pending litigation.

Petitioner respectfully submits that there is no section 314(a) or 325(d) issue that would warrant discretionary denial of the Petition.

§314(a): The *General Plastic* factors are not relevant; this is the first and only IPR petition filed by Petitioner with respect to the ’972 patent.

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

Nor do the *Fintiv* factors support discretionary denial under §314(a). As noted above, the pending litigation involving Petitioner was recently transferred to the Northern District of California. Before transfer, the pending litigation was still in an early pre-Answer stage, with no substantive discovery or claim construction having taken place. Since being transferred to the Northern District of California, the litigation has not substantively progressed. An initial case management conference is currently set for June 20, 2025. There is no schedule or trial date. Petitioner also intends to move to stay the litigation pending resolution of IPR.

§325(d): *Advanced Bionics* does not apply to Fontana or Arakawa as neither was presented during prosecution. Lau was not distinguished by the applicant, relied upon in any ground of rejection, cited in combination with any references, nor considered in combination with Kritt and Fontana. It was mentioned by the Examiner in passing solely as “rendering advertisements with rich media,” consistent with its use in this Petition. (EX1007, p.0069.)

In accordance with the Director’s March 26, 2025 Interim Processes for PTAB Workload Management memorandum and the related FAQs², Petitioner

² See <https://www.uspto.gov/patents/ptab/faqs/interim-processes-workload-management>.

reserves the right to address and respond to any assertions that Patent Owner may raise regarding discretionary factors in a Discretionary Denial Brief or otherwise.

IV. OVERVIEW OF THE PATENT

A. Level of Ordinary Skill

A person of ordinary skill would have possessed a bachelor's degree in electrical engineering, computer science, or similar field, with two years of experience in developing and implementing computer software for processing and/or analyzing multimedia content, such as audio, video, or image data. A person could also have qualified as a person of ordinary skill in the art with some combination of (1) more formal education (such as a master's of science degree) and less technical experience, or (2) less formal education and more technical or professional experience. (EX1002, ¶¶18-23.)

B. Specification Overview

The '972 patent purports to offer one solution to the common problem of how to reduce audio and visual information to written form to provide additional content-matched audio-visual information.

The '972 patent describes its purpose at a very high level. It states as its "Technical Field" that it "relates to a method and a system for generating various and useful data from videos." ('972, 1:10-11.) The summary of the invention confirms the patent's high-level approach: "The present invention is generally

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

directed to a method to generate data from video content, such as text and/or image-related information.” (’972, 1:27-29.)

Audio-visual data such as videos long predate the ’972 patent, and prior artists had long developed various ways to analyze, search, and generate video content. By the time of the alleged priority date (August 15, 2013), the fields of audio and image analysis for video content were well-developed. Indeed, the Background section of the ’972 patent acknowledges known image searching techniques to identify matching and similar images, as well as audio-to-text algorithms for transcribing text from audio. (’972, 1:15-23.)

The specification of the ’972 patent provides primarily a series of high-level, functional explanations for how to implement the alleged invention, with scant information on how to carry out any of its steps. For example, Figure 1 of the ’972 patent, shown below, “illustrates an embodiment of present invention,” and specifically “[a]n embodiment of video-to-text engine operation”:

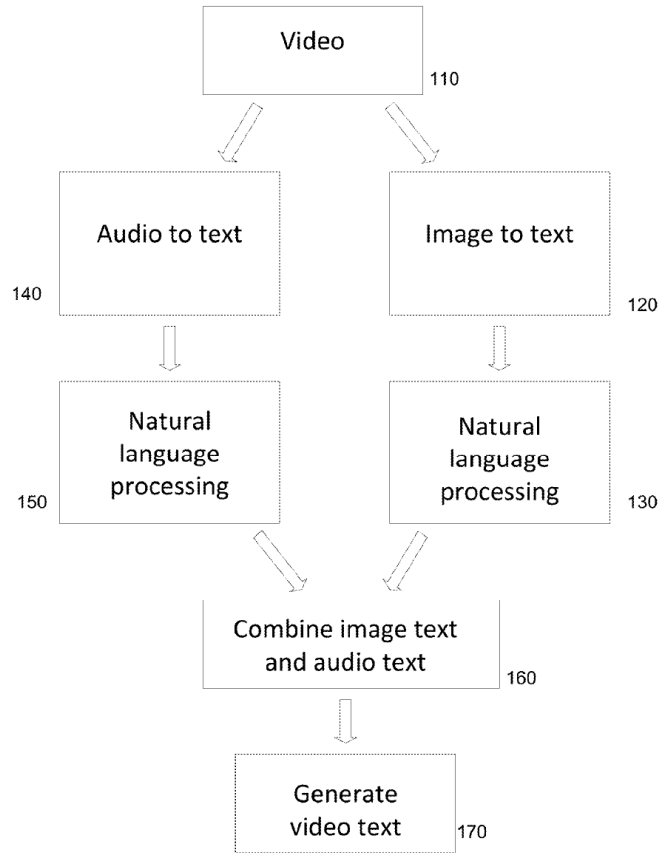


Figure 1

(’972, Fig. 1, 2:5, 3:51-52.) At step **110**, “a video stream is presented.” (’972, 3:52.) “The engine can extract audio data and image data (e.g. images or frames forming the video) from the video stream.” (’972, 3:58-59.) At step **120**, “the video-to-text engine performs an image data processing on the video stream[,]” such as symbol (or object) detection and identification, such as identifying a dog in the video image data. (’972, 3:64-65, 4:3-12.) Then, “[b]ased on the identified symbol, a plurality of instances of the symbol can be compared to a topic database to identify a topic (such as an event)[,]” such as identifying the dog as running or jumping. (’972, 4:20-23.) Thus, “text describing a symbol of the video and topic relating to the symbol

may be generated.” (’972, 4:26-28.) At step **130**, “natural language processing can be applied to the image text.” (’972, 4:61-62.) “For example, based on dictionary, grammar, and a knowledge database, the text extracted from video images can be modified as the video-to-text engine selects primary semantics from a plurality of possible semantics.” (’972, 4:62-66.) Applying dictionary definitions and grammatical rules to text was nothing new in 2013.

“In parallel, at **140**, the video-to-text engine can perform audio-to-text processing on audio data associated with the video.” (’972, 5:4-6.) At step **150**, “natural language processing can be applied to the text” and the text “can be given context, an applied sentiment, and topical weightings.” (’972, 5:41-45.) The specification is largely silent as to how to implement these natural language processing tools, which would have been well-known to a person of ordinary skill in the art (“POSA”). (’972, 1:15-23 (acknowledging audio-to-text was known).)

At step **160**, “the topics generated from an image or a frame and the topics extracted from audio can be combined. The text can be cross-referenced, and topics common to both texts would be given additional weights.” (’972, 5:46-49.) This is the common step of finding a way to turn an apples-to-oranges comparison (audio and video) into apples-to-apples (give everything a text descriptor or “topic” that can be combined and compared).

At step **170**, “the video-to-text engine generates video text, such as text

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

describing the content of the video, using the result of the combined text and cross reference.” (’972, 5:49-52.) This is summarizing the video, such as what a theater critic would do for a play. “For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized.” (’972, 5:52-54.) An “advertisement (text, images, or animation)” can then be generated based on the video text and added to the video stream. (’972, 6:24-29.)

V. CLAIM CONSTRUCTION

Petitioner does not believe express claim construction is necessary at this time. The prior art cited herein, as demonstrated below, renders the challenged claims obvious under any reasonable construction. Petitioner therefore respectfully submits that, for purposes of this IPR, express constructions are not required.

VI. THE CHALLENGED CLAIMS ARE UNPATENTABLE

A. Overview of Grounds

The Petition primarily relies on Fontana (**EX1003**) and Lau (**EX1004**), which like the ’972 patent disclose techniques for parallel processing of the audio and image data from video content and using the resulting analysis to generate additional content to add to the video. (Houh, ¶¶54-57, 60-62.) The Petition cites an additional reference, Arakawa (**EX1005**), in **Ground 2**, which discloses in detail the use of spectrum thresholds to segment audio. (Houh, ¶¶65-66.)

B. Prior Art Status of Relied-Upon References

Because the '972 patent claims priority to an earliest application filed August 15, 2013, AIA law applies to the challenged claims. All the references cited in the Grounds qualify as prior art under §102(a)(1) and §102(a)(2) because they were each patent applications filed and published before August 15, 2013.

C. Ground 1: Claims 1-20 Are Obvious Over Fontana in view of Lau

1. Independent Claim 1: “A method to generate video data from a video comprising:” (Claim 1[pre])

Assuming the preamble provides a claim limitation, Fontana discloses it. Fontana discloses a method that generates video data from a video, such as meta-data describing the video generated from performing audio processing (for example speech-to-text processing), and/or object recognition of visual content of a video. (Houh, ¶¶69-71.) Fontana teaches “constructing a set of text metadata describing an audio portion of the multimedia content, and generating a set of object metadata describing at least a portion of one or more objects appearing in the multimedia content.” (Fontana ¶0008.³) Fontana explains that the “multimedia content can include any type of content containing, for example, one or more of images, video, audio, or a combination thereof.... In the context of the present disclosure, a robust example of multimedia content is used in which video and audio information are

³ Unless otherwise noted, all emphasis added.

included[.]” (Fontana ¶¶0042.) Fontana can generate “text metadata” and “object metadata” from a video, making them “video data from a video.”

Therefore, Fontana discloses “[a] **method to generate video data from a video.**” As explained below, the method as claimed is obvious over Fontana and Lau.

(a) “**generating audio files and image files from the video;**”
(Claim 1[a])

Fontana discloses claim 1[a] by separately processing the audio and video content of the multimedia content, and further extracting thumbnail images (“**image files**”) from the video. (Houh, ¶¶72-75.)

For example, Fontana teaches “processing of multimedia content at an audio processing module 606, a video processing module 608, and a video conversion module 610” where “[e]ach of these modules can be executed concurrently (e.g., in parallel), with jobs associated with each module operating on one or more computing systems as defined by a scheduler (e.g., scheduler **406** of Fig. 4).” (Fontana ¶¶0090.) “The audio processing module 606 is configured to process audio content associated with the multimedia content” while the “video processing module 608 is configured to process the video portion(s) of multimedia content[.]” (Fontana ¶¶0091, 0094.) Fontana discloses that the audio and video may be separately processed by “segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.)[.]” (Fontana ¶¶0135.) As described below, Fontana teaches that

these audio and video processing modules are run on audio and video content extracted (“**generated**”) from the multimedia content. A POSA would further appreciate that a video, separated from its audio, is a series of images, such that video files are a form of “**image files.**” (Houh, ¶73.)

Fontana further teaches that the “audio processing module” can be run on “**audio files**” **generated** from the video. Specifically, Fontana teaches that “an audio separation operation 1128 strips, or extracts, the audio from the multimedia content.” (Fontana ¶0164.) A POSA would have understood that stripping or extracting the audio from the multimedia content generates an audio file. (Houh, ¶74.) Fontana additionally teaches that “that a number of sources provide speech to text conversion program.” (Fontana ¶0169.) One such program disclosed in Fontana is “large vocabulary continuous speech recognition (LVCSR) engines” that “depend on a language model that includes a vocabulary/dictionary for speech-to-text conversion of audio files.” (Fontana ¶0169.) A POSA would thus have appreciated that Fontana’s disclosures encompass “**generating audio files ... from the video.**” (Houh, ¶74.)

Fontana additionally teaches that the “video processing module” **generates** another set of “**image files**” as thumbnail images derived from the video. Fontana teaches “a thumbnail extraction module” that “is arranged to generate thumbnails In some embodiments, the thumbnail extraction module 618 generates a series

of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶0095.) A POSA would appreciate that a thumbnail is an image. (Houh, ¶75.) Such images were commonly stored as image files in 2013, such as in gif, tiff, jpg, or tiff formats, for example, which are common formats for graphical image files. (*Id.*) A POSA would thus have appreciated that by extracting thumbnail images from the video, Fontana’s disclosures satisfy “**generating ... image files from the video.**” (*Id.*)

- (b) “**distributing the image files across a plurality of processors and processing the image files in parallel, wherein processing the image files comprises extracting one or more objects and identifying the one or more objects;**” (Claim 1[b])

Fontana discloses or renders obvious claim 1[b]. (Houh, ¶¶76-86.)

Fontana implements its method using “a plurality of distributed computing systems[.]” (Fontana ¶0045.) A POSA would have understood that a distributed computer system is a form of “parallel” processing—by distributing tasks across a plurality of processors, each processor handles a portion of the tasks, resulting in parallel processing of the set of requests. (Houh, ¶77.) Fontana’s distributed computing system allows tasks, including “video, image or audio algorithms[.]” to be “separated and performed across multiple computing systems in parallel.” (Fontana ¶0053; *see also id.* ¶0090.) As further detailed below, Fontana’s distributed computing system includes both parallel processing of image files

separate from audio files, and processing multiple image files in parallel. (Houh, ¶77.) Therefore, Fontana discloses that its distributed computing system distributes image processing requests, such that it **“distribut[es] the image files across a plurality of processors and process[es] the image files in parallel.”**

For example, Fontana teaches:

The network 200 can, in certain embodiments, correspond to an architecture underlying the multimedia processing system 104 of FIG. 1, for example in a cloud-based or other distributed computing environment. The network 200 includes, in the embodiment shown, a workflow server 202 interconnected to an integration framework 204 and a storage network 206. The integration framework 204 provides interconnectivity and data sharing among a plurality of computing systems, such that the computing systems can share workloads, messages, and other tasks.

(Fontana ¶0048.)

Fontana explains that this workflow server “distributes one or more portions of jobs associated with each data processing request to the integration framework 204 and the storage network 206.” (Fontana ¶0050.) The result “allow[s] computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (*Id.*) Thus, the workflow server divides computationally intensive jobs, such as video processing (including image processing, as previously discussed), into portions, and distributes those portions

across multiple computing systems. A POSA would have appreciated that the result of this system is that these processing tasks, such as image processing, would be divided up and performed in parallel. (Houh, ¶¶78-79.)

Fontana teaches that its method is implemented using “a plurality of computing systems, illustrated as servers 302a-c.” (Fontana ¶0052.) These servers “are interfaced to inbound work ... for coordination and communication of data for processing.” (¶*Id.*) Fontana further teaches that “one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms” and are “configured to share processing jobs, such that tasks can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel.” (Fontana ¶0053.) As a result, “any of those computing systems can perform all or a portion of a processing job as defined by a scheduling algorithm, allowing multimedia content to be processed efficiently when necessary.” (Fontana ¶0055.) A POSA would have appreciated that the described servers that share processing jobs carry out parallel processing. (Houh, ¶80.) In the context of image processing, it would have been obvious to a POSA that image processing within Fontana’s distributed computing system could **process the image files in parallel** by, for example, distributing the image files for processing across multiple servers. (*Id.*)

Fontana further discloses processing images files to “**extract[] one or more**

objects and identify[] the one or more objects” as part of **“processing the image files in parallel.”** Fontana explains that it “includes a scheduler 406” that “receives tasks from the frontend 402 as defined by content providers, for example indicating that multimedia content should be processed to generate one or more objects of interest.” (Fontana ¶0061.) “The scheduler 406 receives and routes the content and processing requests to the desired computing systems within the grid 408” and “provides the ability to equally distribute resources to all jobs that are running at once[.]” (¶*Id.*) A POSA would have appreciated the scheduler is what distributes tasks among servers for parallel processing, and that Fontana’s scheduler distributes tasks to process multimedia content to generate one or more objects of interest. (Houh, ¶81.)

As discussed above for claim 1[a], Fontana’s image processing algorithm generates image files in the form of thumbnail images. Fontana discloses **“extracting one or more objects and identifying the one or more objects”** as part of processing these image files. The objects to be identified by Fontana can be either provided to the system or automatically identified by the system:

The thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest ... The objects of interest module 620 generates one or more objects of interest as defined in metadata to be associated with the multimedia content. In various embodiments, the

objects of interest module 620 can accommodate input from content providers to identify the objects of interest, or can at least partially automatically identify at least candidate objects of interest for confirmation by a user.

(Fontana ¶0095)

Fontana explains that identification of “objects of interest” can occur automatically within the thumbnail images: “boundaries of a number of candidate objects of interest could be automatically detected within one or more thumbnails[.]”

(Fontana ¶0142.) Fontana further teaches “**extracting one or more objects**” by “generat[ing] a ‘filmstrip’ which is a strip of thumbnails containing ‘objects of interest’ from the video. These objects of interest can be items, people, or conditions in the video that the viewer may be interested in[.]” (Fontana ¶0148.)

Fontana additionally teaches “**identifying the one or more objects.**” Fontana discloses, for example, that objects can be identified using OpenCV or other tools that would have been familiar to a POSA:

After the content is received, a candidate object generation operation 906 generates candidate objects of interest from the multimedia content.... The candidate object generation module can be performed by any of a number of object recognition programs, including computer vision programs. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition, gesture recognition, object identification, segmentation, and calibration tools. Other tools, such as MatLab or scale-invariant feature transform (SIFT)

algorithms could be included in the object detection process as well.
(Fontana ¶0139; Houh ¶85.)

Fontana teaches that objects could also be identified through “a neural network or other learning model.” (Fontana ¶0140.) Fontana provides as examples of such neural networks those developed by Numenta, Inc.; Vidient Systems, Inc.; and Behavioral Recognition Systems, Inc. (*Id.*)

(c) “processing the audio files;” (Claim 1[c])

Fontana discloses “**processing the audio files,**” such as processing associated with performing speech-to-text recognition. (Houh ¶87.) Fontana discloses an “audio processing module 606” that “is configured to process audio content associated with the multimedia content. In certain embodiments, the audio processing module 606 is configured to generate a full text transcript of the audio included in the multimedia content.” (Fontana ¶0091.) As discussed for claim 1[a], Fontana discloses that this speech-to-text recognition may be performed on **audio files.**

**(d) “converting audio files associated with the video to text;”
(Claim 1[d])**

As noted for claim 1[c], Fontana discloses “**converting audio files associated with the video to text,**” such as performing speech-to-text recognition to convert audio from the multimedia content (which may be a video, as discussed for the

preamble) to text. As discussed for claim 1[a], this speech-to-text recognition may be performed on **audio files**. (Houh ¶88.)

(e) **“converting the image files associated with the video to video data;” (Claim 1[e])**

Fontana discloses **“converting the image files associated with the video to video data,”** such as performing facial and/or object recognition to generate data and meta-data about the video content from thumbnail images. (Houh ¶¶89-90.)

As discussed for claim 1[b], Fontana discloses “an objects of interest module 620” that “generates one or more objects of interest.” (Fontana ¶0095.) Fontana also discloses a “candidate object generation operation” that “can generate a number of candidate objects of interest defined by the content provider.” (Fontana ¶0141.) Additionally, “boundaries of a number of candidate objects of interest could be automatically detected within one or more thumbnails[.]” (Fontana ¶0142.) These modules are discussed above for claim 1[b]. A POSA would appreciate that the data generated by the objects of interest module and the candidate object generation operation are **“video data”** as they are data about the video content. (*See, e.g.*, ’972, 1:10-11 (“The present invention relates to a method and a system for generating various and useful data from videos.”); *id.*, 1:27-29 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”); Houh ¶90.)

- (f) **“generating a topical meta-data that describes content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio files;” (Claim 1[f])**

Fontana discloses **“generating a topical meta-data that describes content of the video”** in the form of contextual information, an indexed transcript, keywords, and other meta-data derived from Fontana’s audio and video analyses. (Houh ¶¶91-98.)

The ’972 patent explains topical meta-data as follows:

Further, the engine applies the topical meta-data to the original full video. The image topics can be stored as topics for the entire video or each image segment. The topic generation process can be repeated for all identifiable symbols in a video in a distributed process. The outcome would be several topical descriptors of the content within a video. An example of the aggregate information that would be derived using the above example would be understanding that the video presented a dog, which was jumping, on the beach, with people, by a resort.

(’972, 4:33-44.) Thus, the topical meta-data can include at least semantic understanding of identified objects (e.g., dog, beach, people, resort) and events (e.g., jumping). (Houh ¶92.)

Fontana generates metadata describing both the objects recognized in the images and the audio processing. For objects, Fontana discloses “generat[ing] object metadata.” (Fontana ¶0120.) Likewise, “an object metadata operation 806 generates

object metadata corresponding to” *inter alia*, “objects appearing in or mentioned in the multimedia content. For example, the object metadata ... can also define people or objects appearing in the content as well.” (*Id.*) A POSA would understand the object metadata to correspond to topical meta-data, in the form of identifying people and objects appearing in the video. (Houh ¶93.)

A POSA would further understand that Fontana discloses identifying topical meta-data in the form of what the ’972 patent refers to as “events,” namely, actions performed by the identified people or objects. (Houh ¶94.) Fontana teaches implementing “[t]he candidate object generation module” using, for example, “OpenCV, which is a library of motion tracking, facial recognition, gesture recognition, object identification, segmentation, and calibration tools.” (Fontana ¶0139.) Motions and gestures are events, such as “jumping,” that the ’972 patent specification identifies as topical meta-data. (*See* ’972, 4:20-23, 4:41-44.)

For audio, “[a] text metadata operation 808 defines text metadata associated with the multimedia content,” which “can include a transcript of audio data included in the multimedia content, as well as additional textual information that a content presenter would like to display alongside the streamed multimedia content, such as additional contextual information.” (Fontana ¶0121.) Furthermore, “[t]he transcript can be indexed ... to allow content consumers to search the spoken text transcript, as well as other descriptive information related to the multimedia content.” (*Id.*) As

a result, the text metadata can include additional text-based descriptive and contextual information about the video content, beyond the transcript of the audio.

(Houh ¶95.)

For example, Fontana discloses identifying keywords in the transcript, which enables searching of the multimedia content: “The text index information 630 can be used to provide a corresponding transcript ... or can be used to provide keyword searchability of the multimedia content.” (Fontana ¶0107.) Fontana describes example keyword data for facilitating searching the multimedia content:

FIG. 7I illustrates example keyword data that can be used in association with particular content to facilitate searching of that content. In certain embodiments, the keyword data 706 can be used as a substitute for the text information 630, or can be used to reference a particular location within the text information to allow searching of content or metadata describing the content. ... In certain embodiments, the keyword data 706 can be made available to external search engines, to allow the content or portions of the content to be made available for search access by search engines that are remote from and unaffiliated with the systems and methods described herein.

(Fontana ¶0114.)

A POSA would have appreciated that Fontana’s above-described object metadata and text metadata (including both transcript and keywords) are all “**topical meta-data that describes content of the video.**” (Houh ¶¶96-97.)

Furthermore, Fontana’s contextual information and meta-data, which correspond to the “**topical meta-data**” of claim 1, are generated “**by deriving semantic information from the identification of the one or more objects and semantic information from the audio files.**” (Houh ¶98.) A POSA would understand “**semantic information**” to refer to information conveying or associated with the meaning of content. (*Id.*) Fontana’s topical meta-data is generated by deriving semantic information from the object identification and audio processing because it includes words and concepts ascribing meaning to the audio and video content. For example, a POSA would recognize keywords identified from a text-to-speech transcript as semantic information because they convey meaning about the audio. (*See id.*; *see also, e.g.*, Fontana ¶0107 (“The text index information 630 ... can be used to provide keyword searchability of the multimedia content.”).) Likewise, a POSA would recognize identifications of people or objects as semantic information about the meaning of a video or image. (Houh ¶98.)

(g) “adding the topical meta-data to the video; and” (Claim 1[g])

Fontana discloses or renders obvious “**adding the topical meta-data to the video.**” (Houh ¶¶99-106.) For example, Fontana teaches that “specific start and end times can be defined, as associated with specific segments of the transcription text. In this way, the transcript could be linked, portion by portion, to the multimedia content based on the time at which the transcribed words are played in the content.”

(Fontana ¶0107.) A POSA would have appreciated that this specific linking of transcript to video “**add[s] the topical meta-data to the video.**” (Houh ¶99.) The ’972 patent specification discloses a similar linking of topical meta-data to the video: “Data generated from an image and/or from audio transcription can be time stamped, for example, according to when it appeared, was heard, and/or according to the video frame from which it was pulled.” (’972, 4:28-31.)

Similarly, Fontana teaches that “keyword data 706 ... can be used to reference a particular location within the text information... Other information can be included in the keyword data as well (e.g., links to a particular location within the multimedia content, or other associated keywords, etc.).” (Fontana ¶0114.)

Furthermore, Fontana teaches that “text index information 630 can be used to provide a corresponding transcript alongside playback of multimedia content.” (Fontana ¶0107.) A POSA would have appreciated that this too was a form of “**adding the topical meta-data to the video.**” (Houh ¶101.)

Fontana additionally teaches that the topical meta-data can be presented concurrently with the video content, which is another form of “**adding the topical meta-data to the video.**” Fontana’s “text metadata,” which is **topical meta-data**, “can take any of a number of forms, and can include ... additional textual information that a content presenter would like to display alongside the streamed multimedia content.” (Fontana ¶0121; *see also* Fontana ¶0147 (example “actions

associated with each object of interest ... include display of contextual information identifying the object”).) A POSA would have appreciated that presenting text metadata alongside the video “**add[s] the topical meta-data to the video.**” (Houh ¶102.)

Additionally, Fontana discloses a “container operation 810” that “applies a container to the received multimedia content.” (Fontana ¶0122; *see also* Fontana ¶0176.) Such containers were well-known to a POSA at the time, as they were a common way to store the various components of multimedia content such as videos. (Houh ¶103.) Fontana discloses several container formats, such as “an Adobe Flash format” as well as “HTML5, Microsoft Silverlight, or other formats[.]” (Fontana ¶0122.)

A POSA would have known that container files such as those disclosed in Fontana routinely store metadata as well as multimedia content, and thus would have found it obvious to apply Fontana’s container operation to both the multimedia content and its associated metadata. (Houh ¶104.) For example, MP3 was a well-known container file format for audio, and a POSA would have known that “MP3 files are capable of storing a certain amount of ‘meta-data’—extra information about each file—inside the file itself,” which “can be added or edited later on, often directly through your MP3 player’s interface.” (EX1009 p.6; *see also id.*, pp. 44-45, 105, 114, 116, 363-64, 369; *see also, e.g.*, EX1008 (describing Adobe Flash

Format metadata options); EX1010 ¶0040 (describing a method to add a metadata track to video data); Houh ¶104.)

Disclosures in Fontana would have confirmed to a POSA that the container could include both the video and its topical meta-data. Fontana explains that the “container operation 1304 can be performed by the multimedia processing systems of the present disclosure, with the container and associated metadata being stored ... by the multimedia processing systems[.]” (Fontana ¶0176.) As a further example, Fontana teaches “[a] storage operation 812 stores the content and associated metadata for use.” (Fontana ¶0123; *see also* Fontana ¶0066 (describing storage of metadata).) A POSA would have appreciated that a container operation was one such storage operation. (Houh ¶¶105-106.)

**(h) “cross-referencing the text and the video data based on the generated topical meta-data to determine topics;”
(Claim 1[h])**

Claim 1[h] is obvious over Fontana in view of Lau. (Houh ¶¶107-120.)

In the combination, the **topical meta-data** generated by Fontana would be searched using the method of Lau to determine a relevant context for the video and ultimately identify advertisements for display at specific times during a video. This search process **cross-references the text and the video data** by matching advertisements using keywords from **the text** (the results of the speech-to-text recognition) and content from the **video data** (metadata on the identified objects of

interest) **based on the generated topical meta-data** (i.e., by incorporating context and other metadata, *see* claim 1[f]). (Houh ¶108.)

Lau discloses performing a search to determine advertisements whose content and/or topics match the content and/or topics for portions of video content. The search counts the number of keyword and/or concept matches for a search term near a particular time in the video. A POSA would have found it obvious to combine Lau and Fontana by using the **topical meta-data** generated by Fontana (*see* claims 1[f] and 1[g]) as well as **the text and the video data** (*see* claims 1[d] and 1[e]) as keywords and concepts searched by Lau for purposes of identifying ads or ad units to match to the video based on context or subject matter. Each of the ads, ad units, and context or subject matter of the video is a “**topic.**”

As discussed for claim 1[f], the **topical meta-data** (various contextual information and metadata) is derived from **the text and the video data**, such that searching the topical meta-data **cross-references the text and the video data**, and any such search is **based on the generated topical meta-data**.

A POSA would have further found it obvious to use the text and the video data as part of Lau’s matching search, which further **cross-references the text and the video data**. (Houh ¶109.) Lau teaches matching keywords and concepts in a video to ads based on context or subject matter. (Lau ¶0051.) This matching aggregates together the different metadata types and uses the resulting aggregated

information to match the content to an ad: “for each ad, correlation engine 202 finds candidate content that may be relevant. This is done by searching for content in the index to match the keywords, categories, and concepts associated with the ad to information in the content.” (Lau ¶0056.) A POSA would have found it obvious to use the results of Fontana’s speech-to-text recognition (**the text**) to generate keywords and “information” for Lau’s search method. Similarly, a POSA would have found it obvious to use the results of Fontana’s object recognition module (**the video data**) as concepts and “information” for Lau. It would further be obvious to perform Lau’s search with reference to the contextual information derived from Fontana, which is the **topical meta-data** and is additionally “information in the content.” (See Lau ¶0056; Houh ¶110.)

Lau teaches that the search to match advertisements to video content may combine the results of different analyses, such as text with images as well as conceptual ideas like context: “Advertisers may buy correlation information, such as keywords, phrases or concepts... The phrases may be any combination of words and other information, such as symbols, images, etc. The concepts may be a conceptual idea of something.” (Lau ¶0046.) Thus, a POSA would have found it obvious that correlating advertisements to the video could **cross-reference the text** (speech-to-text transcript and any derived keywords) **and the video data** (the identified objects of interest) **based on the generated topical meta-data** (additional

contextual information). (Houh ¶110.)

This contextually aware keyword and topic searching aligns with the '972 specification's teaching regarding cross-referencing:

At 160, the topics generated from an image or a frame and the topics extracted from audio can be combined. The text can be cross-referenced, and topics common to both texts would be given additional weights. ... For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized.

('972, 5:46-54.)

Returning to Lau, Lau discloses using its method to **determine topics**, that is, subject matter for the ads and ad units. For each ad, Lau identifies a set of candidate locations within the video, then scores each based on the strength of the match based on the cross-referencing of keyword/concept matches:

Correlation engine 202 locates the times where the keywords and concepts match. For each candidate time, correlation engine 202 creates an “ad anchor” holding the score for the match. The score may be a linear combination of the following weights:

...

2. Concentration of the match—the more keywords/concepts for the ad matches near the time, the higher the score.

(Lau ¶¶0057–0059.)

Lau additionally discloses that “[a]n advertisement may be broken into ad units” that are each “a subset of a larger advertisement.” (Lau ¶0023.) As a result,

“[c]orrelation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter” to combine as the final advertisement. (Lau ¶0038.)

Searching for candidate matches for ads or ad units based on the subject matter, and assigning weights to those candidates, each involves **cross-referencing the text and the video data based on the generated topical meta-data** to determine ad or ad unit matches based on subject matter or context (“**topics**”). (Houh ¶114.)

Rationale and Motivation to Combine (Fontana with Lau): A POSA would have been motivated to combine Fontana and Lau in order to use Lau’s method for placing advertisements within a video to carry out Fontana’s disclosures of linking advertisements to video content, and would have had a reasonable expectation of success in the combination. (Houh ¶¶115-120.) Fontana and Lau are analogous references to the ’972 patent: all are in the field of processing and generating multimedia content, including video, and specifically address searching within multimedia content. (’972, 1:27-29 (“a method to generate data from video content, such as text and/or image-related information.”); *see also* ’972, Abstract; Fontana ¶0001 (“systems and methods for processing and delivery of multimedia content.”); Fontana ¶0041 (“methods and systems for receipt, processing, and delivery of multimedia content, as well as enrichment of multimedia content for

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

enhanced search and delivery.”); Lau ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”); Lau ¶0006 (“In one embodiment, an advertisement is matched to subject matter in a portion of rich media content, such as digital video”); Lau ¶0053 (“Content may be searched to determine if the content includes the keywords.”).) Both further disclose linking advertisements to the multimedia content. (’972, 1:57-59 (“An advertisement can be placed at a specific time in the video based on the video content and/or section symbol of a video image.”); Fontana ¶0117 (“example advertisement data 716 ... to link one or more advertisements with multimedia content during playback.”); Lau ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”).)

Fontana expressly discloses that advertisement data could be used “to link one or more advertisements with multimedia content during playback.” (Fontana ¶0117.) Fontana further teaches that “the matching of advertisements and content occurs based on a decision process separate from the content delivery system of the present disclosure.” (Fontana ¶0117.) A POSA would have been motivated to combine Fontana with an advertisement-matching method, such as Lau, to practice Fontana’s disclosures of linking advertisements with multimedia content. (Houh ¶116.)

It would have been obvious to a POSA that the text and object meta-data generated by Fontana could be used as the keywords/content to search for matching advertisements to video content according to Lau. (Houh ¶117.) Fontana makes clear that its keywords are suitable for searching within the video, including with search engines outside of Fontana’s disclosures. (*See, e.g.*, Fontana ¶0121 (indexed transcript to allow “search[ing] the spoken text transcript”); Fontana ¶0107 (text index “to provide keyword searchability of the multimedia content”); Fontana ¶0114 (keywords “can be used to reference a particular location within the text information to allow searching of content or metadata describing the content” including by “external search engines ... that are remote from and unaffiliated with the systems and methods described herein”); Fontana ¶0042; Fontana ¶¶0126-0127 (keyword searching of metadata, including by “a remote decision engine”).) A POSA would have found it obvious that Lau’s method of searching for ad matches is one such keyword-based search engine. (Houh ¶¶117-120.) Thus, a POSA would have appreciated that the **topical meta-data** generated by Fontana could be used as the keyword/content information for the video in Lau’s search method, and Fontana indeed encourages such combinations. (Houh ¶¶115-120.)

- (i) **“generating video text based on the cross-referencing, wherein the video text describes content of the video;”**
(Claim 1[i])

The combination of Fontana with Lau renders obvious claim 1[i]. Lau

discloses matching ads to multimedia content through **cross-referencing** by determining a context or subject matter (for simplicity, “context”) (“**video text**”) that “**describes content of the video**” and using that context to determine “ad units” from which the ads are generated. Additionally, Fontana discloses storing “advertisement data” including “topics, keywords, or content” that can be matched or related to the video content as additional “**video text.**” (Houh ¶¶121-127.)

As discussed for claim 1[h], the result of the cross-referencing are the ads, ad units, and context/subject matter of the video. Lau explains that an “ad unit may be created by taking a static ad and augmenting the unit with an advertiser-specified message dependent on **context** and **keywords.**” (Lau ¶0027.) A POSA would have appreciated that this includes determining “context,” which is a way of describing what a video is about beyond just keywords. In other words, a POSA would have appreciated that “context” as used in Lau can be a form of **video text based on the cross-referencing** in claim 1[h] and that **describes content of the video.** (Houh ¶123.)

Lau provides a BMW ad example:

Correlation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter. For example, based on one or more keywords, ad units from the ad matrix are determined. ... One example of this is BMW may provide a general ad unit for their logo and have a different ad unit for different models,

such as the 330 model, 530 model, etc. ... If the content talks about the 330 model then the logo and the 330 ad units may be combined and presented to the user.

(Lau ¶0038.)

In this example, Lau determines that the “subject matter” of the video is the BMW 330 model. This subject matter (or context) is **video text** (BMW 330 model) that is **based on the cross-referencing** (identified “based on one or more keywords,” which can search text and image metadata (*see* Lau ¶0046), as the model appearing in the video content) and that **describes content of the video** (“the content talks about the 330 model”). (*See* Lau ¶0038.) A POSA would therefore have recognized that Lau teaches **generating video text based on the cross-referencing, wherein the video text describes content of the video.** (Houh ¶¶122-126.)

In the combination of Fontana and Lau, information about the advertisements selected through Lau’s method are then incorporated into Fontana’s “advertisement data” to facilitate display to the viewer:

[T]he advertisement data 716 can include an advertiser identifier, a definition of an advertisement, and associated topics, keywords, or content that can be linked to the advertisement. In certain embodiments, the advertisement data 716 is used to link the content to advertisements during playback...

(Fontana ¶0117.)

A POSA would have appreciated that Fontana’s “advertisement data” in combination with Lau is another form of **video text** (e.g., “a definition of an advertisement, and associated topics, keywords, or content”) that is **based on the cross-referencing** (because derived from Lau’s method of matching advertisements to video content) and that **describes content of the video** (such as data on “associated topics, keywords, or content” and other data “used to link the content to advertisements during playback”). (Houh ¶¶126-127.)

(j) “generating a text, image, or animation based on the video text; and” (Claim 1[j])

The combination of Fontana with Lau renders obvious claim 1[j]. Lau teaches that “ad units” are selected based on context (which is **video text**, see claim 1[i]). Once ad units are selected, see claim 1[i], “[t]he ad units are then combined into an advertisement that is correlated to the subject matter.” (Lau ¶0038.) This advertisement may be a **text** (such as “an advertiser-specified message”, Lau ¶0027), **image** (such as the BMW logo, Lau ¶0038), or **animation** (such as a “video that may serve as pre/mid/post-roll”, Lau ¶0027). The advertisement is **based on the video text** as it is the result of combining the ad units, which are selected by matching to context (**video text**), together into an advertisement. (See Lau ¶0038; Houh ¶128.)

(k) **“placing the text, image, or animation in the video.”**
(Claim 1[k])

The combination of Fontana and Lau **plac[es] the text, image, or animation in the video**. Lau teaches placing the advertisement in the video by turning ad units into “video that may serve as pre/mid/post-roll.” (Lau ¶0027.) Lau further explains that “the advertisement may be displayed in serial, parallel, or be injected into the rich media content.” (Lau ¶0034; *see also* Lau ¶0084 (“rendering formatter 204 can determine that an advertisement should be rendered serially relative to the portion of rich media content, in parallel to the portion of rich media content, or injected into the rich media content”); Lau claim 5 (“the advertisement is injected into or laid on top of portion of rich media content”).) A POSA would have appreciated that injecting the advertisement into the rich media content (i.e., video), **plac[es] the text, image, or animation in the video**. (Houh ¶¶129-130.)

The combination of Fontana with Lau therefore renders obvious claim 1.

2. **Claim 2: “The method according to claim 1, further comprising: generating a content-rich video based on the video, the text, and the video data.”**

The additional limitations of claim 2 are obvious over Fontana in view of Lau. The result of the combination—a video indexed to searchable metadata and into which an advertisement has been added—is a **content-rich video**. This content-rich video is generated based on **the video** (the initial multimedia content, see claim 1[pre]), **the text** (converted from the audio files, see claim 1[d]), **and the video data**

(the object metadata, see claim 1[e]). (Houh ¶¶131-132.)

3. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”

The additional limitations of claim 3 are obvious over Fontana in view of Lau. (Houh ¶¶133-136.) As discussed for claim 1[d], Fontana teaches converting the audio files associated with the video to text as a transcript. Fontana further teaches analyzing this transcript using **natural language processing**. (Fontana ¶0184 (“The search performed within the content can in certain embodiments, be performed based on natural language processing of an existing transcript”); *see also* Fontana ¶0100 (“In alternative embodiments, additional search arrangements can be included as well, such as a natural language search[.]”).)

Fontana then determines the topic of the multimedia content—the **context associated with the video**—using natural language processing. Specifically, Fontana teaches that search queries, including those leveraging the above natural language processing, can be used **to determine context associated with the video** in the form of keywords or relevant portions of the multimedia content: “A content request operation 814 receives a request related to multimedia content... such as a search query related to keywords appearing in one or more fields of metadata associated with the content (e.g., titles, authors, producers, genre, etc.) or in the transcript or other text associated with one or more pieces of content.” (Fontana

¶0126.) Additionally, “[t]he provide metadata operation 816 selects at least a portion of the metadata associated with the content (e.g., including ... transcript information,...) for inclusion with the content during playback.” (Fontana ¶0127.)

A POSA would have appreciated that identifying metadata, portions of the transcript, and portions of the multimedia data related to keywords or other queries is **determin[ing] context associated with the video.** (Houh ¶¶134-135.)

Lau also discloses using **natural language processing** to determine context associated with a video for purposes of matching it to an advertisement. Lau teaches that determining ad units based on keywords can use natural language processing: “Recognition engine 212 receives rich media content that may be accessed by a user and uses correlation recognition detection techniques to recognize the content...In another embodiment, it could be [a] natural language processing engine.” (Lau, ¶0041.) A POSA would have appreciated that using a natural language processing engine to recognize the content is **applying natural language processing to the text to determine context associated with the video.** (Houh ¶136.)

4. Claim 4: “The method according to claim 2, further comprising: applying natural language processing to the text to extract the topical meta-data.”

The additional limitations of claim 4 are obvious over Fontana in view of Lau. (Houh ¶¶137-138.) Fontana’s “search query related to keywords appearing in one or more fields of metadata associated with the content ... or in the transcript or other

text associated with one or more pieces of content” are a search within **the topical meta-data**. (Fontana ¶0126.) Fontana teaches that searching can **apply natural language processing**. (See claim 3; Fontana ¶¶0100, 0184.) Fontana additionally discloses “[a] provide metadata operation” that “provides metadata (and optionally the multimedia content) in response to the request.” (Fontana ¶0127.) It was obvious that the method could **apply natural language processing to the text** (transcript, which is part of the text metadata) **to extract the topical meta-data** (i.e., to provide metadata in response to the search request).

5. Claim 5: “The method according to claim 1, further comprising: processing the image files to extract additional text.”

The additional limitations of claim 5 are obvious over Fontana in view of Lau. (Houh ¶¶139-141.) Lau explains that “[i]mage recognition can be used on visual portions of the rich media content. For example, optical character recognition (OCR).” (Lau ¶0040.) OCR was a well-known image detection algorithm that identifies text appearing in images and extracts that text. (Houh ¶¶140-141.)

It was obvious to use Lau’s OCR on Fontana’s **image files** (the thumbnail images) to improve generating video metadata and matching ads. (Houh ¶141.) Fontana teaches that “[o]nce a user has selected one or more objects of interest, a number of optional detection algorithms can be applied to further define those or other objects of interest.” (Fontana ¶0144.) Implementing OCR as an “optional

detection algorithm[]” would capitalize on Lau’s teaching that OCR can be performed as one of the “[c]orrelation recognition detection techniques” that “may be used to determine that the advertisement is correlated to the portion of rich media content.” (Lau ¶0040; Houh ¶141.)

6. Claim 6: “The method according to claim 5, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”

The additional limitations of claim 6 are disclosed by Fontana. (Houh ¶¶142-145.)

Fontana teaches segmenting the image files in several ways. As described for claim 1[b], Fontana teaches a distributed computing system that “allow[s] for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing and resulting usability for content providers.” (Fontana ¶0135.) Fontana’s workflow server “distributes one or more portions of jobs associated with each data processing request.” (Fontana ¶0050.) The result “allow[s] computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (*Id.*) This distributed computing system **segments** tasks including image processing, prior to processing the images in parallel. (Houh ¶143.)

Fontana also teaches **segmenting image files** into “a series of thumbnails

representing scenes throughout the multimedia content” prior to performing the object recognition. (Fontana ¶¶0095.) “In certain embodiments, the candidate object generation operation 906 splits the multimedia content into a plurality of sections, and generates a thumbnail image associated with each of those sections for preview by the content provider.” (Fontana ¶¶0139.) Fontana also discloses using “computer vision programs” such as “OpenCV,” which includes “segmentation” tools, as well as using “a video or other multimedia-editing web service” to “segment, edit, and reprocess the content.” (Fontana ¶¶0139; *id.* ¶¶0058.) Each of these **segments** the image files before parallel processing. (Houh ¶¶144-145.)

7. Claim 7: “The method according to claim 1, further comprising: determining a motion associated with the one or more objects.”

The additional limitations of claim 7 are disclosed or rendered obvious by Fontana. (Houh ¶¶146-148.) Fontana teaches using “OpenCV, which is a library of motion tracking, ... gesture recognition.” (Fontana ¶¶0139; *see also* Fontana ¶¶0144.) Motion tracking and gesture recognition **determine a motion associated with the one or more objects**. (Houh ¶¶147-148.)

8. Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”

The additional limitations of claim 8 are obvious over Fontana. (Houh ¶¶149-152.) Fontana teaches multiple servers with “audio algorithms” that “are configured

to share processing jobs, such that tasks can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel.” (Fontana ¶0053.) As a result, “any of those computing systems can perform all or a portion of a processing job ... allowing multimedia content to be processed efficiently when necessary.” (Fontana ¶0055.) Fontana additionally teaches that segmentation may occur in order to facilitate parallel processing. Fontana teaches that its “distributed computing systems ... allow for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing.” (Fontana ¶0135; *see also* Fontana ¶0058 (“instructions provided ... to segment, edit, and reprocess the content”).) Based on the foregoing, a POSA would thus have found it obvious that to implement audio processing in parallel, Fontana would **segment[] the audio files before processing the audio files in parallel.** (Houh ¶¶150-152.)

9. Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum thresholds.”

The additional limitations of claim 9 are obvious over Fontana. (Houh ¶¶153-156.) Performing the segmentation of audio and image files **at spectrum thresholds** could, for example, segment the audio and image files based on scene or content changes. Fontana teaches using scene or content divisions for thumbnails, “to

generate thumbnails at possible locations the content provider would like to create an object of interest (for example ... immediately following major scene or sound changes in the content). In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶¶0095.) A POSA would have appreciated both that this division of the content corresponds to a segmentation of the multimedia content, and that identifying “major scene or sound changes in the content” could be performed by identifying **spectrum thresholds**. (Houh ¶¶154-155.) A spectrum threshold is just a quantifiable valuation of some aspect of the media content at each moment in time—for example, the overall volume at a given time or frame of video, or its overall color saturation value. (Houh ¶155.) Indeed, the ’972 patent gives as an example of segmentation at spectrum thresholds that “the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data.” (’972, 5:16-20.) One common method for detecting major scene or sound changes was to look for changes above a **threshold**; such approaches were known to a POSA before 2013. (Houh ¶155; *see also* EX1011, 11:44-54; EX1012 ¶¶0008, 0093.)

More generally, a POSA would have known that performing speech-to-text analysis commonly involved a Fourier analysis of the audio and segmenting the

audio based on frequency to divide the speech into individual phonemes. (Houh ¶155.) As Fontana describes: “Phonetic-based applications separate conversations into phonemes, the smallest components of spoken language; they then find segments within the long file of phonemes that match a phonetic index file representation of target words, phrases and concepts[.]” (Fontana ¶0169; *see also* Fontana ¶0171 (listing example “[p]honetic-based applications useable as one or more of the speech to text conversion programs”).) These phonetic-based algorithms described in Fontana **segment audio files at spectrum thresholds**. (Houh ¶155.)

Additionally, periods of silence with low energy are commonly identified as gaps between words. (Houh ¶156.) For example, the “large vocabulary continuous speech recognition (LVCSR) engines” disclosed by Fontana identify gaps between words. (*See* Fontana ¶0169 (describing LVCSR searching text for “target words, phrases and concepts”); Fontana ¶0170 (listing example LVCSR conversions); Houh ¶156.) This common understanding of the use of silence to segment audio aligns with the ’972 patent’s disclosure:

In another example, quiet periods in the audio can be detected, and the segmentation can be defined by the quiet periods. For example the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data. The quiet periods in the audio data can be ignored,

and the processing requirements thereof can be reduced.

(’972, 5:13-22.) Therefore, a POSA would have found it obvious to use a spectrum threshold for segmenting audio and image files based on scene and/or sound changes, as taught by Fontana.

10. Claim 10: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”

The additional limitations of claim 10 are obvious over Fontana in view of Lau. (Houh ¶¶157-158.) Lau generates an advertisement by compiling ad units, each selected by matching the ad units to the video based on the text (i.e., the audio transcript) and the video data (i.e., the metadata generated by Fontana). The details of this matching and advertisement generation are explained above for claim 1.

11. Claim 11: “The method according to claim 10, further comprising: placing the advertisement in the video at a preferred time.”

The additional limitations of claim 11 are obvious over Fontana in view of Lau. (Houh ¶¶159-162.) Lau’s “advertisements are time aligned to correlate to the subject matter” and turned into “video that may serve as pre/mid/post-roll.” (Lau ¶¶0027, 0045; *see also* Lau ¶0034 (“advertisement may be displayed in serial, parallel, or be injected into the rich media content”); Lau ¶0084 (serial, parallel, or injected rendering); Lau claim 5.) These time-aligned advertisements are **placed in the video at a preferred time.**

12. Claim 12: “The method according to claim 6 wherein the additional text includes information regarding context associated with the video.”

The additional limitations of claim 12 are obvious over Fontana in view of Lau. (Houh ¶¶163-164.) It would have been obvious to a POSA that text generated from the OCR of image files (**the additional text** of claim 6) is **“information regarding context associated with the video.”** Lau explains that “correlation recognition detection techniques” including “optical character recognition (OCR)” “may be used to determine that the advertisement is correlated to the portion of rich media content.” (Lau ¶0040.) A POSA would understand that this corresponds to a “context associated with the video.” (Houh ¶164.)

13. Claim 13: “The method according to claim 6, wherein the additional text relates to a symbol appearing in the video.”

The additional limitations of claim 13 are obvious over Fontana in view of Lau for the same reasons as discussed for claims 5 and 6. (Houh ¶165.) A POSA would understand that OCR includes symbols such as letters, numbers, punctuation, and other typographic symbols (e.g., “+”). (*Id.*)

14. Claim 14: “The method according to claim 13, wherein the symbol is a brand logo, and wherein the additional text includes information regarding placement and time of appearance of the brand logo.”

The additional limitations of claim 14 are obvious over Fontana in view of Lau. (Houh ¶¶166-170.)

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

A POSA would have understood that the object recognition and OCR processing of the combination of Fontana with Lau could encompass recognizing brands. The '972 patent explains that techniques such as object recognition and OCR can be used to identify brands:

For example, the normalized image frame files can be analyzed for text identification and/or by optical character recognition. ... Such techniques can be improved by focusing on regions of interest, for example based on brands, logos, objects, and/or features of interest.

('972, 6:62-7:2.)

A POSA would have found this use of object recognition and/or OCR obvious in view of Fontana and Lau. (Houh ¶¶167-168.) Fontana expressly discloses that “[a]dditional objects of interest can be identified by a user,” thus permitting a user to specify, for example, a brand logo for the method to recognize. (Fontana ¶0089.) Similarly, Fontana explains that “objects and individuals appearing in the content” can be included in a script provided with the multimedia content. (Fontana ¶0088.) A POSA would have found it obvious that a brand logo could be one such object specified by a script as appearing within a video. (Houh ¶168.)

Furthermore, Lau expressly teaches that brands may be a particularly useful type of content to search for within the video in order to match ad units to specific locations within the video, by giving an example of identifying ad units corresponding to the BMW model discussed in the video: “BMW may provide a

general ad unit for their logo and have a different ad unit for different models, such as the 330 model, 530 model, etc. ... If the content talks about the 330 model then the logo and the 330 ad units may be combined and presented to the user.” (Lau ¶0038.) A POSA combining Fontana with Lau therefore would have found it obvious to use brand logos as objects of interest for identification within the image files. A POSA would have been motivated to do so in order to achieve the benefits described by Lau of improved matching of ad units to the video content, and would have had a reasonable expectation of success in light of Fontana’s disclosures discussed above that objects of interest may be specified by the user and/or an accompanying script. (Houh ¶169.)

As explained for claims 1[e], 1[f], and 5, the recognized objects (including the results of OCR and any brand logos) are converted into searchable metadata, i.e., **additional text**. And as explained for claim 1[h], Fontana teaches that this metadata is linked to where in the video content it occurs, such that **the additional text includes information regarding placement and time of appearance of the brand logo**.

15. Claim 15: “The method according to claim 1, wherein the one or more objects are letters appearing in the video.”

The additional limitations of claim 15 are obvious over Fontana in view of Lau for the same reasons as discussed for claim 5. (Houh ¶171.) A POSA would

have known that the **one or more objects** identified by OCR would include **letters appearing in the video**. (*Id.*)

16. Claim 16: “The method according to claim 6, wherein the additional text relates to faces appearing in the video.”

The additional limitations of claim 16 are obvious over Fontana in view of Lau as both references teach facial recognition as part of image processing. (Houh ¶¶172-174.) Fontana teaches using OpenCV, whose tools include “facial recognition,” as well as use of “additional detection algorithms” that “can include facial recognition[.]” (Fontana ¶¶0139, 0144.) Similarly, Lau teaches “facial recognition” as a form of “[i]mage recognition can be used on visual portions of the rich media content.” (Lau ¶0040.) The image recognition, including facial recognition, is then used to generate “meta-data about the visual content” (“**additional text**”). (Lau ¶0042; *see also* Lau ¶0040.)

17. Independent Claim 17: “A system for extracting data from a video, comprising:” (Claim 17[pre])

Assuming the preamble provides a claim limitation, Fontana discloses it. For the same reasons as claim 1[pre] and 1[a], Fontana discloses a system for extracting data from a video. To the extent the preamble of claim 1 is addressed to a “method” and the preamble of claim 17 is addressed to a “system,” Fontana teaches both a method and a system. (*See* Fontana ¶0008 (“In a first aspect, a method for providing

multimedia content is disclosed.”); Fontana ¶0009 (“In a second aspect, a system for providing multimedia content is disclosed.”); Houh ¶175.)

(a) “a computer processor having a plurality of processors for parallel processing; and” (Claim 17[a])

For the reasons explained for claim 1[b], Fontana discloses parallel processing. Fontana further discloses implementing its system using a **computer processor having a plurality of processors for parallel processing**:

In addition, electronic computing device 500 comprises a processing unit 504. ... In this first example, processing unit 504 may be implemented as one or more processing cores and/or as one or more separate microprocessors.

(Fontana ¶0074; Houh ¶176.)

(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 17[b])

Fontana teaches that its system is implemented on a **non-transitory computer readable medium containing instructions directing the system to execute** certain steps. Fontana explains:

FIG. 5 is a block diagram illustrating example physical components of an electronic computing device 500, which can be used to execute the various operations described above, ... A computing device, such as electronic computing device 500, typically includes at least some form of computer-readable media. ... [C]omputer-readable media might comprise computer storage media and communication media.

(Fontana ¶0072.) Fontana further explains that this computer readable media can contain “computer readable **instructions**,” and can be implemented in forms of memory such as “RAM, ROM, EEPROM, flash memory or other memory technology,” twchich a POSA would have recognized as **non-transitory computer readable media**. (Fontana ¶0078; Houh ¶177.)

(c) “**converting audio associated with the video to text; ”**
(Claim 17[c])

Fontana discloses **converting audio associated with the video to text** for the same reasons discussed above for claim 1[d]. (Houh ¶178.)

(d) “**converting images associated with the video to video data; ”** (Claim 17[d])

Fontana discloses **converting images associated with the video to video data** for the same reasons discussed above for claim 1[e]. (Houh ¶179.)

(e) “**generating the video data by segmenting image files of the video before processing the image files in parallel; ”**
(Claim 17[e])

Fontana discloses **generating video data** for the reasons discussed in claim 1[e], and **processing the image files in parallel**, for the reasons discussed for claim 1[b]. For the same reasons as discussed above for claim 6, Fontana further discloses **segmenting image files of the video before processing the image files in parallel**. (Houh ¶180.)

(f) “**identifying one or more objects in the image files; ”**
(Claim 17[f])

For the reasons discussed above for claim 1[b], Fontana discloses **identifying one or more objects in the image files.** (Houh ¶181.)

(g) “**generating data topics, from the text and the video data, that describe content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio; ”** (Claim 17[g])

For the same reasons discussed above for claim 1[f], Fontana discloses claim 17[g]. As explained for claim 1[f], a POSA would have appreciated that the various metadata generated by Fontana correspond to the “topical meta-data that describes content of the video” within claim 1 of the ’972 patent. A POSA would also recognize that Fontana’s generated metadata also correspond to “**data topics ... that describe content of the video.**” For example, Fontana’s object metadata identifying objects are “data topics,” as are keywords or other terms derived from the text metadata. (Houh ¶182.)

(h) “**adding the data topics to the video as meta-data; ”**
(Claim 17[h])

For the same reasons discussed for claim 1[g], Fontana discloses **adding the data topics to the video as meta-data**, including by linking the data topics (i.e., Fontana’s text **metadata** and object **metadata**) to specific times in the video,

synchronizing the metadata for simultaneous display to the user, and storing the video with its metadata. (Houh ¶183.)

- (i) **“cross-referencing the text, the video data, and the topics with the video based on the generated data topics; ” (Claim 17[i])**

It would have been obvious to combine Fontana with Lau to **cross-referenc[e] the text, the video data, and the topics with the video based on the generated data topics** for substantially the same reasons explained above for claim 1[h]. The combination of Fontana and Lau and motivation to combine are the same as for claim 1. (Houh ¶¶184-188.)

Minor differences in wording between claim 1[h] and 17[i] do not materially alter the analysis. What claim 1[h] refers to as “topical meta-data” corresponds to the **topics** in claim 17[i], such that for the same reasons discussed for claim 1[h] above, the combination of Fontana with Lau **cross-referenc[es] the text, the video data, and the topics**. (Houh ¶¶185-186.) Furthermore, this occurs **“with the video.”** Specifically, Lau’s search to match ad units to the content identifies a set of candidate locations within the video, then scores each based on the strength of the match. (Lau ¶¶0057–0059.) Thus, the same combination described for claim 1[h] **cross-referenc[es] the text, the video data, and the topics with the video**. (Houh ¶187.) Finally, claim 1[h]’s cross-referencing is “based on the generated topical meta-data to determine topics” whereas claim 17[i]’s cross-referencing is **“based on**

the generated data topics.” Here, the topical meta-data for claim 1[h] corresponds to the data topics for claim 17[i], such that this element of claim 17[i] is obvious over Fontana with Lau for the same reasons as claim 1[h]. (Houh ¶188.)

(j) “generating a text, image, or animation based on the data topics; and” (Claim 17[j])

Claim 17[j] essentially combines the end of claim 1[h] (“to determine topics”), claim 1[i], and claim 1[j] into a single limitation – that the system **generat[e] a text, image, or animation based on the data topics.** For the same reasons discussed above for claims 1[h], 1[i], and 1[j], claim 17[j] would have been obvious over the combination of Fontana with Lau. (Houh ¶189.)

(k) “placing the text, image, or animation in the video.” (Claim 17[k])

For the same reasons discussed above for claim 1[k], Fontana discloses **placing the text, image, or animation in the video.** (Houh ¶190.)

The combination of Fontana and Lau therefore renders claim 17 obvious.

18. Claim 18: “The system according to claim 17, wherein converting the audio comprises natural language processing.”

The additional limitations of claim 18 are obvious over Fontana in view of Lau for the same reason as discussed above for claim 3. (Houh ¶191.)

- 19. Claim 19: “The system according to claim 17, the computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel.”**

The additional limitations of claim 19 are disclosed by Fontana. (Houh ¶¶192-195.) For the same reason as discussed above for claim 1[b], Fontana discloses that **the computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel.** (Houh ¶192.)

As also explained for claim 1[b], Fontana teaches a “distributed computing network” that performs parallel processing. (*See, e.g.*, Fontana ¶¶0052, 0053; Houh ¶193.) Fontana illustrates this system in Figure 3:

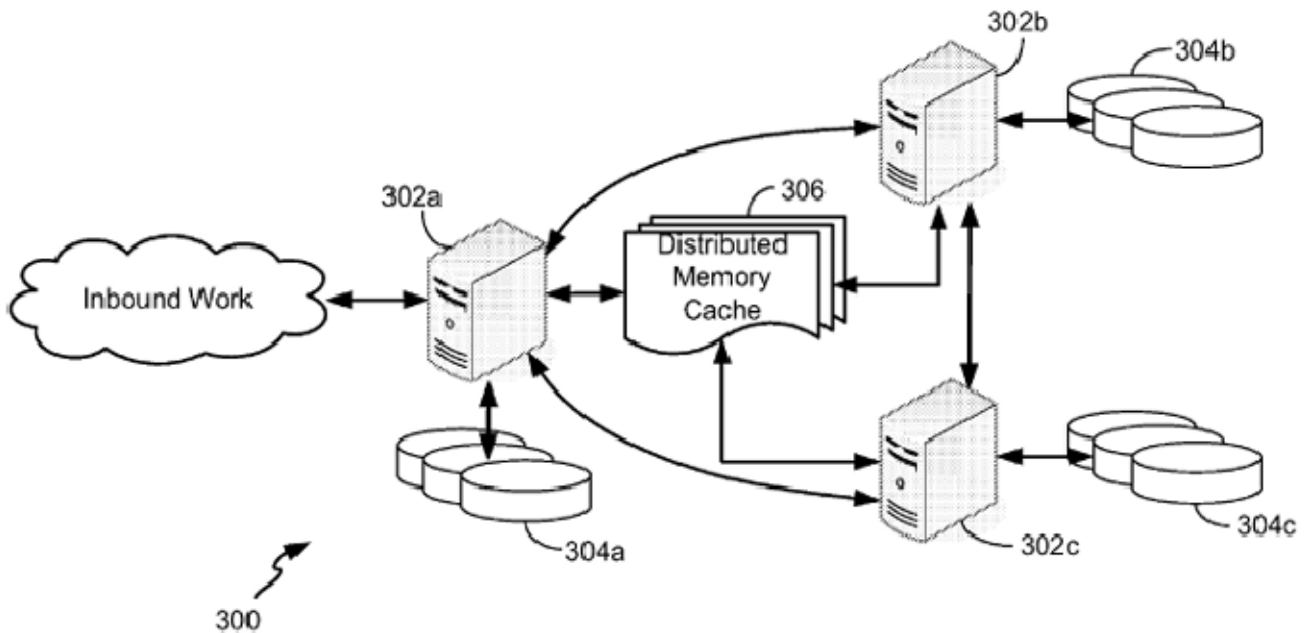


FIG. 3

A POSA would have appreciated that each of the servers, such as servers 302a-c in Figure 3, represents **at least one node of a cluster**, such that the distributed computing system of Fontana directs processing tasks to at least one node of a cluster. (Houh ¶¶193-194.)

Fontana further explains that this distributed computing system and network of servers “allow for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing and resulting usability for content providers.” (Fontana ¶0135.) From this, a POSA would have appreciated that audio and images would be separated and processed in parallel, such that Fontana discloses that **the**

computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel. (Houh ¶195.)

20. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”

For the same reason as discussed above for claim 9, the additional limitations of claim 20 are obvious over Fontana. (Houh ¶196.)

D. Ground 2: Claims 8-9 and 20 Are Obvious Over Fontana in view of Lau and Arakawa

1. Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”

Claim 1 is obvious over Fontana in view of Lau for the same reasons as in Ground 1. The additional limitations of claim 8 are obvious over Arakawa. Specifically, a POSA would have found it obvious to combine Fontana with Arakawa to “**segment[] the audio files before processing the audio files in parallel**” in order to take advantage of Fontana’s teaching that multiple servers can be used to efficiently carry out tasks like audio processing while also increasing efficiency by processing voice clips separately from background noise. (Houh ¶¶197-205.)

Arakawa teaches **segmenting audio** by frame, and further segmenting audio files into voice sections (corresponding to speech) versus non-voice sections

corresponding to other noise, prior to performing speech recognition on the voice sections. The result is “a voice recognition system capable of, while suppressing negative influences from sound not to be recognized, correctly estimating utterance sections that are to be recognized.” (Arakawa, Abstract.) Arakawa’s segmentation into voice versus non-voice sections is performed by comparing each frame to a threshold value, such as Arakawa explains:

The voice segmentation unit 103 calculates a voice segmentation feature value which indicates possibility of being voice for each frame input sound data. Then, the voice segmentation unit 103 classifies each frame into a voice frame or a non-voice frame by comparing a threshold value (hereinafter, it is referred to as threshold θ) and a voice segmentation feature value for each frame. If a calculated voice feature value for a frame is larger than the threshold value θ , the frame is classified as a voice frame. If a calculated voice feature value is less than the threshold value θ , the frame is classified as a non-voice frame. Then, the voice segmentation unit 103 merges connected voice frames classified above into a voice section (hereinafter, referred to as a first voice section). As a voice segmentation feature value, amplitude power for each frame can be used, for example. However, the voice segmentation feature value which indicates possibility of being voice is not limited to amplitude power.

(Arakawa ¶0043.) Arakawa next teaches calculating “a feature value used for voice recognition” for each audio frame, such as, for example, “cepstrum feature or its

derivative feature.” (Arakawa ¶0044.)

Arakawa further teaches using the feature value for each frame to identify words and/or phonemes for speech recognition, as well as to update the threshold for distinguishing voice from non-voice:

The searching unit 108 calculates, based on the voice recognition feature value, a likelihood of voice and a likelihood of non-voice for each frame, and searches for a word sequence using these likelihoods and the above-mentioned models. ...

Also, the searching unit 108 segments a section to be the target of the voice recognition (hereinafter, it is referred to as a second voice section) based on the likelihood of voice and the likelihood of non-voice that have been calculated. ...

Thus, the searching unit 108 obtains the word sequence corresponding to the input sound (a recognition result) using the feature value for each frame, the vocabulary/phoneme model and the non-voice model, and, in addition to that, obtains the second voice section....

According to a difference between length of the first voice section and length of the second voice section, the parameter updating unit 109 updates the threshold value θ . That is, the parameter updating unit 109 compares the first voice section and the second voice section, and updates the threshold value θ to be used by the voice segmentation unit 103.

(Arakawa ¶¶0047–0050.)

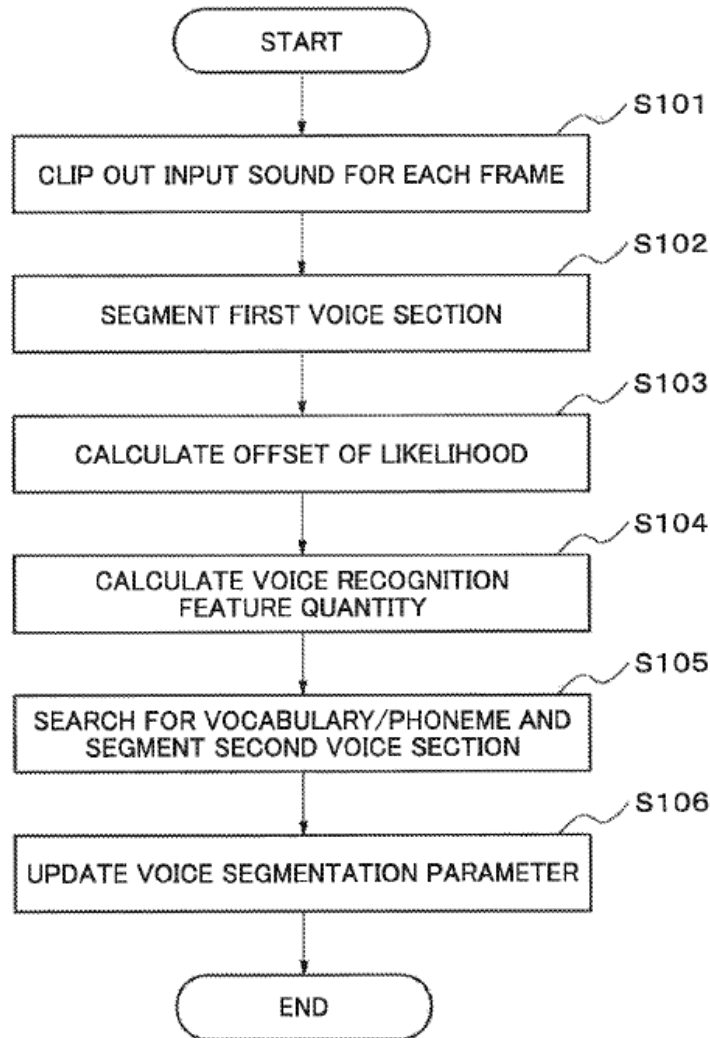
Arakawa thus teaches at least three segmentations of the audio files: the first

segmentation of voice versus non-voice sections, the second segmentation of voice segments based on phoneme recognition, and the frame-by-frame segmentation for purposes of voice segmentation and voice recognition. (Houh ¶¶198-201.)

It would have been obvious to a POSA to combine Arakawa's method of performing voice recognition using segmentation with Fontana to take advantage of Arakawa's improved speech recognition in the presence of background noise. (*See, e.g.*, Arakawa Abstract, ¶0001 (“The present invention relates to a voice recognition system, a voice recognition method and a voice recognition program which recognize voices in an environment where background noise exists.”); *see also* Arakawa ¶0012; Houh ¶202.)

Furthermore, a POSA combining Fontana with Lau would have found it obvious to segment the audio files by frame and into voice sections according to Arakawa's “voice segmentation unit” **before processing the audio files in parallel.** (Houh ¶203.) Arakawa explains that dividing the audio into frames, and segmenting voice versus non-voice sections using the voice segmentation unit, each occur prior to performing the speech recognition functions described in Arakawa. This is illustrated in Figure 2, which shows that “clip out input sound for each frame” and “segment first voice section” each occur before the speech recognition functions (e.g., “search for vocabulary/phoneme”):

Fig.2



(See Arakawa Fig. 2.)

A POSA would have found it obvious that Fontana's distributed computing system could separate the resulting audio segments across multiple servers to process in parallel for speech recognition—either by having each frame processed

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

in parallel by the “voice recognition feature value calculating unit” and “searching unit” of Arakawa, or by having each voice versus non-voice section of Arakawa processed in parallel. (Houh ¶204.) As also discussed for Ground 1 Claim 8, Fontana expressly discloses that processes such as audio processing could be separated and performed in parallel in this fashion—its servers, which include “audio algorithms,” “are configured to share processing jobs, such that tasks can be ... separated and performed across multiple computing systems in parallel.” (Fontana ¶0053.)

Rationale and motivation to combine (Fontana and Lau with Arakawa):

A POSA would have been motivated to combine Fontana and Lau with Arakawa to take advantage of Arakawa’s benefits of improving speech recognition in the presence of background noise, as discussed above. (Houh ¶205.) Additionally, Fontana explains that “a plurality of different speech to text algorithms can be applied.” (Fontana ¶0093.) As such, a POSA would have been motivated to identify audio processing algorithms, such as that disclosed by Arakawa, that provided additional benefits. (Houh ¶205.) Arakawa is an analogous reference to the ’972 patent. Arakawa, like the ’972 patent, addresses audio processing using segmentation. (’972, 1:49-53 (“The text from the audio can be generated by first segmenting images of the audio, and then converting the segments of images to text in parallel. The audio can be segmented at spectrum thresholds.”)); Arakawa,

Abstract (“Provided is a voice recognition system capable of, while suppressing negative influences from sound not to be recognized, correctly estimating utterance sections that are to be recognized. A voice segmenting means calculates voice feature values, and segments voice sections or non-voice sections by comparing the voice feature values with a threshold value.”).) A POSA would have had a reasonable expectation of success in deploying Arakawa’s audio processing algorithms within the context of Fontana’s distributed computing for multimedia processing. (Houh ¶205.) Indeed, as noted, Fontana expressly teaches that “a plurality of different speech to text algorithms can be applied” (Fontana ¶0093) and that audio processing can be done in parallel using a plurality of servers (Fontana ¶0053).

2. Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum thresholds.”

A POSA would have found claim 7 obvious over the combination of Fontana with Lau, as discussed above for Ground 1. Additionally, as discussed above for Ground 2 Claim 8, Arakawa discloses segmenting audio files **at spectrum thresholds** to distinguish voice from non-voice sections of an audio file. It would have been obvious to combine Fontana with Arakawa to **segment[] the audio files and the image files at spectrum thresholds**. (Houh ¶¶206-208.)

A POSA would have been motivated to segment the image files in the same

Petition for *Inter Partes* Review of
U.S. Patent No. 9,940,972 B2

locations as the audio files based on Arakawa's thresholds for the voice segmentation unit. (Houh ¶¶207-208.) Specifically, Fontana explains that it may be desirable to generate thumbnails (i.e., segment the image files) based on sound changes in the multimedia content: "The thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest (for example a first frame, a last frame, and immediately following major scene or sound changes in the content)." (Fontana ¶0095.)

A POSA would have found it obvious that Arakawa's voice segmentation unit identifies "sound changes" based on its use of a **spectrum threshold**. (See Arakawa ¶0043 ("If a calculated voice feature value for a frame is larger than the threshold value θ , the frame is classified as a voice frame. If a calculated voice feature value is less than the threshold value θ , the frame is classified as a non-voice frame. Then, the voice segmentation unit 103 merges connected voice frames classified above into a voice section (hereinafter, referred to as a first voice section).").) Thus, a POSA would have appreciated that the same segmentation into voice sections versus non-voice sections performed for audio files in Arakawa could be used for generating the thumbnail images to segment the image files in Fontana. (Houh ¶208.)

3. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”

A POSA would have found claim 17 obvious over the combination of Fontana with Lau, as discussed above for Ground 1. Additionally, for the same reason as discussed above for claim 9, it would have been obvious to a POSA to combine with Arakawa to **segment the audio and the images at spectrum thresholds.** (Houh ¶209.)

VII. CONCLUSION

Petitioner respectfully requests IPR institution.

Dated: May 10, 2025

Respectfully submitted,

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue NW
Suite 700
Washington, DC 20004
Tel: (650) 843-5001
Fax: (650) 849-7400

By: / Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673
Counsel for Petitioner

CERTIFICATE OF COMPLIANCE WITH WORD COUNT

Pursuant to 37 C.F.R. § 42.24(d), I certify that this petition complies with the type-volume limits of 37 C.F.R. § 42.24(a)(1)(i) because it contains 13,502 words, according to the word-processing system used to prepare this petition, excluding the parts of this petition that are exempted by 37 C.F.R. § 42.24(a) (including the table of contents, a table of authorities, mandatory notices, a certificate of service or this certificate word count, appendix of exhibits, and claim listings).

DATED: May 10, 2025

COOLEY LLP
ATTN: Patent Docketing
1299 Pennsylvania Avenue NW
Suite 700
Washington, D.C. 20004
Tel: (650) 843-5001
Fax: (650) 849-7400

/ Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673

CERTIFICATE OF SERVICE

I hereby certify, pursuant to 37 C.F.R. Sections 42.6 and 42.105, that a complete copy of the attached **PETITION FOR INTER PARTES REVIEW OF U.S. PATENT NO. 9,940,972 B2**, including all exhibits (**Nos. 1001-1013**) and related documents, are being served via Federal Express on the 10th day of May, 2025, the same day as the filing of the above-identified document in the United States Patent and Trademark Office/Patent Trial and Appeal Board, upon Patent Owner by serving the correspondence address of record with the USPTO as follows:

27890 - STEPTOE LLP/DC
1330 CONNECTICUT AVENUE, N.W.
WASHINGTON, DC 20036

And, via Federal Express upon counsel of record for Patent Owner in the litigation pending before the U.S. District Court for the Northern District of California entitled *Cellular South, Inc. v. Google LLC*, Case No. 4:25-cv-01487-YGR (N.D. Cal.) as follows:

Robert Sean Hill
Holland & Knight LLP
1722 Routh Street
Suite 1500
Dallas, TX 75201
214-964-9421
Fax: 214-964-9501
Email: robert.hill@hklaw.com

DATED: May 10, 2025

/ Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673
COOLEY LLP
1299 Pennsylvania Ave. NW,
Suite 700
Washington, D.C. 20004
Tel: (650) 843-5001