

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

META PLATFORMS, INC.,
Petitioner,

v.

EIGHT KHZ, LLC,
Patent Owner.

Case IPR2023-01004
Patent 9,282,196 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 ‘196 Patent.....	3
A. Voice of a computer program necessarily requires speech synthesis	6
B. Processing a voice of a computer program into binaural sound requires use of HRTF filters.....	12
VI. GROUND 1 OF THE PETITION: LI	15
VII. GROUND 6 OF THE PETITION: BEDINGFIELD	17
VIII. CONCLUSION.....	20

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 9,282,196 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 ‘196 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. Attached hereto and incorporated by reference as Exhibit 2008 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 ‘196 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 ‘196 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 ‘196 Patent

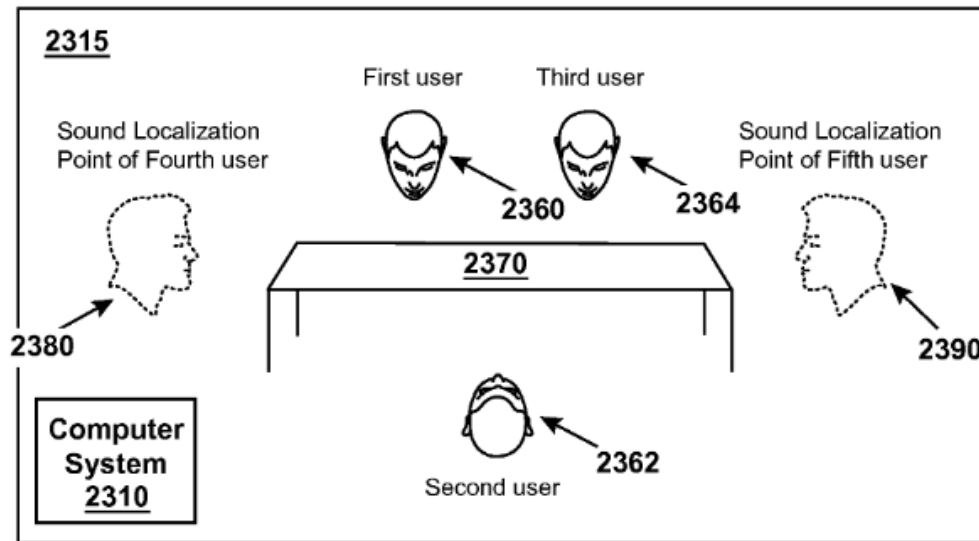
7. The ‘196 Patent discloses a method (or a system) that processes a voice of a computer program with head-related transfer functions (HRTFs)

during a voice exchange between the computer program and the user to provide to the user a binaural sound occurring at an externalized sound localization point. Claim 1 describes “an electronic system that move a sound localization point (SLP) of a **voice of a computer program** during a voice exchange between the computer program and a user.” A ‘computer voice’ must obviously use different means to vocalize sound compared to the mechanism of human speech.

8. Furthermore, a ‘computer voice’ represents more than a simple translation from text-generated speech to a computer speech synthesizer. The voice of a computer combines voice synthesis with a **character model** in order to establish a “personality” related to the computer voice. Whereas a voice model used in speech synthesis might include a vocabulary (lexicon), articulation, prosody, and other features based on examination of the speech signal itself, a character model takes information from a voice model and applies a “personality” that includes cognitive actions and interactions within the virtual world.
9. The patent describes scenarios where the computer voice is **representative of one or more virtual characters** that interact with the user. First, the patent points to the possibility that a space can be occupied by “virtual object, or other manifestation or representation of a remote

real or virtual character or object.” ‘196 Patent, 4:45-46.¹ Second, in discussing Figure 23 reproduced below as Figure 1, the patent makes clear that all human faces appear therein with the exception of that of the second user may also represent people or “computer programs with interactive audio input/output interfaces. ‘196 Patent: 35:64-36:12. In light of this example, a PHOSITA will understand that the computer voice disclosed by claims of the patent may represent a virtual character who listens, responds, and interacts with the user with an appropriate degree of autonomy.

Figure 1.



¹ I understand that the ‘196 Patent has been attached to the Petition as Ex. 1001.

10. In doing so, the ‘196 Patent uses at least two different technologies to implement and interact with a computer voice: speech synthesis and binauralization (i.e., 3-D audio processing using HRTF filters). Below, I discuss each technology in turn.

A. Voice of a computer program necessarily requires speech synthesis

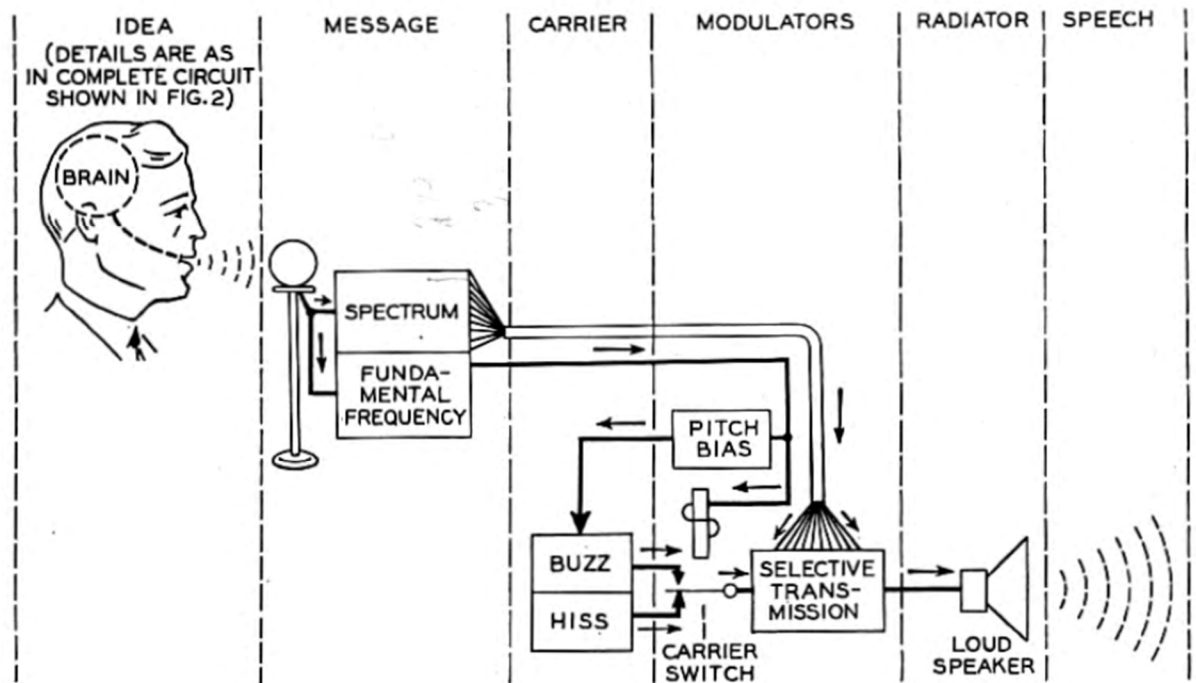
11. Speech synthesis is a distinct component of modern user interfaces to computers. “A computer may ... deliver message to the user by producing intelligible speech, which requires *speech synthesis technology*” (Ex. 2011, Pulkki and Karjalainen, 2015, p. 335). Speech synthesis technology can be connected to artificial intelligence networks to generate speech representative of virtual characters, objects, etc. In this regard, the ‘196 patent considers more than the simple attempt of a computer to convert text to speech.

12. Modern speech synthesis uses processes to create artificial sounds that are electronically transduced by a loudspeaker in order to simulate human speech in an intelligible manner. The synthesis is accomplished by deriving modeling parameters from actual speech to drive sound generators and filters in combination. For instance, many speech *phonemes*, including vowel sounds, can be modeled using analog or

digital processors that are based on the vibrations of the vocal cords and subsequent filtering by the vocal tract.

13. Figure 2 below shows an example of the *vocoder* from the mid-20th century from Bell Laboratories. The vocoder analyzed the spectrum and fundamental frequency of a real talker, and then used sound generators (“buzz”, “hiss”) and modulators to create synthetic speech (Ex. 2012, Dudley, 1940, p. 508). The actual speech of the talker is not reproduced; instead, its parameters are analyzed and then modeled to produce a synthetic signal. One can appreciate that in the current digital era, far more sophisticated means of modeling human speech are used for speech synthesis. Large data bases of talkers with complex means of modeling such as neural networks are currently used to “train” computers to talk in an increasingly realistic manner. Dialect, prosody, individual phonetic characteristics, and many other factors constitute modern speech synthesis.

Figure 2.



(from Ex. 2012, Dudley, 1940, p. 508).

14. The process of creating a 'computer voice' uses electronic means to synthesize speech, compared to the mechanisms of human speech that uses acoustical means. Speech synthesis is fundamentally based on establishing a vocabulary of sounds that can simulate speech on demand, based on *prior* and *real-time analyses* of human speech. This is an analogous concept to the use of oscillators, filters, and other electronic circuits to form musical instrument sounds using a sound synthesizer. Neither type of synthesis involves the presence of a human or a musical instrument.

15. For example, a human can simulate the sound of wind by producing air through the vocal tract without vibration of the vocal cords. They can shape the frequency characteristics of the air sound by changing the position of the tongue and jaw in a manner characteristic of forming vowel sounds. Similarly, a synthesizer can simulate the sound of wind by playing a noise generator through a filter with variable characteristics. However, both of these processes are distinct from using a microphone to capture an actual acoustical signal of the wind.
16. The processes described above differ significantly from the use of a microphone array in a telecommunication system. Rather than a computer as an intermediary for interpreting and analyzing speech, as described above, a microphone array simply captures an acoustic signal usually from a targeted location in actual space of a human, and transmits to a listener. Any processing of a received signal by HRTFs to create a virtual sound localization position (SLP) is immaterial to the distinction between computer speech synthesis and the processing of actual speech. The creation of virtual sound sources from a microphone in a real-time communication scenario is independent of the constraints and consequent advantages of a computer-based speech synthesis system.

17. Relying on the examples given by its expert, Petitioner appears to purport that the claims of the ‘196 Patent are obvious in view of text-to-speech synthesizers used for teleconferencing. Petition, 16; Ex. 1003, ¶¶ 89-93. I disagree, because the text-to-speech examples cited do not include a two-way communication with the voice of a computer system, as described in ‘196.

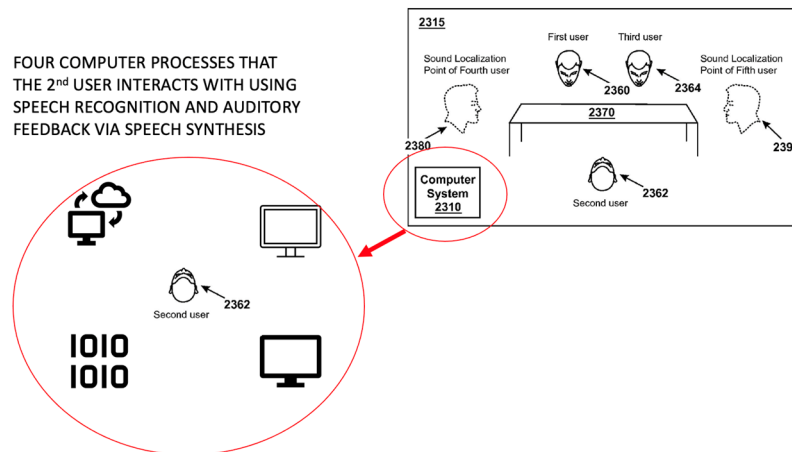
18. By contrast, the ‘196 patent refers to the following scenario that clearly establishes the use of a computer-based speech synthesis system involving multiple virtual characters:

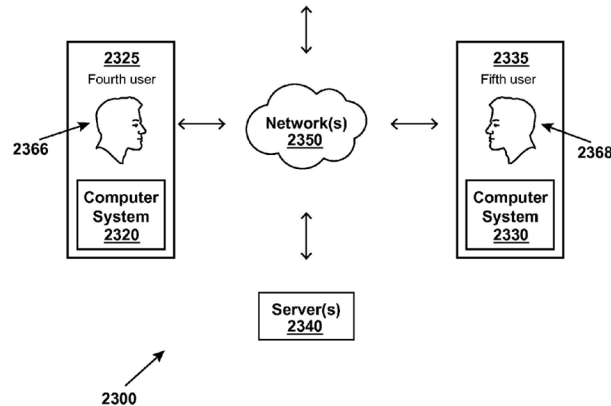
Consider an example in which the second user 2362 is speaking to and listening to the four others (2366 localized at 2380, 2360, 2364, and 2368 localized at 2390). **In this example, however, the others are not people but computer programs with interactive audio input/output interfaces.** Second user 2362 can hear the progress announced for process 2366 on his left, and he can pause, halt, and otherwise control the process by voice control, directing his voice to 2380, and likewise respectively monitoring and controlling each of the four processes by listening in their direction and speaking the voice commands back to the SLP of each. **Each audio window/process could perceive when it was the one being addressed** by the user 2362 by comparing the sound pressure received relative to that received at the other three process/window points and/or by monitoring the head orientation of user 2362.

‘196 Patent: 35:64-36:12. Note that this description is about a human-computer interface (“...computer programs with interactive input/output interfaces”) that uses speech synthesis as the “interactive output” to one or

more “audio window/process[s].” The “process” can represent virtual characters, and are monitored by a human user by means of speech or non-speech synthesis for hearing their “direction” or their “progress”: “Computer user 2362 can hear each of his programs running in a different spot on his large desk.” *Id.*, 36:12-14. Figure 5 below adapted from Figure 23 of the ‘196 Patent illustrates this point:

Figure 5. Top: Shown at left: an annotation of Figure 23 from the patent showing the computer processes representing virtual characters in their interaction. Bottom: From Figure 23, this shows that the potential interaction via a network of two virtual characters, for instance, within a gaming environment.





B. Processing a voice of a computer program into binaural sound requires use of HRTF filters

19. In his declaration, Dr. Kyriakakis mentions that “binaural sound” can be created in two ways, but does not highlight a relevant distinction between the two processes:

52. Binaural sound can be created in two ways. The first is to record binaural sound using two microphones that are spaced at the same distance as the width of a human head (i.e., a binaural recording). This is often achieved by placing microphones on the sides of a mannequin (or dummy) head. When the left microphone signal is sent to the left ear and the right microphone signal is sent to the right ear (using headphones), the listener perceives the recorded scene with precise localization of each source in 3D space.

53. The second is to synthesize binaural sound by processing non-binaural content with the left and right HRTFs of the desired placement angle and position. For example, a monaural sound source, such as a person’s voice, is processed (via convolution) with a left ear HRTF and a right ear HRTF...

Ex. 1003, ¶¶ 52-53. There is a significant difference between the first and second examples in the context of 3D sound and virtual reality, particularly in human machine interfaces.

20. The first example, “binaural recording,” does not allow a user to arbitrarily manipulate the spatial characteristics of the recorded sound. The sound localization points of sound sources in the resultant recording are fixed. In Figure 6 below, two microphones located in the ears of a mannequin are monitored by listeners over headphones. The sound source is the tambourine played by the standing person. A binaural recording made through the microphones could not be modified to place the sound of the tambourine at a different location relative to the mannequin head in post-processing.

Figure 6. From “Binaural Transmission System at the Academy of Music in Philadelphia”, Bell Telephone Laboratories, photo 55336. Reprinted in “Spatial Sound Techniques. Part 1: Virtual and Binaural Audio Technologies” (Durand R. Begault, ed.) New York: Audio Engineering Society, 2004. p. 118).



21. By contrast, the second example to “...synthesize binaural sound by processing non-binaural content with the left and right HRTFs” means that a user can place sound sources at **arbitrary locations**, based on the design and application of a human interface. I wrote in my book *3D Sound for Virtual Reality and Multimedia* (Ex. 1036, Begault, 2000, page x) that 3D sound “... refer[s] to techniques where the outer ears (the pinnae) are either directly implemented or modeled as digital filters. By filtering a digitized sound source with these filters, **one can potentially place sounds anywhere in the virtual space about a headphone listener**. The combination of 3-D sound within a human interface along with a system for managing acoustic input is termed a **3-D auditory display**.” It is the ability to use a computer to update digital filters that represent the transfer function of the outer ears as a function of source direction that takes binaural recording into the distinct domain of 3D sound processing.
22. While the term “binaural sound” is sometimes used as a proxy for 3D audio techniques, it is important to distinguish this concept from that of “binaural recording”, which has existed since before the computer age. Furthermore, one cannot use a binaural recording or a microphone array as a method to create a computer voice in 3D space in more than one fixed location, because there would be no way to interact dynamically with the

spatial configuration of the sound source and the listener. To do so in more than one fixed position, one must use digital filtering representing binaural HRTFs.

23. I note that the ‘196 Patent discloses the use of HRTFs, which by definition will include interaural time and level differences:

For example, the sound is processed so sound localization occurs for the listener to emanate from or originate at the sound localization point. For instance, the sound is processed to alter, add, or generate head-related transfer functions (HRTFs), interaural time differences, and/or interaural level differences to correspond to a fixed or variable distance from the listener, a coordinate location or a GPS location of the sound localization point.

‘196 Patent. 4:24-29.

VI. GROUND 1 OF THE PETITION: LI

24. I understand that Petitioner relies on U.S. Patent No. 9,131,305 (“Li”) to allege that this reference teaches or suggests every limitation of independent claims. I disagree, and for the reasons discussed below, I am of the opinion that Petitioner mischaracterized teachings of Li.
25. Li is directed at generating “a configurable three-dimensional binaural sound, a configurable three-dimensional stereo sound, and a configurable three-dimensional surround sound,” (Ex. 1006, 3:50-52), by processing “any existing or pre-recorded sound tracks,” (*Id.*, 6:5-7), with HRTFs measured in front of a dummy head. Li does not disclose speech synthesis.

26. Li coins new terms such as, “3D sound synthesis” and “synthesize a three-dimensional binaural sound.” But Li consistently used these terms to mean the process of generating a 3D binaural sound from a pre-recorded sound using HRTF filters.

The 3D sound processing application synthesizes and reproduces the user preferred sound field using the measured head related transfer functions (HRTFs) stored in a head related transfer function (HRTF) database 908. The 3D sound processing application performs sound track mapping 907 by convolving each of the sound tracks with corresponding HRTFs stored in the HRTF database 908 to synthesis 3D binaural sound 909 for a headset user.

Ex. 1006, 17:30-37.

27. I understand that Petitioner relies on the following passage from Li to conclude that Li discloses a voice of a computer program.

Moreover, the configurable 3D sound system 900 provides a method of communication among multiple people, for example, in a conference call, by placing different speaking users at different spots to mimic a real conference room environment.

Id., 22:1-5. However, as explained above, processing acoustic speech signals using 3D sound (HRTF processing) is not equivalent to using speech synthesis in a human-computer interface as the “voice of a computer program.” While Li teaches the use of a microphone array to capture speech from humans, Li cannot use this speech to create an arbitrary character model. Furthermore, Li gives examples using pre-recorded sounds in fixed positions; this method

could not be used for a user to interact dynamically in three dimensions with the spatial configuration of one or more sound sources such as computer voices.

VII. GROUND 6 OF THE PETITION: BEDINGFIELD

28. I understand that Petitioner relies on U.S. Patent No. 8,406,439 (“Bedingfield”) to allege that this reference teaches or suggests every limitation of independent claims. I disagree, and for the reasons discussed below, I am of the opinion that Petitioner mischaracterized teachings of Bedingfield.

29. Bedingfield is directed at a communication control system that provides multi-channel sound to a listener in a conference call. Bedingfield’s communication system assigns each participant a position within a stereo or multichannel virtual sound field such that the listener hears each participant’s sound at a particular location. The described means of “synthetic audio placement” or “synthesizing a sound field” for conference participants uses a “synthetic Audio Placement Circuit (SSAPC)” using a plurality of speakers or headphones. Ex. 1005, 8:35-48.

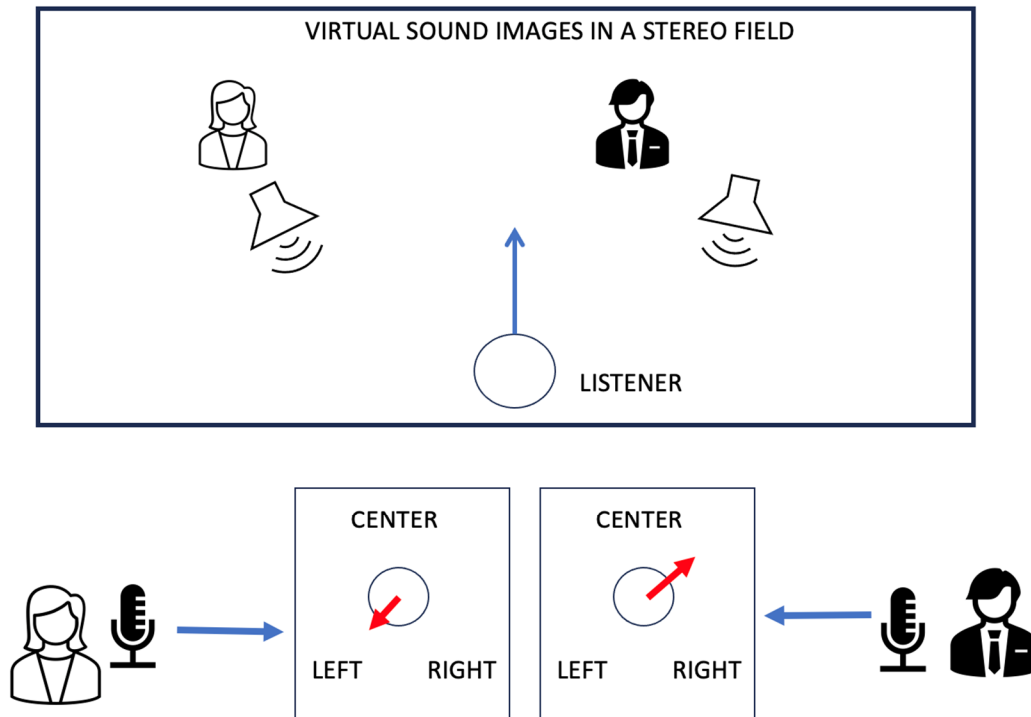
30. The basis of Bedingfield’s SSAPC system is related to the well-known concept of acoustic “pan” or level adjustment between multiple channels.

The word “pan” is an abbreviation of the term “panoramic potentiometer,” a processor found in professional mixing consoles and familiar to most persons who have a stereo system as a “balance” knob or slider. Nowhere does Bedingfield indicate the use of HRTF filtering, binaural processing, or interaural time delay. Furthermore, Bedingfield’s virtual space is by definition two-dimensional, without elevation (3D sound), and furthermore is fixed to both the height and the lateral extent of the actual loudspeakers. Bedingfield merely assigns to each sound input a variable level at the loudspeaker outputs, “thus creating a balance effect to aurally locate the participant.” *Id.*, 9:55-56.

31. Table I of Bedingfield shows the adjustment of the level of different inputs to place sounds in a stereo sound field. Figure 7 below shows how a common mixing console accomplishes the same result in stereophonic reproduction. The microphone of the person on the left is “panned” to create a virtual image at the far left of the stereo field that is defined by the extent of the loudspeakers. The microphone of the person on the right is panned to a position between the middle and the far right of the stereo field. This level adjustment corresponds to the technology taught by Bedingfield (ref. Table I). Note that one cannot exceed the angle defined

by the loudspeakers or create elevated images, using Beddingfield's method.

Figure 7.



32. Furthermore, Benningfield does not consider synthetic speech in a human-computer context, only input from actual talkers in a teleconference, as evidenced by the following:

[W]hen conferences occur over typical communication systems, spatial reference is often lost through the single channel communication. It is often difficult to determine which participant is speaking during a conference call when multiple participants sound similar or when there are a large number of participants on the conference call....

In different aspects, the methods and systems may include a connection port to receive a plurality of remote audio signals included in a conference call, at least one of the plurality of

remote audio connections being a single channel audio signal, a processor configured with the connection port to provide a sound field for placing the plurality of audio signals, the sound field providing synthetic audio placement of each of the plurality of remote audio signals, and a rendering output configured with the processor to project the plurality of remote audio within the sound field to a listener.

Ex. 1005, 1:33:39; 2:5-16. Finally, this method could not be used to interact dynamically with the spatial configuration of a sound source, such a computer voice, and a listener.

VIII. 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.

Dated: October 11, 2023

By:



Durand R. Begault Ph.D.