3-D Ambisonics Experience for Virtual Reality

Cedric Yue, Teun de Planque
Stanford University
{cedyue, teun}@stanford.edu

Abstract

To create an immersive virtual reality experience both graphics and audio need to be of high quality. Nevertheless, while much virtual reality research has focused on graphics and hardware, there has been less research into audio for virtual reality. Ambisonics is a technique that can provide virtual reality users with an immersive 360 degrees surround audio experience. For this project we built a virtual reality application for Google Cardboard in which the user can experience the audio produced with ambisonics. We also made a B-format ambisonics channel visualization. Four particle simulations show samples of each of the four first order ambisonics channels. In addition, we created a particle simulation with 512 particles that show the frequency components of the binaural version of the sound.

1 Introduction

1.1 Motivation

To create an immersive alternate reality both immersive graphics and audio are necessary. Without high quality audio that matches the graphics users do not truly feel part of the virtual reality. Ambisonics and spatial audio have much potential to improve the virtual reality sound experience. Ambisonics is a method to record, modify, and recreate audio in 360 degrees. After being selected by Google as the preferred audio format for Google VR and being supported by leading game engine Unity there has been rising interesting in ambisonics for virtual reality. While most sound recording techniques encode information corresponding to specific speakers, ambisonics encodes the whole spherical soundfield. The major benefit of this approach is that the recorded sound can be reproduced with a variable number of speakers at a variety of positions distanced horizontally and vertically from the listener. In this project we experimented with ambisonics and created an audio and visual ambisonics experience. We obtained an ambisonics audio encoding from Anna Tskhovrebov of the Stanford Center for Computer Research in Music and Acoustics (CCRMA). We then used Unity to create a Google CardBoard virtual reality application with ambisonics based audio. Our project also has an educational component. There still seem to be few virtual reality developers familiar with how ambisonics work. We hope that with our application users can learn more about ambisonics and how to use ambisonics to create a virtual reality application with great audio. The scene we created contains ambisonic channel visualizations. The ambisonics B-format file format consists of four channels W, X, Y, and Z. We created particle visualizations that show the samples of the four B-format channels and change location based on the beat of the binaural version of the music. Each of these four particle visualizations contains 512 particles. The size of each particle changes over time and represents samples of the four B-format ambisonic channels. The particles are animated to behave like broken arcs shaped firework. The radius of the arc is determined by the amplitude of the raw output data. In addition, in our application there is a visualization of the frequency components of the binaural version of the sound in the form of a spiral. We used the Fourier transform to determine the frequency components. The spiral moves around based on the beat of the music, and each of the 512 spiral particles corresponds to one of the frequency components of the binaural version of the sound.

2 Related Work

Many virtual reality research papers have focused on graphics and hardware, but there have been relatively few papers about audio [1-4]. However, research demonstrates that audio is essential for an immersive virtual reality experience. Hendrix et al. showed that 3-D sound greatly increases people's sense of presence in virtual realities [1]. The study showed that people felt more immersed in virtual realities with spatial sound than in virtual realities with non-spatial sound or no sound [1,2]. Similarly, Naef et al. found that creating two separate sound signals for the left and right ear based on the head-related transfer function (HRTF) leads to a more realistic sound experience [2]. The HRTF is based on the shapes and sizes of the ears and head. As head and ear shapes vary between people it is common to use general HRTF, since it is time-consuming and expensive to compute the HRTF for all users. Vljame et al. and Gilkey et al. studied the consequences of suddenly removing sound for someone in the real world. Both Vljame et al. and Gilkey et al. showed that the sudden loss of sound makes people feel lost in the physical world; they lose their sense of being part of their environments [3,4]. Sundareswaran et. al., on the other hand, tried to guide people through a virtual world using only 3-D audio. They found that people can quickly identify the location of 3-D audio sources, meaning that 3-D audio can be an effective way to guide people through an application. Without 3-D audio users could localize audio sources within only 40 degrees, while with 3-D audio the users were able to identify the location of objects within 25 degrees. However, according to the study localizing a sound source is relatively challenging when the sound source is behind the user. Given that audio quality has a large impact on how users experience the virtual world, one might wonder why audio has received relatively limited research attention. In the paper ”3-D Sound for Virtual Reality and Multimedia” Begault and Trejo argue that audio has played a less important role in virtual reality development for two main reasons [5]. First, audio is not absolutely necessary for developing a virtual reality. One can use a virtual reality headset without audio, just like a deaf person can legally drive a car without (mostly) problems [5]. Second, the tools needed to develop 3-D audio for virtual reality have long been limited to audio experts, composer musicians, and music academics [5]. Many people simply did not have the tools to develop audio for virtual reality [5]. Begault and Trejo believe that as audio hardware is becoming less expensive, more people will start developing audio to create more realistic virtual reality experiences [5].

To improve audio for virtual reality researchers have experimented with small microphones placed inside and around the ears of users that measure sound with high precision [6,7]. In this way it is possible to create personalized and highly realistic audio. Subsequently, an audio engineer plays a sound from a specific location. Harma et al. applied this idea of personalized audio to augmented reality [6]. They placed two small microphones in the ear of each user, and then collected and processed audio from different locations and sources [6]. The downside of this approach is that each measurement is only accurate for sound coming from a sound source at one specific location [6]. To create a comprehensive personalized sound landscape the speakers has to be placed at hundreds or even thousands of locations around the user [6]. VisiSonics, the company that is currently the main audio technology supplier for Oculus Rift, aims to tackle this problem by placing speakers in the users ears instead of microphones [7]. The researchers swap the speakers with microphones, play sound through the speakers, and then record the sound with microphones placed at many locations around the user [7]. In this way they can pick up the necessary information to create a personalized audio experience in several seconds [7]. Nevertheless, most of these new 3-D personalized sound recording techniques are still best for headphones instead of 360 degrees surround sound speakers, and without head tracker the speaker needs to sit still in a relatively small space [6,7].

Figure 1: Omnidirectional soundfield microphone for ambisonics

Sennheiser recently released a microphone specifically for 360 degrees surround sound; this Sennheiser microphone can be used for ambisonic recording (see Figure 1) [7]. Research into ambisonics for virtual reality has so far been most limited [5-9]. Since its adoption by Google as the audio format of choice for virtual reality, and Unitys support of ambisonics B-format audio file format, ambisonics has seen an increase in interest [7,10]. Research into ambisonics was started in the 1972 when Michael Gerzon wrote the first paper about first order ambisonics and the B-format [9,11]. In the paper he described the first order ambisonics encoding and decoding process for the

B-format, and how the W, X, Y, and Z channels capture information about the three-dimensional sound field [9,11].

There have been several startups that have made significant progress improving audio for virtual reality. For example, Ossic has developed headphones that have built-in sensors to measure the size and shape of your ears and head [12]. With these measurements the headphones can calibrate sound and create personalized audio [12]. Ossic also developed an audio oriented rendering engine for HTC Vive [12]. Developers can create object based sound to guide users through virtual reality experiences [7]. Accurate head tracking plays an essential role in virtual reality audio. With accurate head tracking audio can provide users with a better sense of direction [7]. Developers can use audio cues to guide users through virtual environments in a natural way [7].

3 Methods

3.1 Ambisonics

Figure 2: With ambisonics the listener can be surrounded by a large number of synthesized speakers [10].

Ambisonics is a technique to record, modify, and recreate audio in 360 degrees. For this project we obtained an ambisonics audio encoding from Anna Tskhovrebov of the Stanford Center for Computer Research in Music and Acoustics (CCRMA). We then used Unity to create a Google CardBoard virtual reality application with ambisonics based audio. In contrast to standard stereo sound, ambisonics can be used to produce sound in horizontal and vertical surround leading to a much more immersive experience. After Google selected ambisonics and its B-format as its preferred audio format for virtual reality, and Unity started to support the ambisonics B-format, ambisonics has seen a surge of interest. In contrast to usual stereo audio, ambisonics B-format does not encode speaker information. Instead it encodes the sound field created by multiple sound sources using spherical harmonics. The B-format encoding can be decoded into a variable number of speakers in all directions around the listener. This flexibility is one of the main advantages of using ambisonics for virtual reality. One of the major audio challenges for virtual reality is matching the sound with the user's viewing direction. With ambisonics speakers can be placed at all locations around the user, so that the sound and virtual speaker sphere around the user can match the viewing direction. The decoding process can turn the B-format into binaural sound for headphones and is relatively light in terms of computation. As a result, the decoding can be done in real-time. The relatively little computation needed for the decoding process makes ambisonics well-suited for devices with limited computing power such as smartphones.

3.1.1 Encoding

Ambisonics encodings contain a spherical harmonics based approximation of the entire sound field. The spherical harmonic $Y_l^m(\theta,\phi)$ of degree l and order m for azimuth angle θ and elevation angle ϕ is commonly denoted using:

$$Y_l^m(\theta,\phi) = N_l^{|m|} P_l^{|m|}(\sin\phi) \cdot \begin{cases} \sin -m\theta, & \text{if } m < 0 \\ \cos m\theta, & \text{if } m \geq 0 \end{cases}$$

where P_l^m is the Legendre polynomial with degree l and order m and N is a normalization term [12]. The ambisonics spherical coordinate system can be seen in Figure 3. The azimuth ϕ is zero along the positive x-axis and increases in the counter-clockwise direction [9]. The elevation angle θ is also zero along the positive x-axis and increases in the positive z-direction [9]. In contrast to the standard spherical coordinate system θ is used for the azimuth and ϕ is used for the elevation angle.

The contribution of sound source s_i with the direction (θ,ϕ) towards ambisonic component B_l^m with degree l and order m is: $B_l^m = Y_l^m(\theta,\phi) \cdot S$. The normalization factor is the length of the

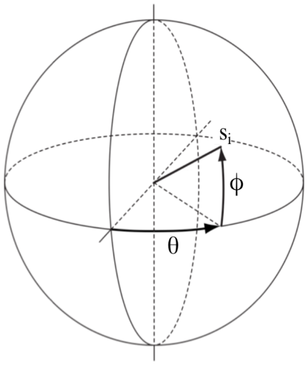

Figure 3: The spherical coordinate system for ambisonics

vector spanning from the origin in the direction of the sound source until it intersects the spherical harmonics. To determine each ambisonic component B_l^m we compute the contributions of each of the sound sources s_i towards the ambisonic component B_l^m and sum them up. We repeat this process for all ambisonic components B_l^m to obtain the encoding.

The first order ambisonic approximation to the sound field is the B-format. This encoding represents the 3-dimensional sound field using spherical harmonics with four functions, namely the zeroth order function W, and the three first order functions X, Y, and Z [13]. The zeroth order function W represents the sound pressure, X represents the front-minus-back sound pressure gradient, Y represents the left-minus-right sound pressure gradient, and Z represents the up-minus-down sound pressure gradient [13,14]. During the recording the W component is captured by an omnidirectional microphone as was shown in Figure 1, and the XYZ components can be captured by a eight recorders oriented along the three axis x,y,z [13,14]. We compute W, X, Y, Z using:

$$W = \frac{1}{k}\sum_{i=1}^{k} s_i[\frac{1}{\sqrt{2}}]$$

$$X = \frac{1}{k}\sum_{i=1}^{k} s_i \cos\phi_i \cos\theta_i$$

$$Y = \frac{1}{k}\sum_{i=1}^{k} s_i \sin\phi_i \cos\theta_i$$

$$Z = \frac{1}{k}\sum_{i=1}^{k} s_i \sin\theta_i$$

3.1.2 Decoding

The ambisonic decoding process aims to reconstruct the original 3-dimensional sound field at the origin of the spherical coordinate system. This point at the origin of coordinate system is at the center of the loudspeakers and is known as the sweet spot. The ambisonic encoding does not require one specific loud speaker setup to recreate the sound field and the encoding does not contain any specific original speaker information. Nonetheless, when selecting a speaker setup for the decoding process it is best to keep the layout of the speakers regular. The sound field can be approximately reconstructed with a variable number of loudspeakers. The only requirement is that the number of speakers L is larger than or equal to the number of ambisonic channels N. When using the B-format which has four channels this means that there needs to be at least four speakers to reconstruct the encoded sound field. Nevertheless, it is better to use more than N speakers; with more speakers the approximation of the original 3-dimensional sound field will be more accurate. If a user is only interested in horizontal surround sound, only the first three ambisonic channels of the B-format need to be used. In this case the Z channel can be ignored since it only contains information about the sound field in the z (up and down) directions.

The loudspeaker signals are obtained from the ambisonic encoding by spatially sampling the spherical harmonics with the speakers. Each speaker gets a weighted sum of the ambisonic channels. The weighted sum of each speaker is the value of its corresponding spherical harmonic. Hence, the signal for the j-th loudspeaker p_j of the L speakers for the B-format is [13,14]:

$$p_j = \frac{1}{L}[W(\frac{1}{\sqrt{2}}) + X(\cos\phi_j \cos\theta_j) + Y(\sin\phi_j \cos\theta_j) + Z(\sin\phi_j)]$$

3.1.3 Higher Order Ambisonics

It is possible to expand ambisonics to higher orders. Higher order encodings can enlarge the quality of the reconstructed sound field and the size for which the sound field is accurately reconstructed [9,15,16]. Higher order encodings are created by including additional components of the multipole expansion of a function on a sphere with spherical harmonics [17,18]. For a m-th order ambisonics encoding there are $(m+1)^2$ channels. As the number of channels increases more speakers are needed for the decoding process to accurately reconstruct the sound field [19,20]. At least m speakers are necessary.

4 Evaluation

Figure 4 shows the four particle simulations on the sides with the samples of each of the four first order ambisonics channels. The particle simulations are all built with Unity. Each of the circles consists of 512 particles. The size of each particle corresponds to samples of the four B-format channels W, X, Y, and Z. Figure 4 also contains the spiral with 512 particles in the middle that show the frequency components of the binaural version of the sound. Figure 5 is zoomed in specifically on this spiral. The spiral moves around based on the beat of the music. Most users who tried out the Google Cardboard application specifically mentioned the high quality ambisonic based sound. They found the experience to be more most immersive as a result of the high quality 360 degrees surround sound. Users also enjoyed learning more about ambisonics and the frequency domain with the ambisonic channels and binaural sound frequency component visualizations.

5 Discussion

Over the course of the project we have learned a lot about the theory of ambisonics. We believe that ambisonics have a lot of potential to create better virtual reality experiences. However, even though developers can create better virtual reality experiences using ambisonics, there still seem to be few virtual reality developers familiar with how ambisonics work. This was one of the main reasons why we decided to display samples of the ambisonics B-format in our scene. We hope that our users

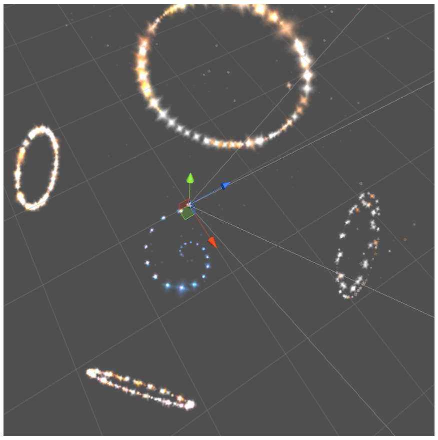

Figure 4: Overview of the scene with ambisonics frequency components

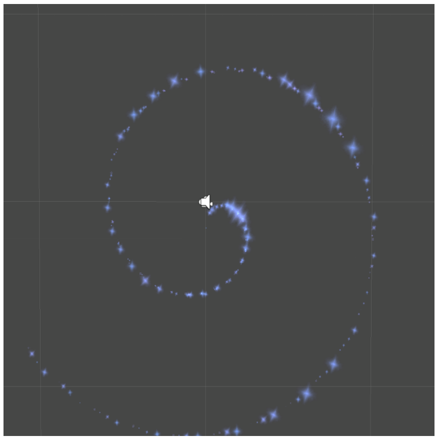

Figure 5: Bottom spiral with stereo sound frequency components

learned more about the theory behind ambisonics from our project. During the demo and while developing our application we tried to explain the theory behind the four channels to our users and testers. We hope that the experience was both enjoyable and educational for our users. During the summer we plan to build an immersive virtual reality ambisonics education experience. Users can experience and learn about ambisonics at the same time. The experience will guide users through a scene during which they learn about the theory of ambisonics. Our plan is to let users experiment with different speaker setups, encodings, and multiple order ambisonic encodings. The applications tells them more about how the different ambisonics elements work. We hope that by building such an educational application we can also learn more about virtual reality and specifically audio for virtual reality. We also plan to experiment more with higher order ambisonic encodings. For this project we only had a access to a B-format encoding. We would like to work with higher 2nd and 3rd order encodings to improve the quality of the reconstructed sound field and enlarge the size of the sweet spot. Another interesting application of ambisonics is gaze guidance. With ambisonics users can accurately localize sound sources [21,22]. Sound sources can be positioned in such a way so that they guide users through a virtual reality experience [21,23]. Traditionally, mostly visual cues have been used to move users through virtual reality applications [21,23]. The main benefit of ambisonic based cues over visual cues is that the ambisonic based cues are more realistic and require no visual changes in the virtual world. Using ambisonics based audio is more similar to the real world than for example artificial arrows that point users around the scene. In the future we would like to experiment with applying ambisonics for user gaze guidance, especially in combination with eye tracking.

Acknowledgements

We would like to thank the EE267 coursestaff for their ongoing support. In addition, we would like to thank Anna Tskhovrebov for providing us with an Ambisonics B-format audio file.

References

[1] Hendrix, Claudia, and Woodrow Barfield. ”The sense of presence within auditory virtual environments.” Presence: Teleoperators & Virtual Environments 5.3 (1996): 290-301.

[2] Naef, Martin, Oliver Staadt, and Markus Gross. ”Spatialized audio rendering for immersive virtual environments.” Proceedings of the ACM symposium on Virtual reality software and technology. ACM, 2002.

[3] Vljame, Er, et al. ”Auditory presence, individualized head-related transfer functions, and illusory egomotion in virtual environments.” in in Proc. of Seventh Annual Workshop Presence 2004. 2004.

[4] Gilkey, Robert H., and Janet M. Weisenberger. ”The sense of presence for the suddenly deafened adult: Implications for virtual environments.” Presence: Teleoperators & Virtual Environments 4.4 (1995): 357-363.

[5] Begault, Durand R., and Leonard J. Trejo. ”3-D sound for virtual reality and multimedia.” (2000).

[6] Murray, Craig D., Paul Arnold, and Ben Thornton. ”Presence accompanying induced hearing loss: Implications for immersive virtual environments.” Presence: Teleoperators and Virtual Environments 9.2 (2000): 137-148.

[7] Lalwani, Mona. ”For VR to be truly immersive, it needs convincing sound to match.” Engadget. Engadget, 14 July 2016. Web. 11 June 2017.

[8] Hrm, Aki, et al. ”Augmented reality audio for mobile and wearable appliances.” Journal of the Audio Engineering Society 52.6 (2004): 618-639.

[9] Hollerweger, Florian. ”An Introduction to Higher Order Ambisonic.” April 2005 (2013).

[10] ”Google VR Spatial Audio.” Google. Google, 7 Apr. 2015. Web. 09 June 2017.

[11] Gerzon, Michael A. ”Ambisonics in multichannel broadcasting and video.” Journal of the Audio Engineering Society 33.11 (1985): 859-871.

[12] Nachbar, Christian, et al. ”Ambix-a suggested ambisonics format.” Ambisonics Symposium, Lexington. 2011.

[13] Gauthier, P. A., et al. ”Derivation of Ambisonics signals and plane wave description of measured sound field using irregular microphone arrays and inverse problem theory.” reproduction 3 (2011): 4.

[14] Malham, David G., and Anthony Myatt. ”3-D sound spatialization using ambisonic techniques.” Computer music journal 19.4 (1995): 58-70.

[15] Daniel, Jrme, Sebastien Moreau, and Rozenn Nicol. ”Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging.” Audio Engineering Society Convention 114. Audio Engineering Society, 2003.

[16] Daniel, Jrme, Jean-Bernard Rault, and Jean-Dominique Polack. ”Ambisonics encoding of other audio formats for multiple listening conditions.” Audio Engineering Society Convention 105. Audio Engineering Society, 1998.

[17] Scaini, Davide, and Daniel Arteaga. ”Decoding of higher order ambisonics to irregular periphonic loudspeaker arrays.” Audio Engineering Society Conference: 55th International Conference: Spatial Audio. Audio Engineering Society, 2014.

[18] Spors, Sascha, and Jens Ahrens. ”A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling.” Audio Engineering Society Convention 125. Audio Engineering Society, 2008.

[19] Braun, Sebastian, and Matthias Frank. ”Localization of 3D ambisonic recordings and ambisonic virtual sources.” 1st International Conference on Spatial Audio,(Detmold). 2011.

[20] Bertet, Stphanie, et al. ”Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones.” Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments. Audio Engineering Society, 2007.

[21] Sridharan, Srinivas, James Pieszala, and Reynold Bailey. ”Depth-based subtle gaze guidance in virtual reality environments.” Proceedings of the ACM SIGGRAPH Symposium on Applied Perception. ACM, 2015.

[22] Padmanaban, Nitish, and Keenan Molner. ”Explorations in Spatial Audio and Perception for Virtual Reality.”

[23] Latif, Nida, et al. ”The art of gaze guidance.” Journal of experimental psychology: human perception and performance 40.1 (2014): 33.