Bachelor's thesis

Information and Communications Technology

2019

Lasse Pouru

THE PARAMETERS OF REALISTIC SPATIAL AUDIO

– An Experiment with Directivity and Immersion

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Information and Communications Technology

2019 | 43 pages

Supervisor: David Oliva

Lasse Pouru

THE PARAMETERS OF REALISTIC SPATIAL AUDIO

- An Experiment with Directivity and Immersion

This work deals with the software implementation of spatial audio, or 3D audio, in today's game engines. Multiple competing software solutions exist, but little research has been done as to which techniques and parameters generally yield the optimal results in regard to the localization of sound sources and the immersivity of the listening experience. This complex problem is approached in this study through comparing the effects of the parameters of the listener directivity pattern of the Resonance Audio spatial audio plugin on the accuracy of directional hearing and the listener's subjective experience of the realism of the spatial impression in a virtual auditory space. Because Resonance Audio is free software, the results of the study are also more generally applicable.

An experiment where test subjects assessed the directions of sounds coming from around them and rated the realism of the listening experience was conducted in a virtual test environment built with Unity engine. None of the tested parameter combinations stood out as clearly better than the rest, but statistically significant differences were nonetheless found between some of the values. Overall the results were in line with the expectations and the empirical studies on human spatial hearing, which gave grounds to assume that the test environment used in the experiment would likely also be suitable for similar experiments in the future.

KEYWORDS:

Spatial audio, immersion, hearing, sound processing, virtual reality

OPINNÄYTETYÖ (AMK) | TIIVISTELMÄ

TURUN AMMATTIKORKEAKOULU

Tieto- ja viestintätekniikka

2019 | 43 sivua

Ohjaaja: David Oliva

Lasse Pouru

THE PARAMETERS OF REALISTIC SPATIAL AUDIO

- An Experiment with Directivity and Immersion

Tämä työ käsittelee tilaäänen eli 3D-äänen ohjelmallista toteutusta nykypäivän pelimoottoreissa. Kilpailevia ohjelmistoratkaisuja on useita, mutta juurikaan ei ole tutkittu millä tekniikoilla ja parametreilla yleisesti ottaen saavutetaan äänilähteiden paikantamisen sekä kuuntelukokemuksen immersiivisyyden kannalta parhaat tulokset. Tätä monimutkaista ongelmaa lähestytään tässä tutkimuksessa vertailemalla Resonance Audio -tilaäänilisäosan suuntakuvion parametrien vaikutusta suuntakuulon tarkkuuteen ja kuulijan subjektiiviseen kokemukseen tilavaikutelman luonnollisuudesta virtuaalisessa kuuloavaruudessa. Koska Resonance Audio on vapaa ohjelmisto, tutkimuksen tulokset ovat myös yleisemmin sovellettavissa.

Unity-moottorilla rakennetussa virtuaalisessa testiympäristössä suoritettiin koe, jossa koehenkilöt arvioivat ympäriltään tulevien äänien suuntaa ja antoivat arvosanan kuuntelukokemuksen luonnollisuudelle. Yksikään testatuista parametriyhdistelmistä ei erottunut selvästi muita parempana, mutta tiettyjen arvojen väliltä tilastollisesti merkittäviä eroja kuitenkin löytyi. Kaiken kaikkiaan tulokset olivat linjassa odotusten sekä ihmisen tilakuulosta tehtyjen empiiristen tutkimusten kanssa, minkä perusteella voitiin olettaa että kokeessa käytetty testiympäristö todennäköisesti soveltuisi vastaavanlaisiin kokeisiin myös tulevaisuudessa.

ASIASANAT:

Tilaääni, immersio, kuulo, äänenkäsittely, virtuaalitodellisuus

CONTENT

FIGURES

TABLES

1 INTRODUCTION

Spatial audio, here understood as the set of techniques that make it possible to create the impression that there are sounds coming from different locations in space around the listener, has been gathering interest in recent years. While the underlying ideas are not new, for many decades the lack of practical use cases has kept the technology from gaining widespread attention outside academic research and a few niche applications. It is largely the recent developments in virtual reality technology that have created a strong demand for it, as the games and other interactive applications built on virtual reality platforms are generally designed to immerse the player in the virtual environment, and in evoking this sense of immersion audio plays a large role.

Because the adoption of the new virtual reality devices has happened relatively quickly, it is not surprising that multiple software solutions for implementing spatial audio exist, each with their own interfaces and customization options. Not much objective and in-depth research has been done to ascertain why the developer should choose one spatial audio solution over the other and with what settings would the most realistic or the most immersive results be achieved. The information currently available leaves many questions unanswered. As there are no clear guidelines for the developers to follow, they might end up selecting settings that produce unnatural acoustic phenomena in the virtual auditory space or settings that do not produce a sense of space at all.

The goal of this thesis is to partially contribute to the solution of the problem of finding the optimal method of implementing spatial audio in a game engine for a first-person game or any other immersive experience. Due to the variety of options and the complexity of the problem, the focus is limited to one piece of software and one specific set of variables that contributes to the overall spatial hearing experience, namely the listener directivity pattern of the Resonance Audio spatial audio plugin. Equivalent settings are found in rest of the currently available spatial audio plugins as well, so the findings of the present study are also more generally applicable.

A virtual test environment was designed and implemented to compare the effects of the directivity pattern with different parameters. In this test environment an experiment was conducted where test subjects would assess the directions of sounds and also rate the perceived realism of the listening experience. One goal of the experiment, in addition to finding the optimal values for the directivity pattern, was to test whether this test setup

would bring out significant differences between the chosen parameters and thus be suitable for expanding this kind of research in the future. The results of this and future experiments will be utilized by the commissioner of the work, Turku Game Lab, in developing a smooth and effective workflow for implementing immersive audio in first-person games and virtual reality applications.

In order to analyze the realism of spatial audio in the virtual environment, it is first necessary to discuss how and how accurately spatial hearing works in the real world. The literature on the subject is extensive, spanning the entire 20th century and beyond. What is most relevant from the standpoint of simulating the spatial hearing experience is how the sound as it arrives at each ear differs from the sound as it emanates from the sound source, and this is also the aspect of spatial hearing primarily concentrated on in this thesis. The data on the accuracy of human spatial hearing in different conditions and for different types of sounds is also of considerable importance, as it can be used to check if the directional estimations in the virtual environment are in line with what they would likely be in a similar situation in the real world.

While the experiment dealt with both the test subjects' perception of direction and their subjective experience of realism, the main point of interest was the realism, and as it turned out, it was in the realism ratings that the most apparent differences between the tested parameters manifested themselves. It should be noted that while realism and the ensuing sense of immersion are often the primary goal in audio design for video games and other virtual experiences, in some applications the precise directivity of individual sound sources may be of greater relative importance. Finding the optimal spatial audio solution in terms of directivity is just one of the many possibilities for future research in this area.

2 THEORETICAL AND EMPIRICAL BACKGROUND

2.1 Spatial hearing and spatial audio

Spatial hearing refers to our ability to locate the sounds that we hear from around us. While this information is often important in itself and as a cue for orienting our visual attention, we also use it to separate sounds coming from different locations when focusing our attention to a particular sound source, for example when having a conversation in a room full of people or listening to a particular instrument in an orchestra [1]. Spatial hearing allows us not only to identify the positions of individual sound sources but also to sense the whole field of sound surrounding us as a whole.

Spatial audio can be defined as the set of techniques that attempt to reproduce a similar hearing experience as in a real-world three-dimensional environment. This is generally done by simulating how sound waves coming from a particular position in relation to the listener would travel in space before reaching each ear. Acoustic calculations are used to assign individual sound sources virtual locations that together form a virtual auditory space [1]. The listener will get the impression of being surrounded by sounds coming from specific directions and distances around them.

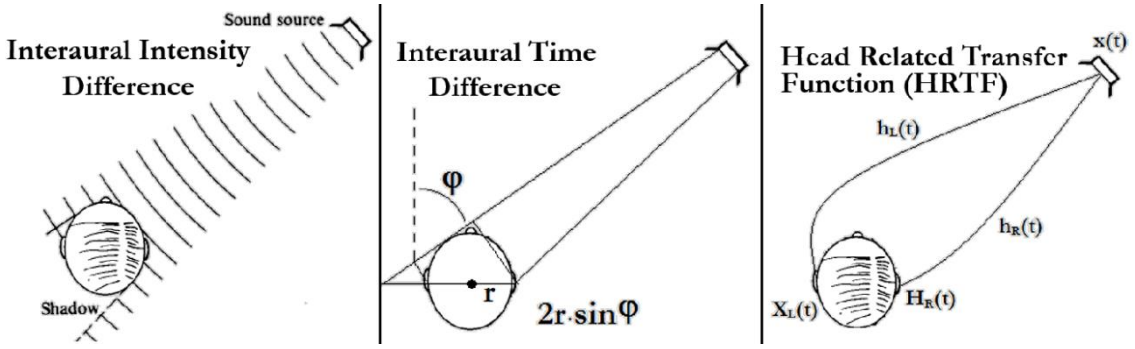

Figure 1. Interaural time and level differences are the most important cues for determining the position of a sound source. Head-related transfer functions are used to simulate these differences and create an impression of spatially localized sounds.

The main cues we use to determine the direction of a sound are time and level differences between the two ears. Therefore one of the most important techniques in audio spatialization is the manipulation of these differences. (See figure 1.) Data applicable for this purpose has been collected since as early as the 1920s from dummy

head recordings, where two microphones are placed within the ear canals of an artificial head. However, all dummy head systems up until the KEMAR model introduced in 1972 have been observably imperfect in reproducing the original auditory space. In general, best results have been achieved with models that seek to accurately replicate not only the shape of the human head but also the inner parts of the ear. [2] From these findings it can be concluded that even minor alterations to the signals sent to the ears can significantly affect the sensation of spatial hearing.

Because the interaural time and level differences are so slight, accurate spatialization can only be achieved by controlling the input to each ear independently. In practice this implies the use of headphones. [1] The problem with loudspeakers, whether in a stereo or a surround setup, is that each ear will hear the sound from both or all speakers, whereas when using headphones the two ears are largely isolated. Although solutions for eliminating the crosstalk between loudspeakers do exist, perhaps the most notable of them being a process called TRADIS (True Reproduction of All Directional Information by Stereophony) developed in the early 1970s [3] [4], they require special compensating circuits in the audio chain and have really only been utilized in laboratory conditions.

A further limitation of stereo or surround sound through loudspeakers is that commonly all of the speakers are more or less on the same level, which means that all of the sounds coming from them will be locked to a flat plane. Representing the full sphere of sound would require an impractically complex speaker array, and the listener would still be bound to the single optimal listening position (the "sweet spot") in the middle of them. [5] These are some of the reasons why spatial audio solutions aiming for accuracy and physical realism focus almost exclusively on headphones.

2.2 Directivity and immersion

The objective of spatial audio can be seen as twofold: firstly to present each individual sound to the listener as if it was coming from a specific direction, and secondly to utilize the totality of all of these sounds together to generate an impression of a virtual auditory space and a sense of being within that space, a sensation that is difficult to rigorously define but is perhaps best captured by the concept of immersion. These two components, directivity and immersion, are very much linked. The immersion and the sense of space can be seen as the result of the directivity of each individual sound.

Let us consider a virtual reality environment experienced through a virtual reality headset and headphones. Whether the objective of a VR application is to generate a virtual representation of an existing real-world environment or to create a whole new world of its own, the experience generally relies on giving the user the illusion of being transported into the virtual space. This is done by isolating the senses of vision and hearing from the real world and replacing everything the user would see and hear with virtual content. The more vivid the virtual content, the more effective the experience. Conversely if the user feels that something, be it visual or auditory, is missing, the experience as a whole will be less impactful. [5] Especially if the visuals don't seem to match with the audio, the immersion is inevitably broken. This is why in any ambitious VR project it is crucial to get the spatial audio right.

Undoubtedly the recent increase in popularity and affordability of virtual reality headsets has also contributed to the growing effort put into spatial audio by game and software developers. If we look at the history of video games, we can see a huge progress from the primitive pixel and vector graphics of the early arcade games to the almost cinematographic realism of today. Not only that, but the steps up in graphics between different gaming console generations, for example, are often significant and easily identifiable. Simultaneous advances in the field of audio have been slower and also less noticeable. In fact the type of simple stereo panning used in games such as Wolfenstein 3D in the early 90s is still not uncommon especially in independent and hobbyist projects.

While all first-person games by definition present the game world as if through the player's eyes, the virtual reality headset that enables the player to freely look around in the game world by moving and turning their head adds a whole new level of immediacy compared to the traditional PC and console games viewed through a regular monitor. It is only natural that as the game's visuals follow the movements of the player's head, the player would expect the game audio to do the same. When the player turns their head up or down, for example, there should be a noticeable change in the surrounding audio. Without a spatialization method that takes into account the full sphere of sound, there is no such change. This is why in virtual reality the difference between left and right panning and fully spatialized audio is especially evident, which in turn might explain why the recent introduction and adoption of consumer VR headsets such as HTC Vive and PlayStation VR has coincided with a sudden appearance of multiple competing solutions for implementing spatial audio. While regular PC and console games may be able to get away with taking some shortcuts audio wise and still keep the player immersed in the

experience, VR almost seems to demand a step up in realism in audio as well as in graphics.

Largely because of the way virtual reality as a media naturally highlights the importance of both immersion and correct localization of sounds, the practical part of this thesis has its focus on an experiment performed in a virtual reality test environment. Many if not all of the findings, however, can be applied more generally to any media that seeks to evoke in the player, viewer or listener a sense of being in a virtual auditory space, whether that space be within the world of a game or a movie or, for instance, the venue of a musical performance.

2.3 The psychophysiology of spatial hearing

2.3.1 General remarks

Spatial hearing is a rich and interdisciplinary area of research with relevant studies coming from fields such as psychology, psychophysics and physiology on one hand and engineering, physics and musical analysis on the other [2]. Within the scope of this thesis it is only possible to scratch the surface, but at the very least a brief introduction of the principal aspects that should be taken into consideration when attempting to replicate the effect is necessary. For readers interested in a more thorough commentary, *Spatial Hearing: The Psychophysics of Human Sound Localization* by Jens Blauert [2] is a comprehensive reference work on the fundamentals of the subject with an extensive bibliography encompassing literature from all of the aforementioned fields of study. A more recent and more general text on hearing, *The Sense of Hearing* by Christopher Plack [1], also deals with spatial hearing at length, primarily from a psychophysiological perspective.

As stated previously, the auditory system relies primarily on interaural time and level differences in establishing the location of the sound source. The relative importance of different attributes of the sound signal, however, varies significantly depending on factors such as the direction and distance of sound source, the waveform and spectrum of the sound, and even psychological matters such as the listener's familiarity with the sound in question. Interaural level differences, for example, are greater for sounds that contain mostly high-frequency components [1] [2], while for sounds that contain a fair amount of

low-frequency components (as most natural sounds do) interaural time differences are more pronounced [6]. Therefore the auditory system uses the combination of time and level differences together to localize sounds across the whole range of audible frequencies, with interaural time differences being of primary importance in most situations [1].

Depending on the type of sound and the environment it is heard in, the perceived sound may be more or less precisely localized. Both the location and the extent of a single click in an anechoic chamber, for example, can be determined very accurately, while for a sustained tone in a reverberant room neither can be established with any precision. [2] In general the ideal accuracy is attained in the case of a distinct sound in an open area where no reflections occur, while multiple simultaneous sounds, whether they be from different sound sources or reflections originating from a single source, tend to hinder our spatial hearing ability. Hence the following chapters concern primarily single sound sources in anechoic (i.e. free field) circumstances.

2.3.2 Spatial hearing on the horizontal plane

The auditory system functions somewhat differently and with different accuracy on the horizontal plane than on the median plane (i.e. the plane that bisects the body vertically into left and right halves, see figure 2). Generally speaking, it is much easier for us to determine the general direction of the sound on the horizontal plane than the elevation on the median plane or, for example, whether a sound is coming directly from the front or directly from the back. Hence it makes sense to consider these two planes separately.

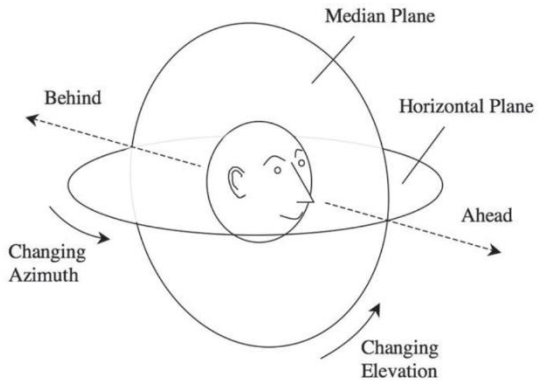

Figure 2. The horizontal and median planes. Human spatial hearing is generally very accurate on the horizontal plane but not so accurate on the median plane. Figure from Plack, Christopher. The Sense of Hearing. 2005.

Discussions on the accuracy of spatial hearing commonly make use of the concept of localization blur, which refers to the smallest displacement of a sound sufficient to produce a change in the perceived direction of said sound. On the horizontal plane spatial hearing is most accurate in the forward direction, where the localization blur is about one degree [2]. At right angles to the listener (i.e. directly on the left or directly on the right) the localization blur increases to between three to ten times its value for the forward direction, while behind the listener the localization blur again decreases to roughly twice the value for the forward direction [7] [8] [9] [10] [11] [12] [2]. It is also worth noting that localization blur varies for different types of sounds such as impulses, sinusoids, noise, and for different frequencies. For impulses and narrow-band signals the localization blur tends to be somewhat lower than for broadband noise. [2]

Spatial hearing on the horizontal plane relies primarily on the time difference between the ears. The difference is maximal directly to the left or to the right and minimal (zero) directly to the front or to the back. There are therefore multiple positions where the time (and level) differences are the same. Often it is our sense of vision that helps us differentiate between these positions. If we see the sound source in front of us, for example, we know the sound is coming from the front, whereas if we don't see the sound source, we will likely assume the sound is coming from the back.

In the lack of visual cues, the ambiguity of the direction of the sound source may also be resolved by rotating the head. If the sound is coming from the front, turning the head left will decrease the sound level in the left ear and cause the sound to arrive to the right ear first, while if the sound is coming from the back, the opposite will happen. If turning the head to either side has no effect, the sound is coming from either directly above or directly below. [1]

In addition to these special cases where both ears receive the exact same sound signal, interaural time and level differences are also equal in directions axially symmetric with the axis passing through the ears, and between these directions as well a similar confusion may arise. For instance, a sound coming from an angle of 45 degrees (i.e. from the front right) may be perceived as coming from an angle of 135 degrees (i.e. from the rear right) [13] [9] [11] [14]. This happens especially with narrow-band signals, because they lack the spectral information that the auditory system normally uses to tell such symmetrical directions apart [2]. The presence of higher frequency content generally makes differentiating between these directions easier [5].

The errors or imprecisions in localization mentioned above rarely occur if the duration of the sound is long enough for the listener to move their head around and listen to the sound from different directions [2]. It can therefore be concluded that as a whole spatial hearing on the horizontal plane is fairly accurate, as in localization blur often being limited to just a few degrees, for most natural sounds and in most situations, with significant errors being more of an exception than the rule.

2.3.3 Spatial hearing on the median plane

Spatial hearing on the median plane is very much different than on the horizontal plane. First of all, in all directions on the median plane the interaural time and level differences will be zero [1]. In other words, the differences that function as the main cues for direction on the horizontal plane do not exist at all on the median plane. The information that the sound signal arrives identical at both ears may help us recognize that the sound source is located somewhere on the median plane, but it provides no clue about the elevation of the sound source or whether the sound source is in the front or in the back.

Perhaps unsurprisingly, then, the accuracy of spatial hearing on the median plane is much lower than on the horizontal plane and varies greatly depending on the characteristics of the sound. For human speech, for example, the localization blur for changes in elevation can be as high as 17 degrees [15] [2]. For sounds narrower in bandwidth and sounds less familiar to us the accuracy will generally be even worse. As an extreme example, for signals with bandwidth less than two thirds of an octave, it is in fact impossible to determine either localization or localization blur on the median plane; there simply is no correlation between the direction of the sound source and the perceived direction of the sound [16] [2].

Interestingly as the direction of the sound source loses its significance in regard to the perceived direction of the sound, the frequency of the sound takes on a more pronounced role. Multiple experiments with different experimental setups have confirmed that tones that are higher in pitch are also perceived as being positioned at a higher angle, and vice versa [17] [18] [19] [20] [2]. In the extreme case it is not the direction but the frequency alone that accounts for the perceived direction of the sound source [2].

In conclusion, while our spatial hearing ability is reasonably accurate and reliable on the horizontal plane, on the median plane it tends to be considerably inaccurate and even deceptive.

2.3.4 Distance hearing

Finally, for the auditory system to precisely locate a sound, directional hearing must be complemented by distance hearing. The main attribute that the auditory system uses to establish the distance of the sound source is the pressure level of the sound signals arriving at the ears. Distance hearing relies heavily on the listener's familiarity with the sound, because there's no way to deduce the distance of the sound source from the level of the sound signal unless that level is known at some reference distance [2]. In general, we tend to perceive loud sounds as being close and quiet sounds as being distant, but these perceptions aren't always reliable. An unfamiliar nearby sound may sound like it is far away when in fact it is just quiet, while a distant but loud sound is likely to be heard as being closer than it actually is. [1]

For familiar sounds such as regular human speech the perceived distance corresponds fairly well to the actual distance of the sound source, but already for less usual types of speech such as shouting and whispering we begin to make mistakes [21] [2]. When evaluating the distance of such somewhat familiar sounds it is common for the listener to perceive the sound as being closer than it actually is [1]. In the case of completely unfamiliar sounds, the perceived distance does not as a rule correlate with the actual distance of the sound source at all. Particularly for narrow-band sounds, and at distances greater than three meters for other sounds as well, the perceived distance depends solely on the loudness of the sound, not on its distance. [2]

There are a few different attributes of sound signals arriving at the ears that depend on the distance of the sound source and that the auditory system uses in evaluating the distance of a sound. For sounds at a distance of roughly three to fifteen meters the distance is determined by the sound pressure level alone [2]. For sounds in this distance range the perceived distance tends to be less than the actual distance. It is also worth noting that the sound pressure level does not only affect the perceived distance but also the perceived loudness and tone color of the sound. As the signal level increases, low-frequency components gain more weight relative to high-frequency components and the tone color becomes darker. [2]

At distances greater than approximately fifteen meters the sound pressure level is still of major relevance, but attenuation caused by air absorption also plays a role. Because air absorbs more high-frequency than low-frequency energy, high frequencies of a sound travelling through air will be attenuated more than low frequencies and the spectral balance of the sound as it is heard will be biased towards low frequencies. This is the reason why distant rolls of thunder or ocean waves have a deep sound. However, the effect is slight and really only becomes noticeable at distances of hundreds of meters. [2] [1]

Finally, for sounds closer than three meters, the pressure level is again the main cue for distance, but the curvature of the sound waves as they strike the listener's head will also have some effect on how the spectrum of the sound changes with distance [2]. Naturally this phenomenon becomes more relevant the closer to the listener the sound source is.

2.3.5 Multiple sound sources and disturbed sound fields

Everything that is said above applies as such for single sound sources in free field. In the case of multiple simultaneous sounds or a single sound in a reverberant environment the situation is a bit more complicated. While from the standpoint of physics the signals arriving at the ears in an environment with multiple sound sources can be conceptualized as a superposition of all the individual sound signals from different sources, the combination of sounds as the auditory system evaluates it is not a superposition of how each of the individual sounds would be evaluated on its own [2]. Thus, the auditory system is unable to simply separate the total sound into the individual components to the determine the positions of each sound source, but instead takes cues from additional phenomena that only occur when there is more than one sound source.

The exact mechanisms for identifying the number of sound sources and differentiating between them are complex and a matter of ongoing research. In addition to time and phase differences, properties of sound such as harmonicity help us determine which frequency components belong together and form a single sound.

Disturbed sound fields, such as in enclosed rooms with obstacles and reflecting surfaces, may be considered a special case of the more general problem of multiple sound sources. The difference is that once the auditory system recognizes the reflections as originating from the same source it focuses on the primary sound signal and largely

ignores the reflections. [2] The primary signal is distinguished from its reflections by something called the precedence effect: Because the path that the reflected sound travels is always longer than that of the direct sound, the direct sound will always arrive at the ears first [22] [1]. The position of the sound source is then determined primarily by the components of the direct sound signal [2].

When the delay between the primary and the secondary signal is long enough, the latter will be perceived as a distinct echo. The length of the delay sufficient to produce an echo effect depends on the type of sound. For clicks and other brief impact sounds it can be as short as two milliseconds, while for speech it is closer to twenty. If the delay is very long, the latter signal is no longer perceived as an echo of the first one but as another, completely independent sound. [2]

While reflections and reverberation are of little utility in and can even be an impediment to localizing sound sources, they provide essential information for the auditory system to form an impression of the dimensions and characteristics of the surrounding space. This is why in music production, for example, it is very common to apply reverberation and delays to studio-recorded audio in order to simulate a more natural listening environment. By varying the amount of reverb and the length of the delays, impressions of different types of environments may be created. [1]

2.4 Simulating the spatial hearing experience

2.4.1 Spatial audio formats

Channel-based audio

The most common and most simple audio format that gives the listener some sense of space is the standard stereophonic audio. While in monaural audio all sounds appear to emanate from a single position, in stereo sounds may be localized anywhere on the straight path between the two speakers. The stereo effect is often more effective with headphones, because the position of the listener and the acoustic properties of the room don't affect the listening experience [5].

Multi-channel surround sound such as 5.1 or 7.1 surround can be considered an extension of the stereo speaker layout. Sounds may now be heard from any direction

around the listener, but because (at least in all common home setups) all of the speakers are situated more or less on the same plane, there is still no sense of elevation. [5] A further limitation of surround sound is that it is tied to a specific arrangement of a specific number of speakers. Just as a stereo recording is mixed into two channels to be listened to either on stereo speakers or on headphones, a surround sound mix is only suitable for the particular speaker setup it is designed for.

Both stereo and surround are channel-based audio formats. This means that the audio is mixed into a fixed number of channels that, when the audio is played back, are routed to specific speakers in the speaker array. [5] The speaker setup is usually designed so that the listener should be sitting somewhere in the middle to get the ideal listening experience. To lessen the impact of this restriction, audio in many of today's movies is produced so that the "sweet spot" in cinemas would be as broad as possible [5]. Even so the spatial impression for the audience near the edges of the room is not exactly how it should be.

Binaural audio, although arguably more of a recording technique than a format of its own, is also worth mentioning in this context. Meant to be listened on exclusively on stereo headphones, binaural recordings are usually done on a special binaural microphone that mimics the shape of and the distance between the ears. The microphone setup is similar as in the dummy-head recordings commonly used in audio research. The microphone captures the soundscape around each ear, and when the sound is played back through headphones it gives the listener quite a natural spatial hearing impression. The technique was invented in the late 1800s and has been in use since then, but it has never gained considerable popularity among mainstream consumers [5].

While channel-based audio can be suitable for situations where the listener is expected to sit still, such as when watching a movie or listening to a record, its limitations clearly manifest themselves when it comes to immersive media such as VR where it is often an essential part of the experience that the subject may freely move around in the virtual environment. If the surrounding sounds don't change as they should when the subject moves or rotates their head, the immersion is lessened. This is why VR needs a spatial audio solution that can present the full sphere of sound for any position and any rotation of the listener. That solution is ambisonic audio.

Ambisonic audio

Ambisonic audio, or ambisonics, was developed in the 1970s, but there weren't really any apparent applications for the technology at that time. It was primarily used by audio researchers until recently when developments in VR technology sparked a demand for such a flexible spatial audio format. [5]

The number one benefit of ambisonics compared to channel-based formats is that it represents the full sphere of sound, which can theoretically be decoded and routed to any number of speakers. In practice the sense of space comes across the best through stereo headphones, especially in the case of VR applications, 360-degree video and other forms of media where the listener may turn their head to listen to the surrounding sounds from different directions.

Ambisonic audio is typically recorded with a special microphone with four or more microphone capsules pointing in specific directions. The capsule arrangement varies by microphone, but what is essential is that together they capture the surrounding soundscape from all directions. Four capsules is the minimum, but more may be added to increase the resolution of the recording. Such higher resolution ambisonic recordings are called second order ambisonics (eight capsules), third order ambisonics (sixteen capsules) and so on [5].

The audio as it is recorded is initially in a so-called A-format, which is then encoded, often with microphone-specific software, to the standard B-format. The B-format has four channels: X, Y and Z for the three dimensional axes, and W for capturing omnidirectional sound pressure. Finally, when the ambisonic recording is being listened to, the B-format is decoded to a suitable format, in most cases binaural audio for stereo headphones. In the past the decoding was done on dedicated hardware devices with circuits designed specifically for this purpose, but modern home computers are powerful enough to easily handle it in real time in software. Ambisonic audio is therefore suitable and often used not only in VR but also in video games in general, both on PC and on consoles.

2.4.2 Spatial audio in game engines

Basic stereo panning

The majority of interactive projects making use of spatial audio, whether they be VR applications or traditional PC games played on a regular monitor, are built on general-purpose game engines such as Unity or Unreal Engine. Hence it is appropriate to briefly examine how spatial audio is generally implemented in game engines.

The most primitive method of spatialization, and also the most common, is simple stereo panning. The game engine calculates the distance between and the relative rotations of the sound source and the player and uses that information to set the appropriate volume level for the left and right channels. Other than through changes in distance, differences in elevation between the sound source and the player have no effect on the sound as it is heard by the player. Thus, it could be said that the spatialization essentially takes place on the horizontal plane only. Furthermore, the frequency spectrum of the sound isn't affected by the positions or the rotations of the sound source and the player in any way; it is only the general volume level per each channel that changes. Acoustic phenomena such as reflections and attenuations caused by walls and other physical objects are usually not simulated either, so, when wearing headphones, a sound coming directly from the left will only be heard in the left ear.

Although recently the shift towards more realistic audio has begun to gain speed, in many popular game engines such as Unity stereo panning is still how the sound system currently works by default. This is also the case with external audio libraries such as OpenAL that are commonly used in conjunction with a separate rendering engine by developers wanting more flexibility and control over the entire codebase than the unwieldy and often proprietary general-purpose engines provide. Consequently, a fair amount of today's game projects, especially by independent and hobbyist developers, still settle for an audio solution that really hasn't changed much in decades.

Full 3D spatialization

While there are signs that full 3D spatialization will most likely be built into all major game engines in relatively near future, the current situation is that there are multiple competing

spatial audio implementations for developers to choose from. Examples include Resonance Audio, Steam Audio, Oculus Audio and dearVR. Many of these provide prebuilt plugins for popular game engines such as Unity or Unreal Engine as well as for audio middleware such as FMOD or Wwise alongside a standalone software development kit for any other use cases.

There are many features and details in which the available solutions differ from one another, but the fundamentals of the spatialization itself are without exception implemented by making use of head-related transfer functions, or HRTFs. The HRTF is a function that describes how a sound from a specific point in space will arrive at the ear. (See Figure 1.) From a signal processing standpoint, it can be considered a filter that is applied to the initial audio signal before sending it to the headphones. For spatial audio a pair of HRTFs is required, one for each ear. For each position of the audio source in relation to the listener there is a specific pair of HRTFs, and when that pair of HRTFs is applied to a sound signal, the listener will get the impression that the sound is coming from that particular position.

In reality, HRTFs vary slightly from person to person due to differences in the shape of the head and the ears. In practice, HRTF-based spatialization software utilizes some general set of data that can be expected to give reasonably good results for the average person. The process of spatializing a sound essentially consists of calculating the position of the audio source relative to the listener and selecting from a spatially sorted set of HRTFs the most appropriate pair of functions to apply for that position.

A fair amount of HRTF data is freely available in the standard SOFA (Spatially Oriented Format for Acoustics) format [23]. The majority of it has been collected from dummy head recordings where a sound source is orbited around a dummy head at a number of different elevations (and, in many cases, different distances) and the impulse response at each ear is measured and recorded for each position of the audio source. The MIT-KEMAR data set is widely considered the reference, but there are also some other ones generated with different equipment that are either higher in resolution as a whole or more focused on some specific area such as near-field or behind-the-ear audio.

Since there is a range of options, it is not surprising that not all spatializers use the same HRTF data. In free and open-source implementations the source of the data is usually clearly visible in the code. Resonance Audio, for example, uses HRTFs from the SADIE database [24]. Some spatializers like Steam Audio give the user the option to replace

the default HRTFs with a SOFA file of their own choosing. In the case of completely proprietary spatializers like dearVR, the user cannot know which HRTFs are being used or if any other processing takes place beyond the HRTF filtering.

In addition to HRTF spatialization, practically all of the spatial audio solutions also simulate the reflection and absorption of sound waves in the environment, which is something that gives the listener a sense of the dimensions and materials of the space and substantially contributes to the immersion. Even though all spatial audio implementations simulate more or less the same set of acoustic phenomena, there are considerable differences not only in the accurate and the complexity of the simulation itself but also in the user interface and the variety of options and adjustable parameters available to the user. For instance, in one plugin the user may be free to create and define the acoustic materials used in the simulation, while in the other there is only a limited number of predefined materials such as wood, concrete and glass to choose from.

The different design decisions are partly explained by the fact that realistic acoustic simulation is simply too resource-heavy to do in real time without greatly simplifying the physics involved with it. To simulate the obstruction and occlusion of sound by environmental objects, for example, we need to keep track of everything that may affect the path of the sound and constantly check by ray tracing how much the sound should be attenuated and how the different materials in its path should shape its spectrum. In a fully realistic simulation this would be the case not only for every direct sound in the environment but also for the multitude of their reflections. [5]

In practice, when the simulation has to be run in real time, as is the case with games and interactive application where both the player and the sound sources may move around in the environment, the reflected sounds aren't usually simulated as accurately as the direct sound. The general acoustic characteristics of the space are often conveyed to the player by a predetermined reverb effect that is applied to all of the sounds that the player hears and that may only change when the player moves from one room to another or from indoors to outdoors. However, deciding which compromises in realism should be made to keep the acoustic simulation light enough is a complex problem with no obvious solutions.

3 FINDING THE OPTIMAL SPATIAL AUDIO SOLUTION

3.1 The scope of the problem

From the perspective of the developer it would be extremely useful to know if one of the many available spatial audio solutions is better than others in replicating the spatial hearing experience and creating a soundscape that feels natural and immersive. Unfortunately, although the adoption of spatial audio plugins appears to be on the increase, not many extensive comparisons between them have been done. One of the only examples of quantitative research on the topic, the master's thesis of Veli Laamanen from 2018 [25], focused only on the spatializer part of each plugin with the rest of features disabled. The less scientific personal comparison done by the sound designer Chris Lane for Richard Gould's *Designing Sound* online publication on the same year [26] did touch on the acoustic simulation side as well, but none of the tested plugins stood out as the clear winner. Furthermore, as Lane admits, not all of the features are directly comparable between different implementations, and many of the more specialized capabilities such as the near-field effect or simulation of air absorption are not present in all of the plugins.

While determining which of the available plugins has the potential to produce the best results is a challenge in itself, the matter is further complicated by the fact that due to the variety of options and parameters it's not even obvious how to get the most out of any one of the plugins in respect to itself. Although each plugin comes with documentation that typically introduces the main concepts and goes through the essentials of the typical workflow [27] [28] [29] [30] [31], there is surprisingly little information that goes into the specifics of every individual setting. Even reference manuals that do list every function of the software rarely explain all of the parameters in sufficient detail. It is not uncommon for a component of the sound system to have an adjustable parameter with some seemingly arbitrary minimum and maximum values and no unit of measurement visible to the user. For some parameters the effect on the sound may either be described in the manual or simple enough to intuitively figure out, but the rationale behind the default value is not explained and recommendations on which value to select to get the most realistic or most immersive listening experience are absent. The developer is left guessing which settings might best reproduce the real-world spatial hearing experience, and for what reasons and in which situations, if any, they might want to diverge from

them. Even outside official documentation, on online forums for example, such discussion is nowhere to be found.

The initial plan for this thesis was to compare the different spatial audio plugins and their adequacy in producing realistic and immersive audio. However, for reasons mentioned above, it soon became evident that this is too complex of a problem to solve at once. It was decided that it would be more fruitful to focus on one plugin only and limit the comparable variables to the ones with presumably the most noticeable effect to the listening experience as a whole. Any information attained about variables of such fundamental significance would in all likelihood be applicable not just for the one plugin used in the test but for many of the other ones as well. Attention was also put into designing the test environment so that it could be used to run similar experiments on other parameters in future research.

3.2 Choice of plugin and comparable variables

Out of the currently available and maintained spatial audio plugins, Resonance Audio was chosen as the one to be used in the test. Other potential candidates included Steam Audio, which at the time of writing is still in the beta phase of development and at the time of testing had some major issues with feedback and real-time performance, and dearVR, which was mainly rejected for licensing reasons and due to the full source code not being provided with the package. A major plus of Resonance Audio is that it's completely free software under the permissive Apache 2.0 license. From the research point of view this is of vital importance because it is not possible to thoroughly interpret and apply the results attained with one particular piece of software on a more general level unless that piece of software be completely transparent in its implementation. For developers the freedom of the source code is a valuable benefit as it allows the adaptation of the software for a variety of environments and is consequently likely to prevent it from becoming obsolete anytime soon. At present, Resonance Audio is in active development but already stable enough for production use, a status that was confirmed in the comparative tests performed during the planning phase of this research where all of its features appeared to function as intended.

As for the variables to be tested in this research, the two parameters of the listener directivity pattern of the Resonance Audio Source component, alpha and sharpness, were the obvious choices. The listener directivity pattern is what is primarily responsible

for how the sound level at each ear changes depending on the direction of the sound source. It affects the spatial hearing experience by shaping the auditory space whether environmental effects such as reflection and occlusion are enabled or not, and can on that account be conceived along with the HRTFs as one of the elementary layers on which the rest of the acoustic simulation is built.

4 TEST METHOD AND PROCESS

4.1 Test environment

4.1.1 General design

The test environment to be used in the experiment was built with the Unity game engine. The version of Unity used for the project was 2018.2.7 and the version of the Resonance Audio SDK was 1.2.1. The test environment was designed to be used with VR devices. The specific device used in the test was the HTC Vive. The integrated headphones of the Vive Deluxe Audio Strap were used for the audio. The built-in motion tracking of the VR headset provided a simple and intuitive user interface by allowing the test subjects to listen to sounds from different directions simply by moving around and turning their head. VR input and output was handled with the SteamVR Unity plugin and the VRTK toolkit.

The principle idea for the test setup was to have a number of test subjects locate sounds coming from different directions around them and evaluate the realism of the spatial impression. Within the bounds of this broad concept various possible arrangements were devised and experimented with.

In the initial version the subject would use the laser pointer on the VR controller to indicate their estimation of the position of the sound source. The pointer could be extended and retracted with the touchpad button on the controller, allowing the subject to point out both the estimated direction and the estimated distance of the sound source. For each sound presented to the test subject, the position of the pointer, the position of the controller and the position of the sound source were recorded and saved to a file. This data could then be used to analyze the magnitude of error in distance and direction, either in total or along the polar and azimuthal angles separately.

While this solution functioned fairly well for the direction estimates, assessing the misjudgments in distance turned out to problematic. The main difficulty was that because the perceived distance of the sound source is dependent on the loudness of the sound, the amplitudes of the different audio clips used in the test would have to have been adjusted so that none of them would have sounded disproportionately loud or quiet in

relation to the others. Furthermore, the overall volume level of the software affected the estimations as well. The latter problem was mitigated somewhat by beginning the test with a calibration phase where the test subject was instructed to adjust the volume of a reference sound with the touchpad until they perceived it at a given proximity. Even then, the data from the distance estimates could not be confirmed as reliable.

It was decided that recording the estimated direction along with a numeral rating of the realism of the sound would have to suffice. Leaving out the distance evaluation had the positive consequence that the user interface could greatly be simplified. Instead of pointing at the sound source with the controller the subject could now simply face towards the sound source and press a button. The estimated direction was captured directly from the rotation of the VR headset. Since the position and rotation of the controller were no longer relevant, the chance of an error caused by small inadvertent motions of the hand when pressing the button was also eliminated.

In the light of how directional hearing works in real life, it was expected that with realistic spatialization the errors the test subjects would make in judging the direction on the horizontal plane would be fairly small, as a rule not more than a few degrees. Therefore, the parameters with which the directional estimations in terms of the horizontal plane would be close to correct could validly be considered to have resulted in a more realistic directional hearing experience than the ones with which the errors would be more substantial.

When it comes to elevation, the situation is not as simple. As has been stated, human hearing is much less accurate in identifying the elevation of a sound than its horizontal direction. Misjudgments as considerable as tens of degrees are not uncommon, and for some sounds the elevation can't be determined with any accuracy at all. Hence it wouldn't have been possible to reliably predict what kind of estimates from the test subjects would have indicated that the spatialization was functioning in the same way as in the real world. As an example, during the design and development phase of the test environment various natural and mechanical sounds were experimented with, and the sound of birds singing was often perceived as coming from somewhere above regardless of what height the sound source was actually at. It wasn't evident whether this was because of the high pitch of the bird song or because of our tendency to associate birds with the sky. Either way, due to erraticisms of this sort, in the final version of the test environment directional angles were measured in regard to the horizontal plane only.

The last major decision concerned the acoustic characteristics of the auditory space. An indoor environment would have offered the possibility to test how reflections and reverb affect the sense of immersion and the accuracy in determining directions of sounds. On the other hand, it seemed likely that an enclosed space with walls and doorways had the potential to unduly confuse the listener with the option that a sound that is heard from the direction of an open door would in fact be coming from somewhere inside the room. This could have greatly complicated the interpretation of the directional estimations, as some of the listeners would likely have attempted to guess the position of the sound source within the room as others would simply have indicated the direction of the sound as they heard it. To avoid any chance of confusion, an open outdoor area was chosen as the visual setting of the virtual test environment and all of the sounds were played back to the listener as in free field. The screenshot in Figure 3 shows what the landscape looked like. Running a similar experiment in a different type of setting, or perhaps a comparison of the same sounds with the same relative positions in two different environments, is still a possibility for future research.

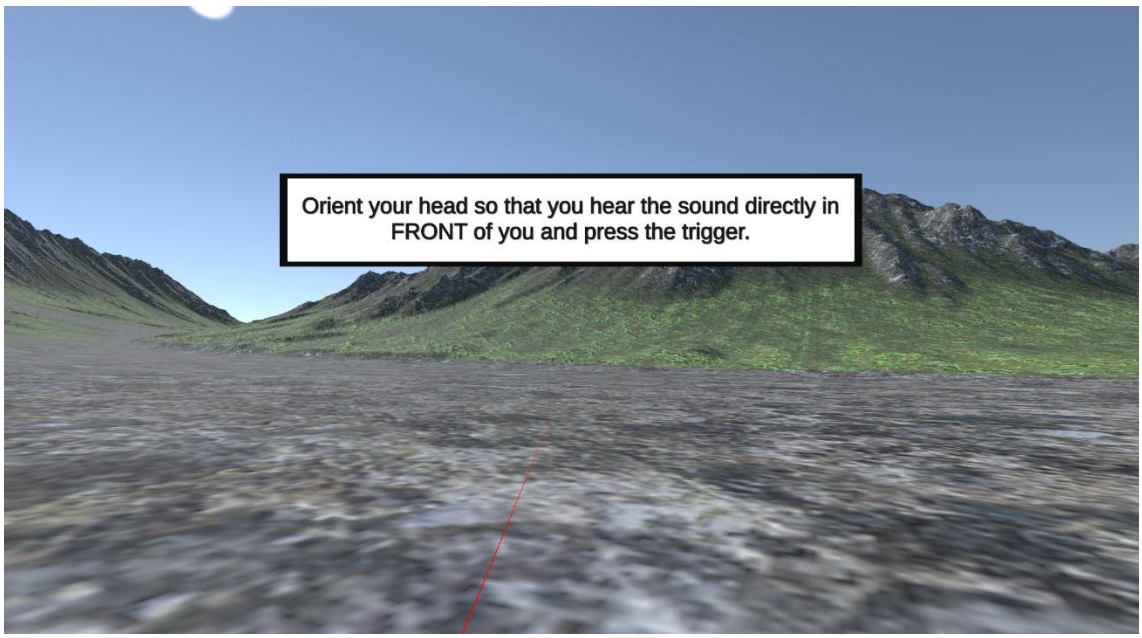

Figure 3. Screenshot from the direction assessment phase of the test environment.

4.1.2 Configuration details

Sounds

Because the waveform and frequency spectrum of the sound were known to affect the accuracy of directional hearing and potentially the perceived realism of the spatial impression as well, a few different types of sounds were selected to be used in the test. It was presumed that with real life recorded sounds it would be easier to judge the realism than with synthesized signals. Three sounds were selected: a meowing cat, a typewriter and flowing water. These three were selected primarily because they were distinctly different from one another in their waveform. The typewriter sound consisted primarily of short impulses of the typist hitting the keys and the hammers striking the ribbon with occasional sounds of the bell and the carriage return. The cat sound was more melodic but with intermittent pauses between the meows. Finally, the flowing water made a steady burbling and splashing sound with little fluctuation in amplitude. The waveforms are depicted in Figure 4.

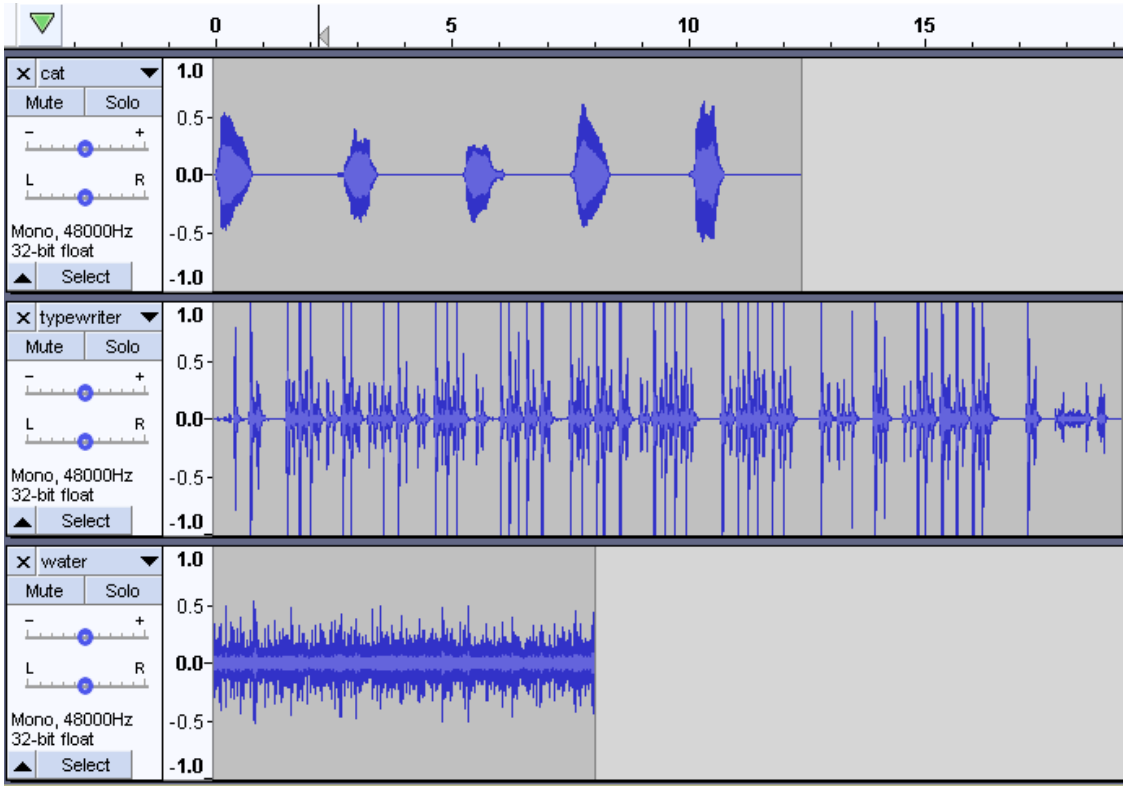

Figure 4. Waveforms of the audio clips used in the experiment. From top to bottom: cat, typewriter, water. The timeline at the top shows the duration of the clips in seconds. Screenshot from Audacity.

Along with waveform the three sounds also differed from one another in spectrum. The sound of the cat's meow was the narrowest in spectrum and mostly centered in the midrange. The typewriter was relatively richer in the high and low ends with a bit of a dip in frequencies from roughly 500 to 3000 Hz. The water sound was markedly the most balanced, having a fairly even amount of components from all audible frequencies. The frequency spectra are depicted in Figure 5.

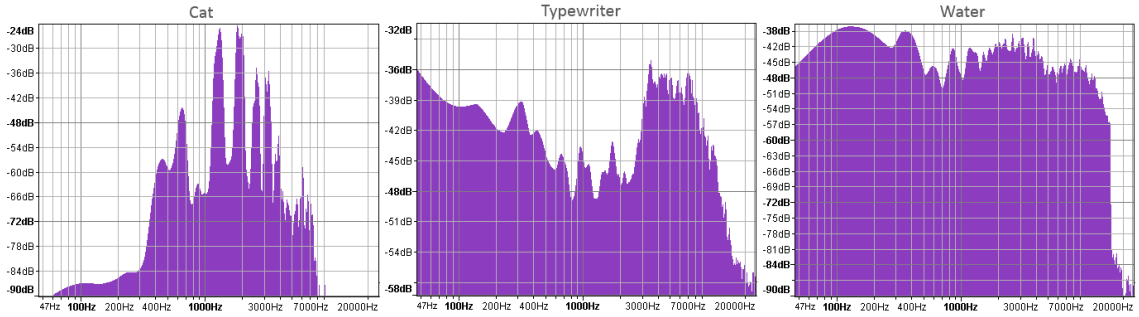

Figure 5. Hann windows of size 1024 of the three audio clips, showing the frequency spectrum from 47 to 20,000 Hz on a logarithmic scale. Screenshots from Audacity.

Settings

The listener directivity pattern of the Resonance Audio Source was tested with three values for alpha (0.3, 0.4 and 0.5) and two for sharpness (1.0 and 1.3). All of the possible permutations of these values were generated at the beginning of the program and shuffled into a random order. This resulted in 18 instances of sounds in total. In addition, there were three dummy sounds, one for each audio clip, that were used in a practice phase that preceded the actual test. The visual representations of the directivity pattern with the different combinations of the alpha and sharpness values used in the test are seen in Figure 6.

With the default values of alpha 0 and sharpness 1 the directivity pattern is perfectly round, meaning the sound will be heard at the same loudness regardless of its direction. Increasing the alpha value up to approximately 0.5 will gradually attenuate the sounds coming from the back so that with the value 0.5 the sounds coming from directly behind the listener will be completely attenuated. When increasing the alpha even further, the pattern will begin to change so that the sounds from the back become audible again but the sounds from the sides are attenuated. With the maximum value of 1 the sounds coming from the front are heard as equally loud as the sounds coming from the back,

while the sounds directly to the left or directly to the right of the listener will not be heard at all.

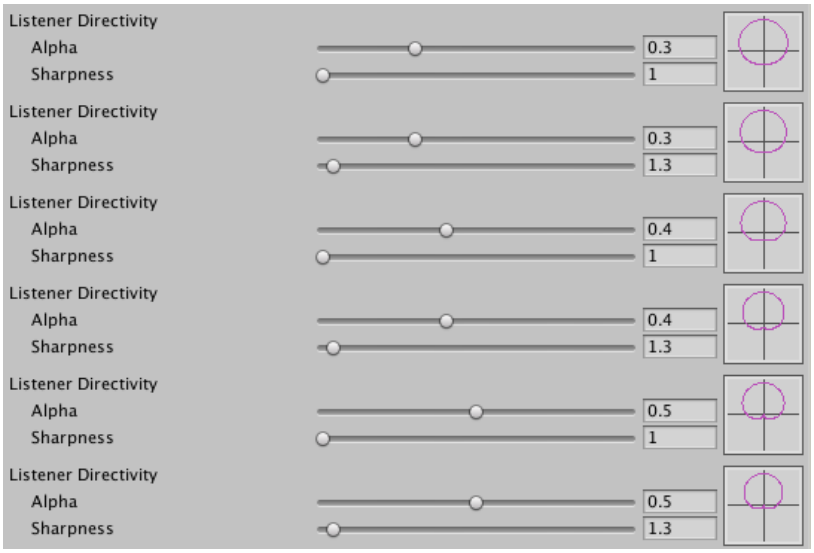

Figure 6. Parameters of the Resonance Audio Source listener directivity pattern.

The sharpness parameter affects the width and intensity of the directivity pattern. Increasing the sharpness will make the pattern narrower and push it towards the front. Both of these effects are somewhat dependent on the alpha value. With the maximum alpha value of 1, adjusting the sharpness will only change the width of the pattern. With the minimum alpha value of 0, sharpness has no effect.

The rest of the parameters for both the Resonance Audio Source responsible for the spatialization and the native Unity Audio Source required for switching between the different sounds and actually playing them were kept constant and largely left at the default values. For the Resonance Audio Source, the parameters of the source directivity pattern were left at the default values of alpha 0 and sharpness 1, meaning that the sound would emanate from the source with equal volume in every direction. Occlusion and near-field effect were left off, and the quality of the simulation was set to "high" for third-order HRTF binaural rendering. The settings of the Resonance Audio Source can be seen in Figure 7. With the exception of enabling spatialization and setting the spatial blend to full 3D, all of the settings of the native Unity Audio Source component were left on default. For volume roll-off the built-in logarithmic curve was used. The settings of the Unity Audio Source can be seen in Figure 8.

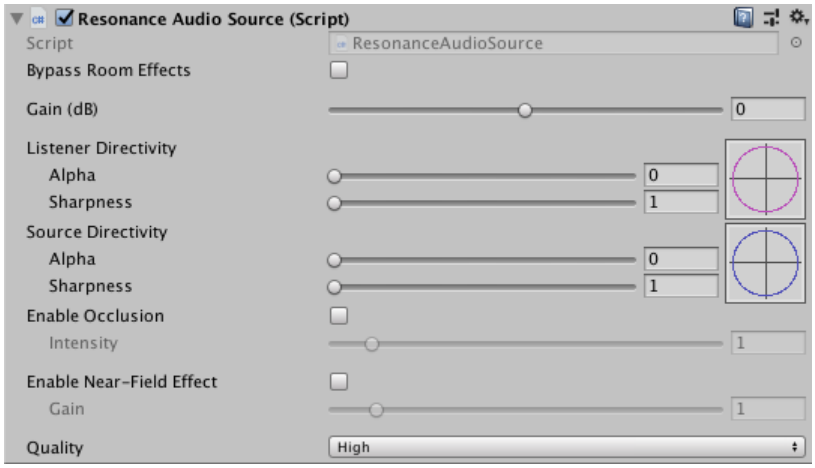

Figure 7. Settings of the Resonance Audio Source component.

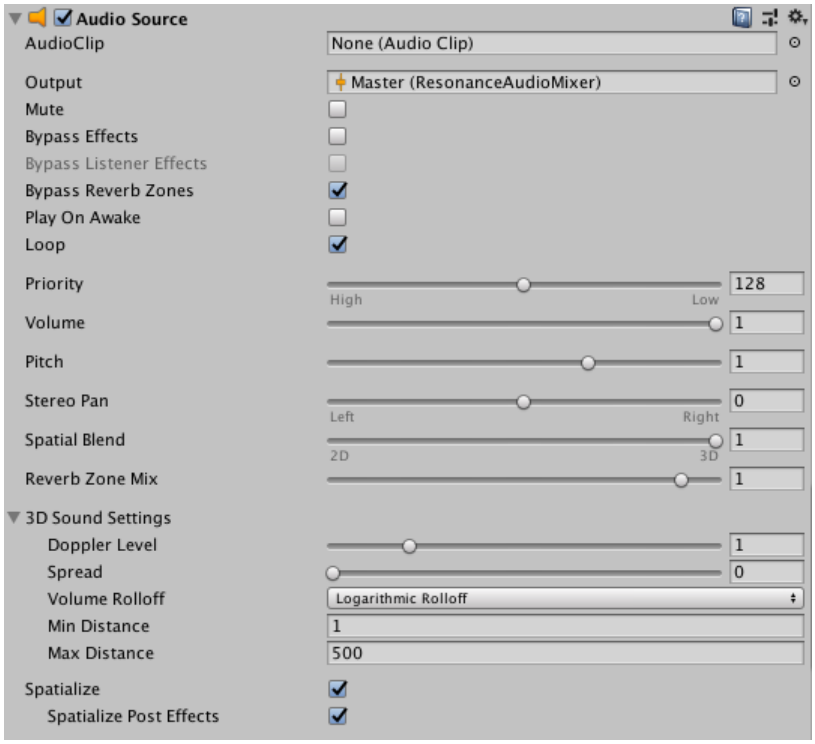

Figure 8. Settings of the Unity Audio Source component.

All of the sounds were programmatically set to start playing at a distance of 3 meters from the listener and at ear level. To accommodate test subjects of different heights, the elevation of the sound source was set to match that of the VR headset. The initial angle of the sound source in relation to the listener was set to a random value between -45 and 45 degrees so that each new sound could start from any direction outside the 90-degree sector directly in front of the listener.

4.2 Test subjects

The test subjects were recruited at random from the students and staff at ICT City campus of Turku University of Applied Sciences. The students were from the information technology and medical fields, while the rest of the test subjects included teachers and administration. There were 20 test subjects in total, with ages ranging from roughly 20 to 50 years. Five of the test subjects were women and fifteen were men.

A hearing test performed in connection with the experiment confirmed all of the test subjects to have normal hearing in both ears. The results can be seen in Table 1.

4.3 Test process

4.3.1 Hearing test

The test was performed in a soundproof studio room. Prior to the test proper that took place in the virtual test environment, the hearing threshold of the test subject was measured with an audiogram. Tones of five different frequencies, 1000 Hz, 2000 Hz, 4000 Hz, 500 Hz and 250 Hz, were played in order to both ears through headphones. For each frequency the intensity of the tone was dropped in steps of 5 decibels to the point of being inaudible to the subject, and raised again by 5 decibels to confirm the previous perception. The lowest intensity detected twice by the subject was recorded as the threshold.

The purpose of the hearing test, in addition to establishing that none of the test subjects suffered from hearing loss that might impede their spatial hearing ability and thus distort the test results, was to find out and make a record of the interaural differences in hearing in case such differences might noticeably affect the directional estimations. It was suspected that considerable differences in hearing thresholds between the ears could have a tendency to skew the directional estimations to the right or to the left depending on which ear is stronger. As can be seen in Table 1, for most of the test subjects the differences turned out to be minor.

Table 1. Results of the hearing test. The table shows the hearing thresholds in decibels for the frequencies on the top row. Values of 20 or below indicate normal human hearing.

	Right					Left				
No./Hz	1000	2000	4000	500	250	1000	2000	4000	500	250
1	0	0	0	0	5	0	0	5	0	5
2	0	5	0	5	10	0	0	10	10	15
3	5	10	5	5	5	5	0	5	5	5
4	10	0	15	10	10	0	5	0	10	10
5	5	0	0	0	0	5	-5	10	5	0
6	10	10	15	20	20	15	10	10	20	20
7	10	10	10	5	5	5	20	5	10	10
8	5	10	5	5	5	5	5	5	5	10
9	10	0	0	0	0	10	5	5	0	5
10	0	0	0	5	5	5	5	5	5	10
11	10	10	5	10	10	10	5	20	10	20
12	10	10	0	5	15	10	10	5	10	15
13	5	0	0	5	15	10	-5	5	0	10
14	15	10	10	10	15	5	15	10	5	5
15	5	5	-5	10	10	5	0	0	10	15
16	5	10	5	5	5	5	0	5	5	10
17	10	0	15	5	10	5	0	10	0	5
18	5	10	10	5	5	5	5	10	5	10
19	5	5	10	5	5	5	10	10	0	5
20	10	15	0	10	15	10	10	0	10	15

4.3.2 Communication

After the hearing test, the test subject was given written instructions for the test proper. Once they had read the instructions, the main points were also relayed orally, and any questions were answered.

The written instructions went through the different phases of the test, explained the user interface of the virtual environment and gave some general guidelines for rating the realism of the sounds on the 1 – 5 scale. Particular emphasis was put on instructing the test subject to focus on how realistically the sound levels change when they move their head. The subject was encouraged to turn around and listen to the sound from different directions until they are ready to give their answer. They were also told not to let the visual environment affect their estimations of the directions of the sounds or the realism of the listening experience but instead to imagine hearing each sound in any quiet environment. To avoid any confusion as to whether the test subject should attempt to determine the direction of the sound source or simply indicate the direction of the sound as they perceive it, it was made clear in the instructions that the sound source is always located in the direction the sound appears to be coming from and that the position of the sound source won't change until they have given both the directional estimate and the rating for the realism of the sound.

4.3.3 Test structure

The test started with a short practice phase where each of the audio clips (the cat, the typewriter and the water) appeared once in a predetermined order and with predetermined parameters, and the test subject would evaluate these sounds in the same way as in the actual test. The parameter combinations used for the three dummy sounds were close to but not equal to the ones used in the actual test, which allowed the subject not only to get used to the interface and structure of the test but also to get an idea of how and how much the range of sounds presented in the test might differ from one another in terms of realism of the spatial impression. The dummy sounds also provided the test subject with an opportunity to adjust the headphones so that the sound in both ears would be as clear and loud as possible.

At the end of the practice phase the subject was asked for the final time to confirm that they have understood what to do and are ready to begin the test. In the test the 18 permutations generated from the 3 audio clips and the 3 alpha and 2 sharpness values were played back in a random order and from random directions, and the test subject would assess each of them in three steps: At first, they were asked to face the direction where they heard the sound directly in front of them and press the trigger on the controller (see Figure 3). Next, they were asked to face the direction where they heard the sound

directly behind them and, again, press the trigger. Finally, a 1 – 5 scale would appear on the screen (see Figure 9) and the subject would rate the realism of the listening experience by pointing at any value between 1 and 5 with the laser pointer on the controller and pressing the trigger. These three steps would repeat in the same order for all 18 sound instances.

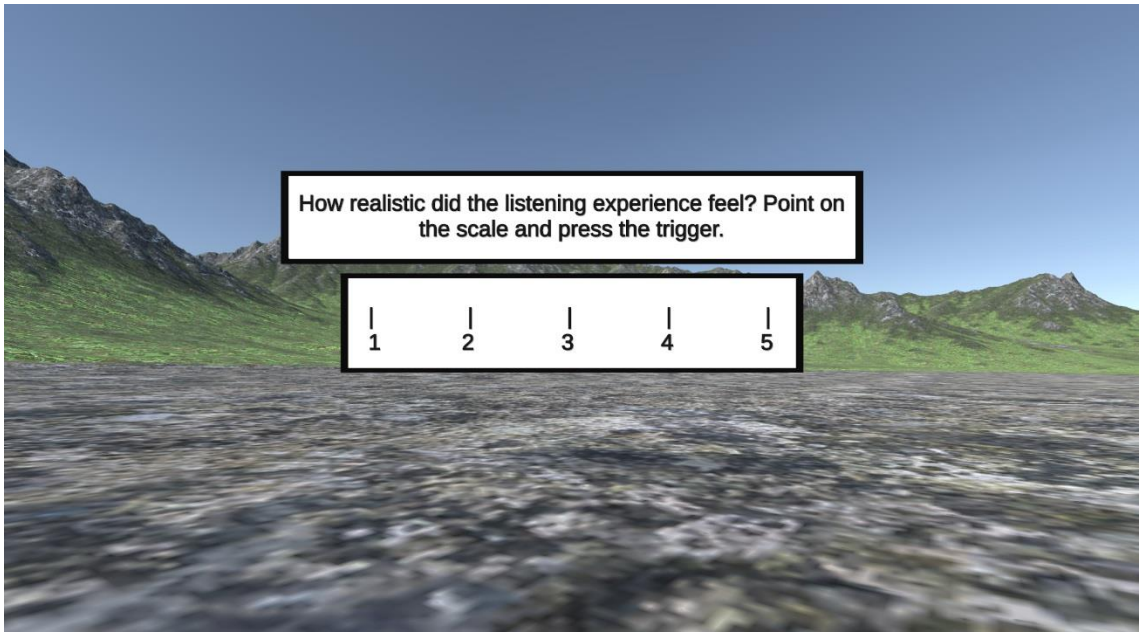

Figure 9. Screenshot from the realism evaluation phase of the test environment. The test subject would use the laser pointer of the controller to select a rating from the scale on the screen.

The main reason for having the subject estimate the direction of the sound from both the front and the back was to give them a reason to turn around at least once during each sound and hear the transition between the front and back directions. It also made it possible to explore whether there would be differences in the accuracy of directional hearing for one or both of the directions with different parameter combinations and whether there would be a correlation between the accuracy of the directional estimations and the perceived realism of the listening experience. In particular it was predicted that the directivity alpha value of 0.5, the highest used in the test, might make it easier to judge the direction especially from the back while also negatively affecting the realism ratings.

5 RESULTS AND DISCUSSION

The realism ratings were significantly worse when the alpha parameter was set to 0.5 than when it was set to 0.3 or 0.4, regardless of the value of the sharpness parameter. The statistical significance of the results was confirmed by sorting the average ratings from high to low, first for each of the three sound types separately and then for all three together, and calculating the P value for each consecutive pair for the significance level of 5%. Box plots of the realism ratings can be seen in Figure 10.

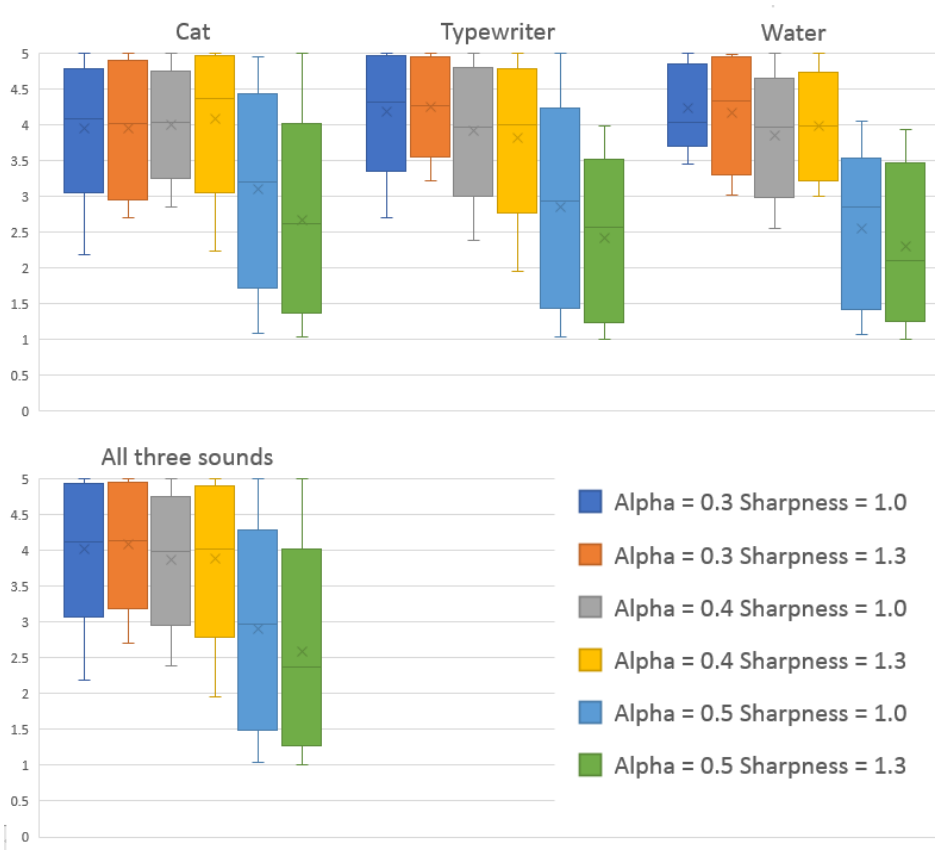

Figure 10. Box plots of the realism ratings. The bottom and top ends of the box represent the lower and upper quartiles, while the horizontal band and the X within the box represent the median and the mean, respectively. The ends of the whiskers outside the box show the minimum and the maximum values.

For the directional estimations there was more variation between the three sound types than with the realism ratings. Overall the top three combinations with the most accurate directional estimates for both front and back directions were (alpha = 0.5, sharpness = 1.0), (alpha = 0.5, sharpness = 1.3) and (alpha = 0.4, sharpness = 1.0). For the rest of the combinations, (alpha = 0.3, sharpness = 1.0), (alpha = 0.3, sharpness = 1.3) and

(alpha = 0.4, sharpness = 1.3), the estimates were less accurate, but perhaps not significantly so. The directional estimates and the corresponding realism ratings are depicted in Figure 11.

It is not surprising that with the alpha value of 0.5 the direction of the sound was the easiest for the test subjects to judge, as with this value the sound coming from the back is attenuated to the point of being completely muted. The subject could then simply reason that when they could not hear the sound at all, it must have been directly behind them. This does not, however, explain why the directional estimates with this alpha value were more accurate than with the other two not only from the back but also from the front. The most likely explanation is that since the test subjects were free to move and turn around during all evaluation phases, they had the option of first finding the deaf spot and then making a 180 degree turn to face towards the sound source. It is also possible that the effects of the alpha parameter to the directivity pattern from the front, although less pronounced than from the back, were noticeable enough to make a difference. Either way, it may be of some interest that the two parameter combinations with the worst realism ratings were also among the ones with the most accurate directional estimates for both front and back directions.

The statistical significance of the differences between the directional estimates for each parameter combination, and consequently the possible correlation between the accuracy of the directional estimations and the realism ratings, is not simple to determine. The P value test on its own is not reliable in this case, as the average and median values are skewed by a few extreme errors such as mistaking a sound coming from the back as coming from the front, or the other way around. Such errors occurred especially with the combination of alpha value of 0.3 and sharpness value of 1.3 (five times in total with this combination, and no more than once with any other combination), but the sample size was not large enough to decide whether this had something to do with some characteristic of that particular directivity pattern or whether it was just a coincidence.

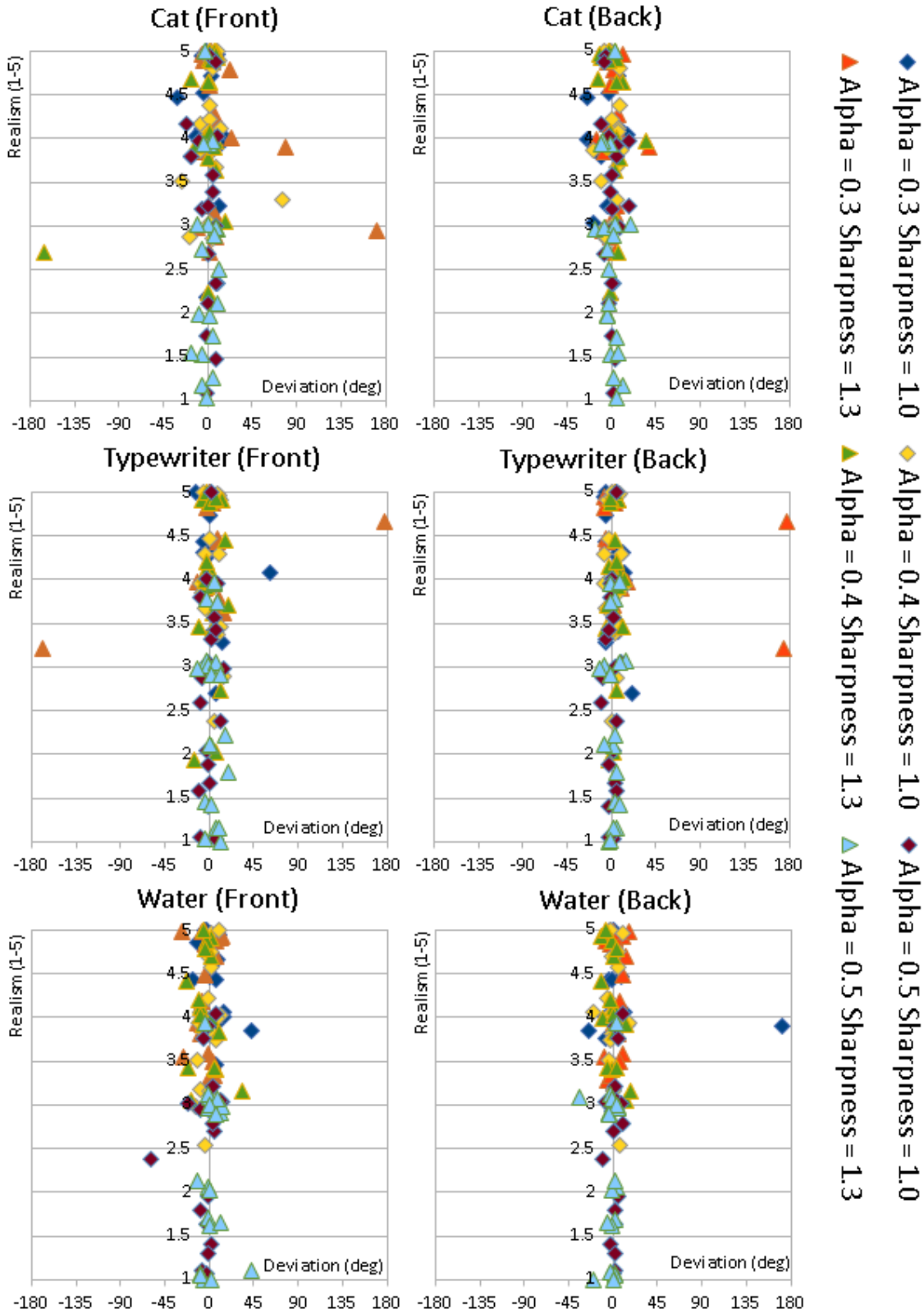

Figure 11. Directional estimations and the corresponding realism ratings. The charts show all of the recorded data, assessments by every test subject for all six parameter combinations.

The directional estimates for all three sound types are fairly evenly spread out around zero, which would support the conclusion made earlier from the hearing test results that the differences in hearing thresholds between the ears were minor and did not significantly affect directional hearing for any of the test subjects. A more detailed analysis reveals that only two or three of the twenty test subjects had a noticeable bias towards either left or right, and among the twenty there were none that consistently misjudged the direction of every sound in the same direction.

The average error for most sounds and parameter combinations was under 10 degrees in either direction, and many of the estimates were just a few degrees off from the correct answer. This is fairly well in line with although not quite as good as what was expected based on how accurate human directional hearing is in the real world. The larger inaccuracies in the virtual environment are at least partly explained by the fact that although HRTFs vary slightly from person to person in the real world due to differences in factors such as the shape of the head and the ears, Resonance Audio as well as most other spatial audio plugins uses the same generic HRTFs for everyone. As a result the virtual spatial hearing experience at its best is close enough to real for most people, but not exactly ideal for anyone. Another factor that may have affected the results was the supra-aural headphones. Although the test subjects were instructed to adjust the headphones so that they hear the sound in both ears as clearly as possible, there was no way to confirm that both sides were positioned exactly the same way on top of the ears.

6 CONCLUSION

Although none of the parameter combinations clearly stood out as the best one, one of the values used in the test yielded significantly lower realism ratings than others. This difference was expected, as the attenuation from behind with the value in question was distinctly more intense than in real life. No predictions had been made prior to the experiment about which of the remaining four combinations might get the best ratings, and in the end the differences between them turned out to be minor. This, again, was not unexpected. While the problem of finding the one optimal directivity pattern for an immersive spatial audio experience remains unsolved, at the very least the experiment confirmed that the directivity pattern does have a major effect on the perceived realism and the immersivity of the listening experience, and that a virtual environment such as the one used in the experiment is suitable for bringing out and examining these effects.

The directional estimates were reasonably well in line with how accurate human directional hearing on the horizontal plane generally is. The majority of the larger errors were cases of confusions that also occur fairly often in real life, such as mistaking a sound coming from the back as coming from the front. Perhaps the most interesting observation was the possible negative correlation between the accuracy of directional estimates and the perceived realism of the listening experience. Confirming this premise would, however, require a larger group of test subjects and quite likely some adjustments to the test setup and equipment to diminish the chance of observational error.

One purpose of the experiment was to find out whether the virtual test environment would be suitable for measuring the realism of spatial audio and the subjective experience of immersion in the virtual auditory space. Based on the apparent validity of the results and the fluency of the test process itself it would seem that the test environment could be used with success for running similar experiments in the future. In particular the user interface was intuitive and easy to use even for test subjects who had not used VR devices before, and from a developer standpoint the built-in motion tracking of the VR headset provided a simple way to accurately capture positional and rotational data. With minor changes the same test environment could be used for performing tests with different variables, plugins or setups. For example, the effect of reflections or multiple simultaneous sound sources on the level of immersion on one hand and the accuracy of spatial hearing on the other would make an interesting topic for future research.

REFERENCES

[1] Christopher Plack. *The Sense of Hearing*. Psychology Press, Taylor & Francis Group, New York, 2014.

[2] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1983.

[3] Peter Damaske & Volker Meller. Ein verfahren zur richtungstreuen schallabbillung des oberen halbraumes über zwei lautsprecher. *Acustica*, 22: 154-162, 1969.

[4] Peter Damaske. Head-related two-channel stereophony with loudspeaker reproduction. *The Journal of the Acoustical Society of America* 50 (4): 1109-1115, 1971.

[5] Stephen Schütze & Anna Irwin-Schütze. *New Realities in Audio: A Practical Guide for VR, AR, MR and 360 Video*. CRC Press, 2018.

[6] Frederic L. Wightman & Doris J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91: 1648-61, 1992.

[7] A. Politzer. Studien über die paracusis loci. *Arch. Ohren-Nasen-Kehlkopfheilkunde*, (11): 231-236, 1876.

[8] Emil Bloch. Das binaurale hören. *Z. Ohren-Nasen-Kehlkopfheilkunde*, (24): 25-83, 1893.

[9] W. E. Perekalin. Über akustische orientierung. *Hals-Nasen-Ohrenheilkunde*, (25): 443-461, 1930.

[10] P. H. G. van Gilse. Untersuchungen über die schallokalisation. *Acta Oto-Laryngologica*, 14(1):1-20, 1930.

[11] S. S. Stevens & E. B. Newman. The localization of actual sources of sound. *The American Journal of Psychology*, 48(2):297-306, 1936.

[12] F. M. Tonning. Directional audiometry. I. Directional white-noise audiometry. *Acta oto-laryngologica*, 69:388-94, 1970.

[13] Lord Rayleigh, M. A. & F. R. S. Acoustical observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 3(20):456-464, 1877.

[14] H. Geoffrey Fisher & Sandord J. Freedman. The role of the pinna in auditory localization. *Journal of Auditory Research*, 8:15-26, 1968.

[15] Jens Blauert. Ein versuch zum richtungshören bei gleichzeitiger optischer stimulation. *Acustica*, (23):118-119, 1970.

[16] Jens Blauert. Ein beitrag zur theorie des vorwärts-rückwärts-eindruckes beim hören. *6th International Congress on Acoustics, Tokyo*, A-3-10, 1968.

[17] C. C. Pratt. The spatial character of high and low tones. *Journal of Experimental Psychology*, 13(3):278, 1930.

[18] Otis C. Trimble. Localization of sound in the anterior-posterior and vertical dimensions of “auditory” space. *British Journal of Psychology. General Section*, 24(3):320-334, 1934.

[19] Suzanne K. Roffler & Robert A. Butler. Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1260-1266, 1968.

[20] Suzanne K. Roffler & Robert A. Butler. Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1255-1259, 1968.

[21] Mark B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1):47-53, 1969.

[22] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost & Sandra J. Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633-1654, 1999.

[23] Sofa (spatially oriented format for acoustics) files. https://www.sofaconventions.org/mediawiki/index.php/Files. Accessed: 2019-05-20.

[24] Sadie (spatial audio for domestic interactive entertainment). https://www.york.ac.uk/sadie-project/index.html. Accessed: 2019-05-20.

[25] Veli Laamanen. Audio design for immersive virtual environments using researched spatializers. Master's thesis, Aalto University, 2018.

[26] Richard Gould & Chris Lane. Let's test: 3D audio spatialization plugins. *Designing Sound. The Art & Technique of Sound Design*, 2018. http://designingsound.org/2018/03/29/lets-test-3d-audio-spatialization-plugins. Accessed: 2019-05-20.

[27] Developer guide for Resonance Audio for Unity. https://resonance-audio.github.io/resonance-audio/develop/unity/developer-guide. Accessed: 2019-05-20.

[28] Developer guide for Resonance Audio for Unreal. https://resonance-audio.github.io/resonance-audio/develop/unreal/developer-guide. Accessed: 2019-05-20.

[29] Steam Audio Unity plugin. https://valvesoftware.github.io/steam-audio/doc/phonon_unity.html. Accessed: 2019-05-20.

[30] Steam Audio Unreal Engine 4 plugin. https://valvesoftware.github.io/steam-audio/doc/phonon_unreal.html. Accessed: 2019-05-20.

[31] Oculus documentation. https://developer.oculus.com/documentation. Accessed: 2019-05-20.