

Case No. IPR2023-1024  
Patent No. 11,290,836 B2

UNITED STATES PATENT AND TRADEMARK OFFICE

---

BEFORE THE PATENT TRIAL AND APPEAL BOARD

---

META PLATFORMS, INC.,  
Petitioner,

v.

EIGHT KHZ, LLC,  
Patent Owner.

---

Case IPR2023-01024  
Patent 11,290,836 B2

---

**DECLARATION OF DURAND R. BEGAULT, PH.D.  
IN SUPPORT OF PATENT OWNER'S PRELIMINARY RESPONSE**

## TABLE OF CONTENTS

|                                                                    | Page |
|--------------------------------------------------------------------|------|
| I. INTRODUCTION .....                                              | 1    |
| II. QUALIFICATIONS .....                                           | 1    |
| III. SUMMARY OF OPINIONS .....                                     | 3    |
| IV. RELEVANT LEVEL OF ORDINARY SKILL IN THE ART .....              | 3    |
| V. The '836 Patent .....                                           | 3    |
| VI. GROUND 1 OF THE PETITION: FINNEGAN AND BEGAULT.....            | 11   |
| A. Finnegan .....                                                  | 12   |
| B. Begault.....                                                    | 16   |
| VII. GROUND 2 OF THE PETITION: NORRIS, BEGAULT, AND<br>ALGAZI..... | 17   |
| A. Norris .....                                                    | 18   |
| B. Algazi.....                                                     | 19   |
| VIII. GROUND 3 OF THE PETITION: SHIVAPPA AND ZHOU .....            | 19   |
| A. Shivappa .....                                                  | 20   |
| B. Zhou.....                                                       | 21   |
| IX. CONCLUSION.....                                                | 22   |

I, Durand R. Begault, declare as follows:

**I. INTRODUCTION**

1. I am currently employed as Director, Audio Forensic Center, a wholly-owned subsidiary of Charles M. Salter Associates, an acoustical consulting firm located in San Francisco, California. I am also currently employed as a research scientist in the Human Systems Integration Division of NASA Ames Research Center, Moffett Field, California.
2. I have been retained as an independent expert consultant in the *inter partes* review (“IPR”) proceeding involving the 11,290,836 Patent by Eight kHz LLC (“8kHz”). I will refer to this patent using common shorthand. I have been asked to consider the arguments raised by the Petition regarding the validity of the ‘836 Patent in this proceeding.

**II. QUALIFICATIONS**

3. I am the director for Audio Forensic Center, where I perform investigations with regards to acoustics; video and audio evidence recordings; audio patent infringement analysis; acoustic measurement; speech intelligibility and warning signal audibility; human factors and psychoacoustics; musicology and copyright infringement; voice comparison; and recorded gunshot analysis. I also serve as a research scientist for NASA Ames Research Center specializing in

psychoacoustics, human factors, multimodal systems, and communications engineering research, applied to aerospace safety, virtual reality, and speech communications. I obtained my Ph.D. in 1987 from the University of California, San Diego with a focus in Computer Audio, and was made a fellow of the Audio Engineering Society in 2002 “for contributions to understanding of spatial hearing”.

4. For over 30 years I have conducted analyses and testified in cases for attorneys, law enforcement agencies, and corporate internal investigations. I have qualified as an expert in numerous State and U.S. Federal Courts; the United States International Trade Commission (USITC); and courts in Canada and international arbitration panels. I have over 100 publications in the area of acoustics, psychoacoustics and forensic audio, and have testified over 85 times in court or deposition, including in the area of digital audio recording authentication and metadata (e.g., the International Centre for Settlement of Investment Disputes, The World Bank, Washington, D.C.; California Superior Courts; United States District Court). I have served as an adjunct faculty member of the Sound Recording Department at McGill University in Montreal, Canada, and have received a Fellowship Award from the Audio Engineering Society. I

have also served on numerous journal review boards, technical committees and Ph.D. dissertation committees in the area of acoustics and audio engineering. Exhibit 2007, attached to the Patent Owner's Preliminary Response is a true and correct copy of my CV.

### **III. SUMMARY OF OPINIONS**

5. My opinions in this Declaration are based on my review of the '836 Patent, its prosecution history before the U.S. Patent and Trademark Office, the IPR Petition concerning this patent, and references cited therein in support of the Petition. I draw upon my experience and knowledge of acoustical engineering, psychoacoustics, signal processing, human-machine interaction, and virtual reality systems.

### **IV. RELEVANT LEVEL OF ORDINARY SKILL IN THE ART**

6. In my opinion, as relates to the '836 Patent, a person having ordinary skill in the relevant art (a PHOSITA) would have several years of academic and/or professional experience in a technical field relevant to digital audio signal processing, such as electrical engineering, computer engineering, or computer science; and a familiarity with relevant literature and concepts concerning virtual "3D" audio processing, which were widely available at the time of disclosure.

### **V. The '836 Patent**

7. Through various embodiments, the '836 Patent discloses a method (or a system) that aims to solve “[t]he problem of aligning a binaural sound with a visible source of the sound ... when the source of the sound is within a near-field distance of the listener.” ‘836, 3:18-26. This addresses in particular the situation where a device such as a PDA or controller is located within the visual field, on or near the frontal median plane that extends directly in front of a listener. A practical example is the visual and auditory location of a caller in a “telecall” communication scenario, which requires virtual simulation of a sound source located directly in front of a listener. Such virtual simulations are typically located within the “near-field” of a listener (50 cm–1 m), corresponding to real-world conversational distance.
8. In both real-world and virtual spatial hearing, users who rely entirely on auditory cues to estimate the distance of a sound source typically underestimate the distance of the sound source from themselves (egocentric center), particularly for sources located in front of a listener. When visual cues are present in real-world communication, visual location dominates over acoustic cues, a phenomenon known as “visual capture.” However, visual capture does not work similarly in virtual simulations due to a comparative paucity of visual and proprioceptive

information, and practical limitations on creating virtual visual imagery as realistic and robust as experienced in the real-world. Hence, there is a potential conflict between visual and auditory cue location in virtual simulations, caused by underestimation of the auditory distance. This underestimation can dramatically impact user experience or the potential quality of a telecall or a virtual simulation of a visual object. To provide a more realistic user experience, it is desirable to match or align the perceived locations of sound and image within a virtual simulation.

9. When a real-world target sound source is located in front of a listener within a near-field distance, psychoacoustic research has shown that both azimuth and distance localization errors are increased, relative to lateral or far-field sound source positions (Brungart, 1998). Ex. 2013, pp. 218-219. To mitigate these distance and azimuth errors, researchers have proposed a technique of using a particular set of near-field head-related transfer functions (“HRTFs”) “tailored” to one or more sound source distances (e.g., Arend, 2016). *See generally* Ex. 2011, 1, 6-8.
10. Another effect observed in HRTF-based virtual audio simulation particular to frontally-located sound sources is an overall perceived compression of distance relative to side positions, and in some cases

“inside the head localization”, the opposite of “externalized” sound source location outside the head. Relative to non-frontal positions, the left and right ear HRTFs for frontal sound source positions have maximal similarity both in interaural time and amplitude differences and in overall spectra. This minimization of interaural differences causes the sound to be minimally differentiated at the two ears, similar to a “monaural” single channel of sound produced through stereo headphones. In contrast to externalized lateral virtual images, where the auditory level, time and spectral cues are different, a significant challenge noted throughout the literature for virtual sound localization has been to create imagery that is located at an externalized position **in front** of a listener, instead of inside the head or with compressed distance. *See* Figures 1 and 2 (from Begault, 2000) showing the compression of distance from the front, and in some cases, inside the head or “verged cranial” (edge of the head) localized positions, that are most significant for frontal positions. Ex. 1006, 63, 81.

FIGURE 1

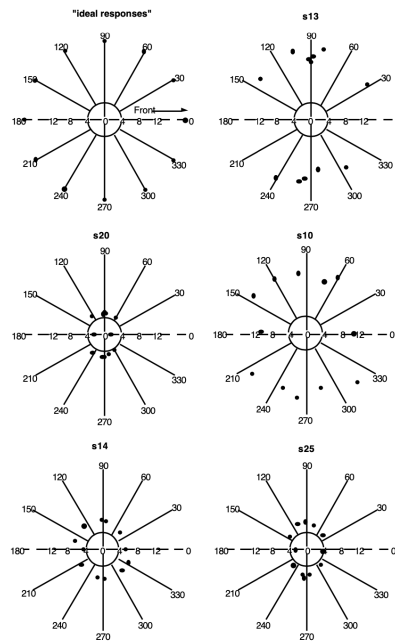


Figure 2.30. Azimuth and distance judgments for speech stimuli (Begault and Wenzel, 1993; see text). Upper left picture shows “ideal responses.” *Reproduced with permission from Human Factors, Vol. 35, No. 2, 1993. Copyright 1993 by the Human Factors and Ergonomics Society. All rights reserved.*

FIGURE 2

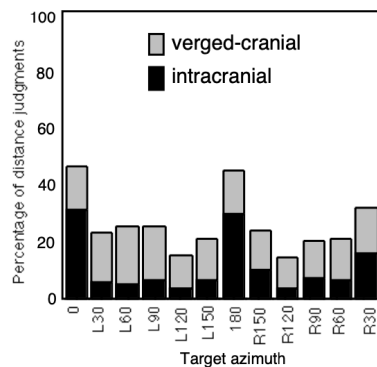


Figure 3.10. IHL speech stimuli as a function of azimuth; results from 11 subjects (Begault and Wenzel, 1993). *Reproduced with permission from Human Factors, Vol. 35, No. 2, 1993. Copyright 1993 by the Human Factors and Ergonomics Society. All rights reserved.*

11. To my knowledge, deterministically simulating a sound source located  
in the near-field by convolution with far-field HRTFs within an

auditory display has not been proposed or tested prior to the '836 Patent. Given this lack of disclosure in the literature, it is my opinion that it would not have been obvious to a PHOSITA that using far-field HRTFs to simulate near-field sounds would be a solution to improve the spatial coherence between frontally-located visual and auditory virtual sources, by “compensating” for the distance compression factors described in the previous paragraphs.

12. I understand that Petitioner challenges independent claims 8 and 15 of the '836 Patent. Both independent claims recite the following limitation: “an image that occurs within a near-field distance from a head of a user.” Both claims further recite “generate binaural sound that has a sound localization point (SLP) with a coordinate location that occurs at a far-field distance located behind the image and along a line-of-sight from the user.” Lastly, both claims disclose that generating the binaural sound require “process[ing] sound for the image” with HRTFs.
13. A clarification needs to be made at the outset regarding what exactly is “an image that occurs within a near-field distance from a head of a user,” and if the location of this image is necessarily the same as the location of the claimed electronic device (or the display). This is critical because it allows evaluation of whether or not the invalidity arguments

raised by Petitioner are germane to the claimed subject matter. Petitioner does not clarify the meaning of this limitation in view of the specification—nor does Petitioner propose a construction necessarily needed to be able to best understand its invalidity arguments. On one hand, Petitioner argues that “the electronic device and the displayed image are co-located.” Petition, 9-10, 28-29, 52-53. On the other hand, Petitioner argues that because the location of an image is “the location of a perceived image,” the image and the display are not co-located. Petition, 10. Nowhere in the Petition, does Petitioner attempt to resolve this inconsistency based of the specification of the ‘836 Patent.

14. Through various embodiments and disclosures, the specification of the ‘836 Patent makes clear that a virtual image that occurs “within a near-field” is not necessarily co-located with the electronic device (or its display) that is located “at the near-field.” The claim speaks in terms of an image that “occurs” within a distance from a user—as opposed to at the display’s location. The use of the term “occur” in this context clearly indicates the user’s perception of the location of the virtual image. As such, it is immaterial where the display is located—what is important is location of the virtual image perceived by the user. A PHOSITA would understand this by reading the specification. For

example, the specification discloses “a location where the sound source is visually perceived” (‘836, 3:36-37) or “percept of the image” (‘836, 19:3). The specification also discloses that “[a]s another example, the listener wears a HPED (e.g., a smartphone) or wearable electronic device (e.g., a HMD) that displays an AR or VR<sup>1</sup> image of the sound source rendered to be perceived at a particular distance with respect to an origin, such as a location of the wearer.” ‘836, 26:66-27:3.

15. In fact, the ‘836 Patent addresses a problem specific to virtual image perception. “Another problem with electronically generated binaural sound occurs when the perceived location of where the sound originates does not match a location of an image of the sound.” ‘836, 2:41-44. The solution of the ‘836 Patent is to make “[b]inaural sound or 3D audio ... more realistic” in its combination with virtual visual images by aligning “a perceived location of where the sound originates ... with a visible source of the sound, such as an image or object.” ‘836, 3:12-17. The specification further discusses the possibility of convolving “... sound to a location where the sound source is visually perceived.” ‘836, 3:35-37. Moreover, in describing FIGS. 6A and 6B, the ‘836 Patent discloses

---

<sup>1</sup> AR= augmented reality; VR= virtual reality

that the images are to be located at different perceived distances from the user:

The listener, however, perceives the images at different distances with respect to the listener. Images 620A and 620B appear to be located at the surface of the display 630. Image 620C, however, appears to be located farther away than a wall or virtual wall 622 at which the display 630 is located.

‘836, 22:11-15.

16. All these examples clearly communicate to a PHOSITA that the distances from the user of the near-field image and of the claimed electronic device (or the claimed display) are not co-located. The claimed device allows a user to perceive and estimate the distance between the virtual image and the user’s egocentric center in an improved way in a virtual display. A PHOSITA would appreciate that the claimed display provides the user with means for improved co-located virtual and audio imagery.

## **VI. GROUND 1 OF THE PETITION: FINNEGAN AND BEGAULT**

17. I understand that Petitioner relies on Daniel J. Finnegan et al., *Compensating for Distance Compression in Audiovisual Virtual Environments Using Incongruence*, Chi’ 16: Proc. of the 2016 CHI

Conf. on Hum. Factors in Computing Sys. (2016) (“Finnegan”) and Durand R. Begault, *3-D Sound for Virtual Reality and Multimedia*, NASA/TM—2000-209606 (Aug. 2000) (“Begault”) to allege that these references teach or suggest every limitation of independent claims. I disagree, and for the reasons discussed below, I am of the opinion that Petitioner mischaracterized teachings of Finnegan and Begault.

**A. Finnegan**

18. Finnegan provides conclusions that in fact teach away from using a virtual simulation of sound at a location at a greater distance than a visual source for near-field stimuli. Finnegan hypothesized that an “incongruity function” would yield more accurate estimate of distance (i.e., perception of distance by an observer): “Rendering the audio at an incongruent position further from the observer than the visual stimulus will result in more accurate distance perception.” Ex. 1005, 6. Finnegan asked subjects to judge the perceived relationship between a virtual visual stimulus and an auditory stimulus for sound sources “close to the observer” defined as 3 +/- 2.5 meters. Ex. 1005, 10. In fact, their results showed that “incongruent presentation did not reduce distance estimates when the targets were close to the observer” (Ex. 1005, 10) and therefore rejected their incongruency hypothesis for near range

sounds (“Thus, H1 is supported by the results of our analysis in the far range but not in the near range”). Ex. 1005, 9. Therefore, a PHOSITA would not think the concept of the ‘836 Patent would be workable as applied to “near-field” ( $< 1$  m) visual stimuli if they understood Finnegan.

19. As described in Begault (2000) and exemplified in later papers such as Huang et al. (2021) and by many others, a PHOSITA would know to use reverberation techniques as a proven solution to help externalize sound and improve distance estimation for sound located in the near-field. Begault (2000) cites multiple studies confirming this: “Researchers have almost consistently found that reverberation, either natural or artificial, enhances the externalization of 3-D headphone sound (Plenge, 1974; Toole, 1970; Sakamoto, Gotoh, and Kimura; 1976; Begault, 1992a).” Ex. 1006, 81; *see also* Figure 3 reproduced below from Begault (2000) (Ex. 1006, 81). Therefore, and consistent with the common understanding regarding the externalization advantages of reverberation, a PHOSITA would interpret Finnegan as teaching away from using “incongruity function” for sources located in the near-field. For example, Huang et al (2021), cited Finnegan to conclude that “the incongruence of audio and visual object distances

can compensate distance compression only when the object is at a longer range [i.e., distances greater than 5 m from the listener].” Ex. 2010, 4.

FIGURE 3.

Table 3.2. Data summarizing IHL judgments of virtual speech, for five subjects under dry and reverberant conditions (adapted from Begault, 1992a). Columns 2 and 3 show percentages of all judgments by each subject that were unexternalized—heard less than or equal to 4 inches distance. The stars indicate a very high (> 99%) probability that the difference between anechoic and reverberant listening was significant.

| Subject Number | IHL Anechoic | IHL Reverberant |
|----------------|--------------|-----------------|
| 1              | 46%          | 10.5% ***       |
| 2              | 59.5%        | 2% ***          |
| 3              | 3.5%         | 0.5%            |
| 4              | -0-          | 0.5%            |
| 5              | 15.5%        | -0- ***         |
| Mean           | 25%          | 2.7%            |

20. Until the late 1990s, near-field HRTFs were not commonly available for use in virtual acoustic rendering systems. Arend wrote in 2016 that “set of high resolution near-field HRTFs that is publicly available could be used for further experiments regarding auditory localization of nearby sound sources” (Ex. 2011, 2) and made such near-field measurements between 0.25-1.5 m from a dummy head publicly available. Arend unequivocally explains that “[p]roper auralization of nearby sound sources requires near-field HRTFs.” Ex. 2011, 8. Thus, a PHOSITA reading Arend and other works that pre-date the priority date of the ‘836 Patent used such near-field measurements in an auditory

display only to understand the problems associated with near-field frontal imagery from the perspective of spatial hearing research.

21. Regarding the Soundscape Renderer described in Geier et al. (2008) and utilized by Arend (2016) (Ex. 2011, 6), the Soundscape Renderer manual (Ex. 2014, section 5.2, p. 32) clearly indicates that a single set of HRTFs are used in any single instance of rendering, and that a user may either use a single default set of measurements or load their own single set of HRTF measurements, as done by Arend (2016) separately for each of the different near-field distances they evaluated. The Soundscape renderer itself does not perform a deterministic transition between two or more HRTF sets based on different distances to a listener; it only works with a single HRTF set at a time that is specified by the user.
22. The Soundscape Renderer uses the same HRTF set specified by the user for any and all source distances, modified by a gain factor. Note there is no equivalence between actual near-field and a gain-adjusted far-field HRTF measurement, as noted in Brungart (1998) (Ex. 2013, 30-31). “At distances greater than 1 m (the “far-field”) the HRTF is roughly independent of distance, while at distances less than 1 m the HRTF changes significantly with distance” (Ex. 2013, 30-31).

23. Therefore a PHOSITA familiar with Arend (2016) and the Soundscape Renderer would be motivated to use one or more separate near-field HRTF sets to simulate sources in the near field. A PHOSITA would recognize the change in near-field HRTF spectra as a function of distances less than 1 m from reading both Arend (2016) and Brungart (1998), and would not be motivated to use far-field HRTFs for near-field source locations.

**B. Begault**

24. As its author, I can state that Petitioner mischaracterizes teachings of Begault (2000). Petitioner states “the use of HRTF spectral shaping is typically featured as the key component of a 3-D sound system” (quoting Ex. 1006, 41). While this is in itself true, nowhere in Begault (2000) is there a disclosure regarding near-field HRTFs, nor is there any discussion on how to move a virtual sound source located in a near-field to a far-field location behind the virtual image source in the near-field. Nowhere does Begault (2000) discuss near-field HRTF measurements or a distinction of near-field versus far-field HRTF measurements. Further, Petitioner cites Begault (2000) to argue that a database of HRTFs known as KEMAR has been by researchers; however, Begault (2000) clearly states that the publicly available

HRTFs of KEMAR from Gardner (Ex. 1006, 207) were measured at 1.4 m, and not in the near-field. There is no suggestion in Begault (2000) of using near-field measurements of KEMAR's HRTFs, or using far-field HRTFs of KEMAR for near-field sources.

25. In fact, Begault expressly teaches away from doing so: “Changing HRTF spectra as a function of head movement is generally considered to improve overall localization performance. It does not necessarily follow that virtual source movement will improve performance.” Ex. 1006, 99. Given this, a POSITA would not think the use of far-field HRTFs to move a near-field virtual sound source to a far-field location behind the virtual image source (i.e., incongruent placement) would be workable.

**VII. GROUND 2 OF THE PETITION: NORRIS, BEGAULT, AND ALGAZI**

26. I understand that Petitioner relies on U.S. Patent No. 9,226,090 (“Norris”), Begault (2000), and V.R. Algazi et al., *The CIPIC HRTF Database*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2001) (“Algazi”) to allege that these references teach or suggest every limitation of independent claims. I disagree, and for the reasons discussed below, I am of the opinion that Petitioner mischaracterized teachings of these references.

**A. Norris**

27. Norris is directed at a method to “localize the sound to determine a distance and a direction of the sound.” Ex. 1007, 3:37-38. Norris does not disclose using visual stimulus to determine the distance between the user and the source of sound. To the extent that Norris discloses image or video, all such disclosures involve displaying a 2-D images on a conventional screen worn by the user (e.g., “video of a friend appears on a display” during a call (Ex. 1007, 7:17-19)). In other words, Norris’s display (or WED) only provides audio cues used for depth perception. Because Norris does not disclose use of visual cues for depth perception, Norris does not disclose “an image that occurs within a near-field” claimed by the ‘836 Patent. To the extent Norris uses objects or real-world images of them located away from the user to localize the sound, the sound being localized is processed to the location of the objects not behind them. For example, Figure 20 in Norris localizes the sound at the actual physical object (i.e., a chair). Norris discloses no audiovisual stimulus for which the visual component is perceived by the listener to be within a near-field. For example, when Norris describes the example of localizing a bird

chirping recording, Norris does not disclose any image associated with the recording.

**B. Algazi**

28. Algazi's paper is completely concerned with HRTF measurements from an acoustical perspective, and not concerned with questions of auditory perception. It does not in any way follow that Algazi's publicly available HRTF set existing prior to the disclosure of '836 patent would cause a PHOSITA to consider different HRTF measurement distances for near and far field acoustic sources. I have never encountered a scientific or publicly available document where Algazi's CIPIC data base was used to simulate near-field virtual sound sources. To do so in a practical implementation would require a set of HRTFs measured at multiple near-field distances, as addressed by Arend (2016).

**VIII. GROUND 3 OF THE PETITION: SHIVAPPA AND ZHOU**

29. I understand that Petitioner relies U.S. Patent No. 10,979,843 ("Shivappa") and Zhiying Zhou et al., *An Experimental Study on the Role of Software Synthesized 3D Sound in Augmented Reality Environments*, 16 Interacting with Computs. 989 (2004) ("Zhou") to allege that these references teach or suggest every limitation of independent claims. I disagree, and for the reasons discussed below, I

am of the opinion that Petitioner mischaracterized teachings of these references.

**A. Shivappa**

30. Shivappa's claims and teachings have to do with compensating latency in a 3D audio device, for situations where the rate of change in position of a listener exceed the processing required for a binaural signal. It accomplishes this via a "position predictor" algorithm (e.g., Ex. 1010, 5:10-12). Nowhere is a prediction algorithm described in the '836 patent. Shivappa appears to simply mention well-known components of VR systems as part of a larger novelty argument regarding an invention having to do with compensating latency. As part of its disclosure, Shivappa uses a conventional display and describes functions of the same: "The VR system (or the AR system) may be configured to display media content, such as movies, and to provide interactive content, such as video games, presentations, virtual meetings, etc." Ex. 1010, 5:42-45. This disclosure has nothing to do with the claimed display in the '836 Patent that provides important visual cues for depth perception (i.e., "an image that occurs within a near-field distance"). In fact, Shivappa does not disclose any virtual image, let alone one that is perceived to be within a near-field distance.

Shivappa merely discusses benefits of spatialization in general terms but does not suggest how spatialization should be accomplished in order to create the optimum user experience.

**B. Zhou**

31. Zhou is irrelevant to the petitioner's arguments. Zhou simply asserts that 3D sound "improves the accuracy of depth judgment and improves task performance." Ex. 1011, Abstract. They used an augmented reality system with a single – camera and explored "perceptual feeling of sound depth exaggeration." Ex. 1011, 4. The experiment based correct judgments on virtual objects ("which telephone is nearer") and used a very unnatural "exaggeration" of the sound cues from a virtual telephone. A PHOSITA could only conclude from Zhou that using 3D audio is better than not using 3D audio for discrimination of two artificial cues (telephone sounds at two far-field distances), which is irrelevant to '836's distinction between near-field and far-field HRTFs for simulating virtual sound source positions. In particular, Zhou never discloses that the sound should be localized behind the marker (or the augmented virtual telephone). In Zhou, "the 3D sound appears to emanate from the marker," not behind it. Ex. 1011, 7.

32. I further note that Zhou uses a single camera system that allows only for monocular vision, thereby eliminating importance cues to visual depth perception. Therefore, experiments in monocular vision are not applicable to stereo-vision head mounted displays capable of providing visual cues for depth perception. Turner (2011) confirms this: a PHOSITA would understand that the interaction of virtual acoustic cues with binocular visual cues does not translate to interaction with monoscopic visual cues, as was practiced by Zhou: “the use of headphones and the monoscopic display mean the conclusions of this study [i.e., Zhou] cannot be transferred directly to a stereoscopic 3D viewing situation.” Ex. 2012, 4.

## **IX. CONCLUSION**

33. I declare that all statements made herein of my knowledge are true, that all statements made on information and belief are believed to be true, and that these statements were made with knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code. I reserve the right to modify my opinions as additional information becomes available to me.

Case No. IPR2023-1024  
Patent No. 11,290,836 B2

Dated: September 14, 2023



By: \_\_\_\_\_  
Durand R. Begault Ph.D.