State of the Art: Virtual Reality Audio

Understanding Sensory Phenomena in Immersive Computer Interaction

by Murray Sandmeyer

Keywords: binaural audio, ambisonics, head mounted display, virtual reality, human computer interaction

Abstract

The following paper establishes a literature review of the state of audio sensory research within Virtual Reality (VR) technology in 2020. The state of visual sensory VR research is outlined, and several notable phenomena of visual cues, such as egocentric distance estimation, are brought to light. Several explorations of the audio sense are discussed in VR, and the need for deeper understanding of the way people perceive sounds in VR applications is illustrated. Finally, several potential areas of research into this field are recommended.

Introduction

Immersive computing in 2020 finds a stronghold in Virtual Reality (VR), a method of human computer interaction that involves immersion of many of the user's senses, such as stereoscopic vision, stereo sound, and haptic feedback. As VR becomes more widespread and accessible, it is increasingly an exciting medium for human computer interaction. According to a report by Nielsen's SuperData (a company that tracks the consumption of games-as-media), spending on VR hardware jumped from $1.6 billion in 2018 to $2.1 billion in 2019, largely attributed to the introduction of Facebook's novel standalone headset, the Oculus Quest (Sherr, 2019). VR and PC gaming software juggernaut Valve, meanwhile, released a high-end VR system in June of 2019 aimed to push the limits of immersive controllers and high resolution Head Mounted Display (HMD), which acts as a set of screens or monitors attached to a user's head as part of a VR system. Valve is also reported to be developing three high-budget video games for the system which are projected to rival mainstream console video games in terms of budget and sales (Campbell, 2017). While VR technology shows the most rapid growth in entertainment, the technology has been repurposed in a variety of fields and shown to be effective in myriad circumstances of human computer interaction.

Entertainment

As with any immersive or interactive software, the use of audio is essential, and many VR experiences incorporate sound either as primary or secondary aspects of the software. VR rhythm games, for example, are a lucrative and exciting field that has disrupted the industry and pushed sales forward, as evidenced by Facebook's decision to purchase the popular game *Beat Saber* which tasks players with cutting virtual cubes along to the rhythm of musical tracks. The game has since exploded in popularity, and some savvy users have even pushed the limits of what is believed to be humanly possible, with developers at Valve discovering that a "properly motivated human" could flick their wrist extremely quickly, and if this speed were expanded to a longer period of time, the calculation arrives at approximately 3,600 degrees per second (Liao, 2019). The argument can be make that the immersiveness and amusement of rhythm video games in VR can generate such extraordinary motivation towards propelling game users to enhance their

physical skill sets, and the role of audio in VR plays a significant role in generating such motivation. Therefore, the role of adequate audio programming in VR games can not be understated.

Military Training

Audio considerations in Virtual Reality are vital beyond mere video games. Within the past few decades, U.S. military organizations have employed VR training systems, and the multinational Eurofighter project has been vocal about its use of Aircrew Synthetic Training Aids. Hailed as one of the best aviation training systems in Europe, this simulation technology is said to provide "leading edge, high fidelity and 360 degree field of view training" (Lele, 2011). In these training sessions, military personnel can practice operating vehicles in land and water. Correct audio signals and cues are necessary, in this instance, to provide realism and audio-based prompting, and when these military skills are transferred to the real world defense scenarios, using audio cues correctly could mean the difference between success and failure, or even life and death.

Physical Therapy

Northeastern University's Rehabilitation Games and Virtual Reality Lab has conducted studies involving VR and physical therapy, with a focus on children, particularly children with cerebral palsy. In 2017, this lab found that children who practiced motor learning tasks in a VR environment "acquired the task more quickly and achieved better acquisition and retention scores" than those who completed the same motor actions in a real-life physical environment (Levac). In this way, VR has been shown to be a promising technology to potentially treat neuromotor disabilities.

Psychiatry

Other academic pursuits have shown promising results of treating mental and social disorders; one study found VR effective in treating aspects of social anxiety (Anderson et al., 2013). This study placed subjects in simulated public speaking environments, such as a virtual classroom or auditorium, and recorded alleviation of symptoms of social anxiety through the trials. Another

study placed adults with high functioning autism in practice interviews using VR technology, and discovered that the experience was helpful for improving social cue recognition in this population (Kandalaft et. al, 2012). Because realism is a noted benefit of the effectiveness of immersive virtual environment treatments and training, special attention ought to be paid to the use of visual and audio sensory cues in therapy and beyond.

Visual Phenomena in Virtual Reality

Because the bulk of sensory research involving VR is fixated on the *visual* domain, understanding the research that has been done in this field is key to laying the groundwork for the *auditory* domain of VR research. Several quirks involving visual perception have been discovered in research involving virtual environments, and these phenomena outline interesting questions for audio perceptions in immersive computing.

One consideration for making VR systems more realistic and of higher quality is studying how people perceive spaces in virtual environments. One question emerges, however: how do researchers measure a subjective, qualitative interpretation of immersive software? Many researchers have turned to distance perception as a metric because it is a convenient method of determining whether or not a user is interpreting a virtual environment as they would a physical space. A study from the Dresden University of Technology reviewed years of research and data involving distance perception in VR, and found that participants tend to view virtual objects as 74% as far away as the engineered distance. A virtual apple, for example, might be viewed as 7.4 feet away when it is in fact intended to be 10 feet away in the virtual environment (Renner, Velichkovsky, Helmert, 2013).

In an effort to remedy this recorded phenomenon, Renner et. al recommend making ground textures—which serve as a sort of distance "ruler" for a typical participant—higher resolution and more noticeable. A similar tactic to improve distance estimation is to add replicas of familiar objects to the space; a virtual book, for instance, could help a user better comprehend distances in this imaginary environment. The researchers posit that perceived "realness" of a space plays a

role in the accuracy of egocentric distance perception as well, because subjects tended to improve their estimations after being shown a small, physical replica of the room they were about to enter virtually. Because these subjects had been presented with a “real” version of a virtual space, Renner et. al acknowledge the possibility that their belief in its “realness” facilitated more thoughtful estimation and more accurate perceptions of the space.

When it comes to the role of audio in distance estimation, a study from Paquier, Côté, Devillers, and Koehl found that audio cues were less significant than visual cues for determining distance estimation of a visible object (2016). Even though hearing may be less pertinent than sight in this context, the role of the auditory sense in object distance estimation is not fully understood, and more studies ought to be conducted using solely audio cues in a virtual environment.

Egocentric viewing height is another visual phenomena that has been shown to have effects on a user’s experience within a virtual environment. In 2013, Freeman et al. placed study participants in a virtual train ride within a HMD. An independent variable in the experiment was modifying the user’s egocentric viewing height, or in other words, modifying how tall they are in the VR experience. The researchers found that lower viewing heights tended to increase paranoia in the user and decrease their perception of self compared to others. One reason for this could be that height in our society is typically attributed to greater social standing and deserving of greater respect. Regardless, because viewing perspective was shown to affect emotional and social states, audio listening perspective ought to be given similar research attention.

Audio Distance and Direction Perception

Although studies involving distance estimation of an object in a virtual environment are an open question for research, there exists some peer-reviewed knowledge on the way audio is perceived when triggered from loudspeakers, and this knowledge can be used as an entry point for discerning audio perception in VR. After all, audio from loudspeakers are “simulated” and one can make the argument that this kind of simulation is equivalent or analogous to visual VR simulations.

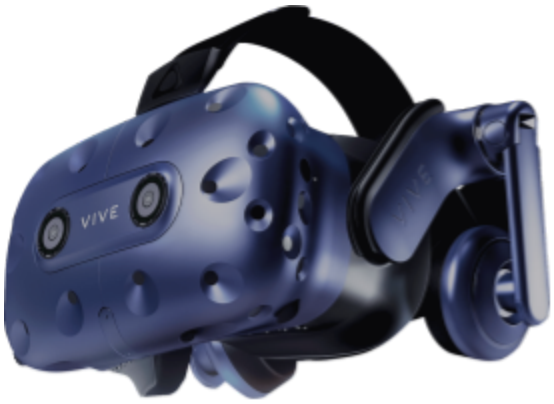
One metric of interest related to audio perception is the perceived distance of a sound from oneself. A study from Laitinen, Politis, Huhtakallio, and Pulkki in 2015 showed that distance perception can be manipulated using the placement and orientation of loudspeakers. This inquiry put participants into a room with loudspeakers that emanated individual sounds, and then manipulated to which direction the loudspeakers were pointing. In order to interpret the results, the researchers employed directivity as an independent variable and discovered that greater directivity of the speakers towards the participant caused the perception of less distance. In other words, sounds from speakers that were pointed directly at the participant were perceived to be closer than sounds from speakers that were pointing away from the participant. Even though these sounds have the same egocentric distance, one is seen as more "direct," and therefore closer to the naive participant, likely due to the fact that additional room reverberation suggests greater distance. These findings highlight the importance of users being in a position such that room speakers are oriented directly at them.

If a sound is in motion, localizing its source proves to be a similarly challenging task. Another 2015 study by 3D audio theorists Franz Zotter and Matthias Frank placed participants in a room with loudspeakers in an effort to understand how direction factors into a sound's perceived location. The study found that participants localized the sounds not at their source, but rather at their reflection paths that occur milliseconds later. These findings suggest that sound reflections in a room are perceptually significant and pose barriers for users to properly understand where a sound is coming from. Misinterpretations of a sound's distance or location, therefore, is dependent on factors such as the room in which the sound populates.

Audio in VR: Headphones or Room Speakers?

How should users experience audio in VR? As shown through consumer headsets such as the HTC Vive Pro or the Valve Index, headphones are the preferred method of VR listening in the industry. Only cheaper headsets such as the Oculus Go include speaker systems. In addition to privacy reasons, the research suggests that perception issues might occur if a user is in a HMD

connected to room speakers that deliver sound. The above findings about audio distance estimation and direction highlight inaccuracies regarding the user moving off axis from the room speakers, for example. Because perception complications happen with the use of speakers outside of an HMD, stereo headphones are the preferred solution for ensuring aural immersion. Logically, a question poses itself: how can a stereo signal in headphones be used to create realistic 3D audio? A solution exists within the field of binaural audio.



Vive Pro VR Headset manufactured by HTC

Binaural Audio

Recorded Binaural Audio

Binaural audio, or the process of creating sounds with head and ear position in mind, is a realistic format that makes signals sound three dimensional, and its realism is a sought after feature for making VR experiences more immersive. This kind of audio has historically been created via recordings where two microphones, perhaps shaped like human ears, are placed beside an artificial head. This setup simulates how sound travels to the hearing organs in the real world perception of audio. An example is shown in *Figure 2*. As a result of this process, recorded sounds through these artificials ears are more closely aligned with real life hearing, and thus the resulting signal sounds more three-dimensional and "real" than a typical stereo signal recording. Due to their realism, individual binaural sounds can be easily pinpointed to a horizontal position in space. The same cannot be said for the vertical dimension, as human hearing has notoriously poor localization of sounds above or below the horizontal axis.

2: A binaural microphone setup with artificial ears and head

Simulated Binaural Audio

While binaural audio can be recorded via mic setups such as those in *Figure 2*, this kind of audio can also be simulated, which is a desirable technology for VR experiences where sounds might be created within the environment itself. Head-Related Transfer Functions (HRTFs) can be used to accomplish this task, as they can convert individual signals at a certain location into binaural audio for a user. HRTFs process the signal by imagining a 3D environment and taking input such as the location of the sound and the location of the head. Through appropriate calculations, the function returns a generated stereo sound with one channel to each ear, and the resulting signals closely mimic real-life sound localization.

Simulated binaural audio is of interest to VR applications because not all sounds in VR environments will be recorded binaurally from the real world. In fact, many VR applications are generated through software and are purely virtual, sound sources included. Immersive 360 degree videos depicting scenes and sounds from the real world—which are a typical vehicle for recorded binaural audio—are a small subset of applications that run on HMDs; high-budget video games, for instance, employ 3D virtual environments with sound and visual sources created and triggered from within that same virtual environment. Perfecting binaural simulations are thus a key aspect of VR audio engineering.

Creating Simulated Binaural Audio

Binaural audio can be created or simulated in a number of ways separate from binaural recording. One widely used method is employing binaural software within a Digital Audio Workspace (DAW) such as Avid Pro Tools or Logic Pro X. There exist a number of free and paid binaural plugins for DAWs, including Ambeo Orbit from Sennheiser, or Binaural Surround from Blue Ripple Sound. DAW plugins such as these employ HRTFs in order to convert stereo or mono signals into binaural audio for headphones. However, implementation methods and details vary across plugins, so the scientific validity of commercial software ought to be studied before assumed to be an accurate application of HRTFs. After creating binaural sounds using plugins within a DAW, these files can be transferred to a VR experience in order to have the desired effect.

Another method for creating and facilitating virtual environments with binaural audio is software within game engines. VR experiences—not all of which are considered video games—are often created in video game engines because of the seamless tools they provide to create three dimensional environments and interactivity. Game engines, for instance, have built-in 3D models and out-of-the-box tools to control lighting and physics, among myriad other software tools and resources. The most popular game engine available to the public is Unity, which has been installed on over 3 billion devices worldwide and is used in 60% of AR/VR projects, according to the company. Another widely used development platform for VR is Unreal Engine, which has been available to the public since 1998 and has since maintained its standing in the game and VR development industry.

Both Unity and Unreal Engine have built in audio capabilities allowing developers to place sound sources at specific locations in three dimensions. Within these capabilities, a sound will get louder as a user approaches a sound source, for instance. Like DAWs, however, a binaural plugin is required to process these sounds into a more realistic binaural signal with horizontal localization for a user with stereo headphones. One such plugin for Unity is called RealSpace 3D, and the organization behind the plugin claims that their software is “the key to VR immersion.” Trivial demos and videos of the software demonstrate its ability to allow a user to localize sounds via HRTFs in addition to the impressive feat of simulating the doppler effect, but its purpose is fundamentally commercial, and the scientific validity of its HRTFs are not yet known. A similar plugin exists for Unreal Engine 4, called Steam Audio, and is developed by Valve Corporation. Even though Steam Audio is attributed to a massive, well-respected games company like Valve, its plugin for Unreal Engine 4 ought to be similarly studied for scientific validity. If these tools can be proven to closely mimic real-life audio phenomena, their use in applications for VR research can be assured.

VR research with binaural audio

The effects of incorporating binaural audio into VR environments is already becoming clearer with emerging research. Similar to the VR neuromotor task being studied at Northeastern's Rehabilitation Games and VR Lab, academics at Aalborg University in Denmark utilized binaural audio as a key aspect of a novel VR searching task. In this inquiry, subjects wearing an HMD were asked to look around in the virtual environment and search for a line segment of a specific color and length. Previously being aware that visual effects can help users find objects in cluttered environments, the researchers added stereo and binaural audio cues to the searching task, and discovered that participants were able to find line segments quickest when the task is accompanied by three-dimensional binaural audio cues (Hoeg, Gerry, Thomsen, Nilsson, Serafin, 2017). Users were able to synthesize the binaural audio signal with the visual domain in order to reduce their reaction times, and the helpfulness of an immersive binaural signal—moreso than a typical stereo signal—was demonstrated in the results.

In addition to the positive effect of reduced reaction times in a searching task, binaural audio can have a negative effect on users, as evidenced by a traffic noise study conducted at the Second University of Naples. In this 2013 inquiry, Ruotolo et. al employed binaural audio in a VR environment in which participants were placed near a virtual roadway, and the binaural signal supplemented the scene with additional traffic noise. In an effort to determine in greater detail how people are affected by this kind of roadway noise, the binaural audio was engineered to be a realistic auditory image of a nearby roadway with traffic. As a result of the added noise, participants were negatively affected, and the study noted that the overall wellbeing of users declined when exposed to binaural traffic sounds in the experience. Because the study echoes the existing research of poor wellbeing in proximity to noise, the researchers concluded that "immersive virtual reality could be considered a valid tool to simulate the multisensory way in which the environment, with embedded sounds, is perceived in everyday life and can offer innovative applications" (Ruotolo et. al, 2013, p. 18). Much like unpleasant noise in real life, binaural VR audio can be harsh, discordant, and tangibly affect the people exposed to it.

Ambisonics

Ambisonics, a full-sphere audio format that takes the horizontal and vertical axis into consideration, is a field of interest for VR audio because of its capabilities regarding the representation of realistic ambient audio. Similar to recorded binaural audio, different orders of ambisonics can be recorded through specific microphone arrays, the simplest and most common being first-order ambisonics, in which four capsule microphones are employed to capture audio in 360 degrees (*Figure 3*). In first order ambisonics, the microphone setup captures four directions: W, X, Y, and Z, and each direction contributes to the full picture of the spherical audio space. Second-order and third-order ambisonics, by comparison, require 9 channels and 16 channels, respectively.

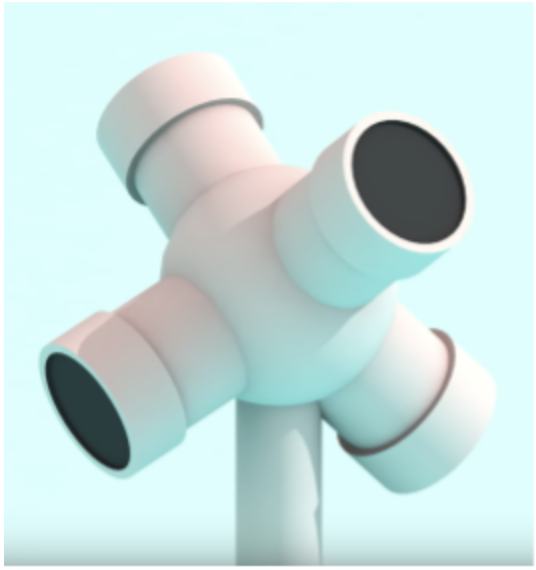


Figure 3: A four microphone setup for recording first-order ambisonics

In a VR environment, ambisonics are typically used in the process of placing ambient sounds in a virtual environment. The workflow is as follows: first, ambisonic microphones are used to record the ambient sounds. Second, the ambisonic signals are converted and routed to *virtual* speakers in the VR experience. Finally, because the ambient audio is playing through virtual speakers in the environment, binaural audio plugins and HRTFs are used to convert the virtual speaker sounds into binaural audio in the user's headphones. This creates the effect of simulating 360 degree audio for a user in a virtual environment.

Several questions arise as a result of this virtual speaker workflow: How many virtual speakers should be used in the environment? Where should these virtual speakers be placed? An intuitive solution is to place an equal number of virtual speakers to the number of capsule microphones used to record the ambisonic audio, in addition to placing them in a similar orientation and layout. One study in 2018 reinforces this intuitive solution with first order ambisonics (Chang, Cho). The study analyzed sound signals processed through HRTFs and aimed to uncover the

optimal virtual speaker solution for maintaining the integrity of the ambisonic signal, and the researchers discovered that “the lesser virtual loudspeakers are used, the better results can be obtained, as long as the number of the loudspeakers is equal to or greater than the number of the components of ambisonics” (Chang, Cho, 2018). This means that in first-order ambisonics, where four microphones are used, it is ideal to use four virtual loudspeakers, as additional loudspeakers reduce the quality of the ambisonic audio. In a practical setting, these virtual speakers could be placed invisibly in a game engine such as Unity. If a participant is in a virtual beach environment, for example, the invisible speakers could be playing the ambient sounds of waves and seagulls, and to the user, the ambient sound would appear seamless to the virtual environment.

Conclusions

The year 2020 and beyond is an exciting time to be involved in the field of VR audio. Although much remains to be discovered, such as the role of audio in egocentric distance estimation, or the role of virtual speakers in higher-order ambisonics, there is a gradually growing repository of knowledge on the subject of audio in virtual environments. Binaural audio has been shown to affect VR users—for better or for worse—and the information thereof will only continue to grow.

Consequently, many notable knowledge areas are ripe for additional research. Because there is a breadth of data surrounding distance estimation using visual cues in VR environments, it follows logically that VR distance estimation using *audio* cues is an opportunity for academics to document the relationship between sensory input and physical assessment of virtual spaces. By combining these domains, there can be an improved understanding with regards to facilitating the accurate perception of virtual environments.

Lastly, additional research ought to be conducted on the measurable outcomes of employing binaural audio in interactive VR spaces. While the Hoeg et. al study found that binaural audio reduces reaction times, other measurable outcomes come to mind, such as fine motor skills, or

hand eye coordination. The preliminary research with binaural cues ought to prompt a variety of research questions involving the positive—or negative—impact of the technology. Nevertheless, VR audio is already proving to be a fruitful domain for knowledge gathering.

References

Anderson, P. L., Price, M., Edwards, S. M., Obasaju, M. A., Schmertz, S. K., Zimand, E., & Calamaras, M. R. (2013). Virtual reality exposure therapy for social anxiety disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 81*(5), 751-760. doi:10.1037/a0033559

Campbell, C. (2017, February 10). Valve is working on three full VR games. Retrieved from https://www.polygon.com/virtual-reality/2017/2/10/14580932/valve-is-working-on-three-full-vr-games

Chang, Ji-Ho, and Wan-Ho Cho (2018). Impairments of Binaural Sound Based on Ambisonics for Virtual Reality Audio. 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM) 2018: 341-45. Web.

Hoeg, E. R., Gerry, L. J., Thomsen, L., Nilsson, N. C., & Serafin, S. (2017). Binaural sound reduces reaction time in a virtual reality search task. *2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*. doi: 10.1109/sive.2017.7901610

Kandalaft, M. R., Didehbani, N., Krawczyk, D. C., Allen, T. T., & Chapman, S. B. (2012). Virtual Reality Social Cognition Training for Young Adults with High-Functioning Autism. *Journal of Autism and Developmental Disorders, 43*(1), 34-44. doi:10.1007/s10803-012-1544-6

Kelly, Jonathan W., Lucia A. Cherep, Brenna Klesel, Zachary D. Siegel, and Seth George (2018). Comparison of Two Methods for Improving Distance Perception in Virtual Reality. *ACM Transactions on Applied Perception* 15.2: 1-11. Web

Laitinen, M.-V.,Politis, A., Huhtakallio, I., Pulkki, V. (2015). Controlling the perceived distance of an auditory object by manipulation of loudspeaker directivity. *J. Acoust. Soc. Am.* 137(6) EL462–EL468.

Lele, A. (2011). Virtual reality and its military utility. *Journal of Ambient Intelligence and Humanized Computing, 4*(1), 17-26. doi:10.1007/s12652-011-0052-4

Levac, D. E., & Jovanovic, B. B. (2017, June 19). Is children's motor learning of a postural reaching task enhanced by practice in a virtual environment? Retrieved November 12, 2017. <http://ieeexplore.ieee.org/abstract/document/8007489/>.

Liao, S. (2019, February 11). Humans playing VR game Beat Saber move faster than what Steam thought was 'humanly possible'. Retrieved from https://www.theverge.com/2019/2/11/18220993/vr-valve-steam-beat-saber-fast-speeds

Paquier, Côté, Devillers, and Koehl (2016). Interaction between Auditory and Visual Perceptions on Distance Estimations in a Virtual Environment." *Applied Acoustics* 105.C: 186-99. Web.

Renner, R., Velichkovsky, B. & Helmert, J. (2013). The perception of egocentric distances in virtual environments - A review. *ACM Computing Surveys (CSUR)*, 46(2), pp.1–40.

Ruotolo, F., Maffei, L., Gabriele, M. D., Iachini, T., Masullo, M., Ruggiero, G., & Senese, V. P. (2013). Immersive virtual reality and environmental noise assessment: An innovative audio–visual approach. *Environmental Impact Assessment Review*, *41*, 10–20. doi: 10.1016/j.eiar.2013.01.007

Sherr, I. (2019, December 11). VR sales on the upswing, thanks to Facebook's Oculus Quest. Retrieved from https://www.cnet.com/news/vr-sales-on-the-upswing-thanks-to-facebooks-oculus-quest/

Zotter, F., Frank, M. Investigation of auditory objects caused by directional sound sources in rooms. Acta Phys. Pol. A 128(1-A) (2015), http://przyrbwn.icm.edu.pl/APP/PDF/128/a128z1ap01.pdf

Acknowledgements

My instructor, Prof. Anthony De Ritis, and my classmates were essential in my process of completing this report. Special acknowledgements go to Kyle McCrosson and Jan Zimmerman for offering guidance on the title of the paper and leading me to sources of additional research, respectively.