TH Köln

PROJECT REPORT

Realizing spatial listening tests in Virtual Reality using Unity

Alexander Müller

Supervised by
M.Sc. Melissa Andrea Ramírez Caro
&
Prof. Dr. Christoph Pörschmann

November 14, 2021

Contents

1 Abstract

Central Auditory Processing Disorder (CAPD) is described by the American Speech-Language-Hearing Association (ASHA) as a condition, which "may lead to or be associated with difficulties in higher order language, learning, and communication functions" without being caused by actual hearing loss or inabilities [5].

Based on the research done by Cameron and Dillon about Spatial Processing Disorder (SPD) and Listening in Spatialized Noise (LiSN) a similar test shall be developed based on the Unity development engine with support for Virtual Reality (VR) peripherals.

2 Introduction

Spatial hearing describes the ability, to localize the origin of a given sound event, by using binaural clues of the respective signal. The more obvious advantages of this capability include an aid in orientation and improved possibilities to react to events (like an approaching car), which aren't in the filed of view.

But apart from this, spatial hearing also allows to separate different sound sources and helps managing noisy environments. This is especially useful if the incoming audio signals are similar to each other. A popular example for this is the so called "cocktail-party-effect", which basically described a situation where a listener has to distinguish between a lot of speech signals and focus on a single one in order to be able to maintain a conversation.

Even though this scenario does not cause too much problems for the average person, it contains a lot of problems for everyone with hearing inabilities. Unfortunately, in many cases these impairments affect the ability of spatial hearing and therefore reducing or eliminating information that could otherwise be used to separate the different audio sources from each other. Even modern day hearing aids still can't offer the nuances required to precisely determine the location of a given sound source [4].

But, apart from physiological issues causing problems with spatial hearing, it has been found that children with normal hearing thresholds still might not be able to correctly interpret binaural clues. This effect is described as SPD. Since the actual hearing abilities of the affected children is not impaired, they are - in contrast to persons with typical hearing inabilities - provided with all the information required to localize a given sound source. Based on this knowledge the idea was formed, to treat SPD by *training* affected children and therefore teaching them, how to correctly interpret binaural clues.

Within the *Development and Evaluation of the LiSN & Learn Auditory Training Software for Deficit-Specific Remediation of Binaural Processing Deficits in Children: Preliminary Findings* [1] paper by Sharon Cameron and Harvey Dillon it is described how such a training process could look like and already gave some hints [1].

[1] The original experiments have only been done with a small sample size. Within the conclusion of the paper the requirement of a clinical trial is mentioned, to validate the efficacy of the training process.

Building upon this approach the training concept described by Cameron and Dillon shall be transferred into a virtual reality application. This shall bring the additional advantage of combining auditory and visual clues and also allows to further extend the process.

3 Introduction B

Spatial hearing describes the ability to use clues in a given sound event to localize its source. This can be achieved by taking advantage of three main features: the difference in level (1) and phase (2) of a signal between the left and the right ear and the spectral colouring (3) of a known noise, which is dependant on the path of the sound waves around the listeners head and upper body.

Apart from obvious advantages of spatial hearing, like localizing the origin of a given sound (e.g. a nearby car), it also provides the listener an option the filter sounds based on their direction. This is particularly important in noisy environments. This situation is most commonly described in the "cocktail-party-effect" (see [3] for further details). A listener can use the information of position of a given audio source to reduce noises from other directions within their perception.

Since noisy environments with a lot of similar audio signals (in particular speech) are very common within a lot of every day occasions, the advantage of using binaural clues has a great importance. In most cases the lack of this ability is found on people with hearing disabilities. Even if hearing aids have advanced throughout the last decades, they still can't offer their users the same subtle binaural clues as a person with intact hearing abilities.

But apart from physiological issues causing problems with spatial hearing it has been found, that children with perfect hearing still might not be able to correctly interpret binaural clues. This effect is described as SPD. Since the actual hearing abilities of the affected children is not impaired, they are - in contrast to persons with typical hearing inabilities - provided with all the information required to localize a given sound source. Based on this knowledge the idea was formed, to treat SPD by *training* affected children and therefore teaching them, how to correctly interpret binaural clues.

Within the *Development and Evaluation of the LiSN & Learn Auditory Training Software for Deficit-Specific Remediation of Binaural Processing Deficits in Children: Preliminary Findings* [1] paper by Sharon Cameron and Harvey Dillon it is described how such a training process could look like and already gave some hints [2].

Building upon this approach the training concept described by Cameron and Dillon shall be transferred into a virtual reality application. This shall bring the additional advantage of combining auditory and visual clues and also allows to further extend the process.

[2] The original experiments have only been done with a small sample size. Within the conclusion of the paper the requirement of a clinical trial is mentioned, to validate the efficacy of the training process.

4 Motivation

Realizing this listening test as a Unity-based VR application offers several advantages. The first one being the open nature of this project. Since only free-to-use assets have been included, both the compiled application and the source files can be made publicly available. Combined with simple setup required to perform the listening experiments, this will hopefully allow a variety of interested groups to perform these tests to collect further data and offer affected children access to work on the condition.

Furthermore the Unity framework in conjunction with the prefabs and scripts, that have been created for this project, it would be relatively easy to extend upon the original training concept. For example the addition of multiple distracters within a given scene could very easily be realized.

Apart from the previously mentioned aspects the usage VR peripherals for this project also offers a lot more possiblilites. Some of them will be discussed in **??**. But especially adaptive 3D audio in combination with head tracking would allow for a much more realistic scenario. Also the novelty of the VR headset itself will most likely already spark a lot of interest in the participants in comparison to a regular learning/training game.

5 Fundamentals

Before discussing the approach towards this project, some basic principles of modern day audio processing shall be explained.

5.1 Auditive localization

The human ability of locating audio sources is based upon two main principles. The first being the perception of the time and level differences between an acoustic event on the left and right ear. The different time points at which the signal is perceived at the two ears, translates to a different phase at which the audio wave is registered. On the other side the level differences mainly derive from the shadowing caused by the head. Both of these effects are not frequency independent, for lower tones the localization via phase differences works better, while at higher frequencies the level variations offer a better indication.

The second way of locating a sound source is based upon the spectral differences that occur through the different paths a sound wave can take around the listeners head and upper body. This effect is strongly influenced by the shape of the outer ear, but also the general geometry of the head makes a difference. Basically the head and the ear can be described as directional filters. This kind of localization is mostly based on experiences. The listener can learn to associate the spectral differences in a known noise to the location from which the noise is originated (e.g. by looking for the sound source).

5.2 Auralization

Auralization describes the process of recreating not only a given audio signal but also the spatial information. With correct auralization a recorded scene (e.g. a classical concert in a concert hall) can be rec
This requires advanced recording processes to capture the spatial information of a scene (e.g. a concert hall)
Auralization describes the process of adding spatial features to audio signals. The simplest example for this would be the move from mono to stereo audio. However, since these initial steps the options of simulating and re-creating spatial features in digital audio processing have advanced quite a lot. This can be easily seen in many consumer electronic products such as AV-Receivers, Soundbars or Headphones supporting 5.1, 7.1 etc. or in particular VR peripherals coming with their own 3D audio frameworks for developers.

The basic principle remains more or less the same for all these products. Within a given *listening space* the original auditory environment shall be recreated.

5.3 Hearing perception (energy and similarity)

5.4 Training abilities

6 Approach

Before evaluating possible improvements of virtual reality support in listening tests, it has to be made sure that there are no major other influences alternating the results of this test. Therefore the first iteration of the new test environment will be developed as close as possible to the original LiSN software. This allows to compare test and training results from this test environment with the data already available from previous experiments. If the new software allows to recreate similar test results, it can be assumed that the unenviable difference don't affect the results in a too drastic manner. Afterwards the data from these initial tests can also be used to validate whether VR exclusive features like head tracking add further benefits to the training process.

Development tools As already mentioned, the major part of the development will be done in Unity with the addition of Visual Studio as IDE. The other major tool is the Oculus VR SDK, which will be used for audio spatialization and of course rendering the visuals to be properly displayed on the given VR peripherals.

6.1 Unity overview

Unity is a development framework, which is mainly used as a engine for video games. However due to the requirements to realism in modern video games, the amount of audio processing included, is already pretty elaborate. In conjunction with the large user base and support by hardware manufacturer (e.g. Oculus), Unity offers a great starting point for this project.

6.2 Orientation on LiSN project

6.3 Audio assets

As basis for this project the word lists given in the appendix of the LiSN paper [?] are used. Since due to the huge variety of options, the only feasible way to construct the sentences is by assembling a sentence through seperate audio files for each word. This however comes with a problem. Simply recording single words and then playing the files one after another strongly alter the speech flow and accentuation, which would normally be present when the sentence is spoken.

To handle this problem, the words won't be recorded separately but instead a subset of the possible sentences from each list will be recorded. Within this subset all words from the list are included within the recordings. Afterwards the recorded sentences will be cut into the individual words and then will be used the assemble randomly generated sequences based on these assets. This should work fine, since the structure of all sentences within any list remains the same. So the speech flow and pronunciation between the sentences should not be too larger. GET SOME REFERENCES FOR THIS ASSUMPTION!

6.4 Visuals

Since this project shall be used to perform research with children as participants the graphical presentation is of great importance. Especially since Virtual Reality (as of today) is in general a quite unfamiliar setting, it is very important to give the user a sense of space within the setting.

Matching audio parameters to visuals The audio scenario shall be based upon free-field sound propagation. To avoid confusion of the listener the environment presented within the VR application should represent this setting. So a closed room would not be a good option, since we inherently expect reflections and reverberation within such a setting. A better approach would be an open field, where the lack of reflexions feels more natural.

Avatars Another important part is the graphical representation of the audio sources. Since we have the visual component given through the VR headset, the sound should not simply come from an invisible sources, but have a origin which the user can identify through the graphics. This however includes another challenge. Of course it would be possible to create humanoid avatars with complex animations - including lip syncing - to convey that the object is indeed active and not just a passive talking rock. However this would take a huge amount of effort to pull of. Instead a different path will be used here, which is also pretty common in budget oriented game design (e.g. indie games). Through abstract avatars and simple animations/movements it is possible to convince the user, that the object is alive and active, while at the same time requiring a lot less effort to pull of. SOURCE...REFERENCES

6.5 User interface

6.5.1 Audio Sources

Within the level both interactive and regular audio sources are implemented. All of them use 3D rendering, so that the preceived sound changes based on the position of the player to the object, emitting the sound. The advantage of adding interactive sound sources is that,...

6.6 Training-Game

6.6.1 Audio files

Target sentences are constructed in a way that the co-articulation in between the sentences of a list is similar enough the allow randomly scrambling up sentences will maintaining a natural sounding result.

6.6.2 Game loop

Continuous distracter stories creating a noise with similar characteristics as the target sentences.

Both the target sentences and the distracter stories shall be omitted by actual objects within the game world. This association between audio and visual representation shall create further immersion (FIND REFERENCE!).

Whenever a target sentence was played, four word options will be displayed. (IMAGES OR WORDS?). When the correct result is picked, the selected option turns green and a celebratory sound is played. Otherwise if an incorrect option is selected, it is marked red and a failure sound is played. Alternatively an "uncertain" option is presented.

SNR: For every correct guess, the SNR is decreased by 2.5 dB. For every incorrect guess, SNR is increased by 1.5 dB. At uncertain at first the same sentence is repeated with 1.5 dB higher SNR. At a second consecutive *uncertain* a new sentence will be played, with again increased SNR.

7 Concept

While section 6 only described the general design principles and foundations of the application, this section shall be used to discuss how these requirements actually be fulfilled and which parts of the project might cause issues.

The biggest part will be achieving flexibility in the projects structure, to allow other users to adapt the application to their requirements.

7.1 Asset import

Icons and audio clips, namespacing!
How does an "outsider" determine whether a word shall be selectable?
Option A: Only offer icons for selectables and compares names of clips and icons.
Option B: Add a prefix or suffix to selectable words, like: "sel-apples". This might work better within a folder?

Refacto the way files are imported! Folder names are not feasible for flexibility! Create a list for each available folder and separate groups via fileNames... Or for now just support one group at a time, which can be replaced...

7.2 Word selection

Even though the initial scope of the application shall only feature a single word list from the original paper (see Appendix A of [**?**]), the implementation should be flexible enough to support any desired word database. This is particularly important when considering the desire to offer the application in different languages. To further elaborate on how this should be done, List 1 from the original paper will be used as an example. Within table 1 all highlighted words are can be used as an *option* to question the user. When observing this table three parameters can be determined:

	Subject	Verb	Count	Adjective	Objects
The	**baby**	bought	**two**	big	**apples**
	boy	carried	**three**	blue	**bottles**
	clown	cleaned	**four**	borken	**cars**
	doctor	drew	**five**	green	**chairs**
	girl	dropped	**six**	little	**crayons**
	lady	had	**seven**	old	**cups**
	man	liked	**eight**	orange	**shoes**
	nurse	saw	**nine**	red	**spoons**
	teacher	watched	**ten**	yellow	**trucks**

Table 1 LiSN - List 1

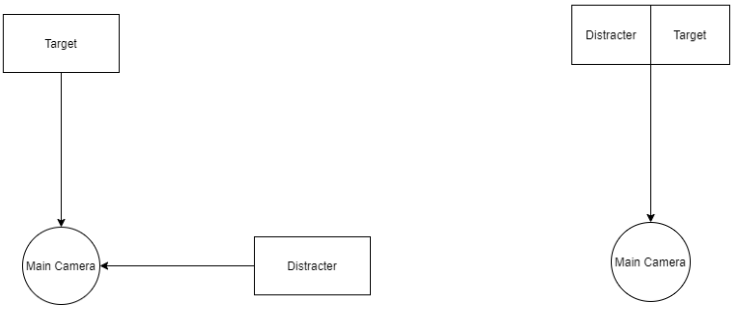
- The sentences length: 6 words
- The possibilities for each group: 9 options
- The number of 'selectable' groups: 3 groups

All three of these parameters have to be considered when creating a dynamic framework, which shall be able to support different word lists. Additionally this example also provides and interesting anomaly. All *options* within a given *group* are sorted alphabetically, except the **Count** - words. It is important to not rely on any order in which the list might be created, or any other parameter that depends on the actual content of a given word list.

However, issues like co-articulation effects have to be considered by the creator of a given group.

7.3 Training Game

Here are the training game audio setups (0 and 90 degrees:).



(a) Setup A

(b) Setup B

Figure 1 Training Game Setup

8 Conclusion

Within the scope of this project a simple binaural recreation plug-in has been implemented using the Juce framework and the FFTW library. Unfortunately the focus of this project has been moved heavily onto software development and typical coding issues, instead of digging deeper into the actual subject of audio auralization. But nonetheless the current state of the plug-in might be a good starting point for further projects, at cutting some of the very time consuming setup procedures. Of course this would require to encapsulate the currently implemented processing further into individual functions, or possibly even better an additional class. With a bit added abstraction, a generic processing chain could be implemented (similar to the one already implemented by Juce), which would allow to add and move individual auralization processes much more easily. Another working point would be to amp up the current convolution, to also feature e.g. elevation. This not too much would have to be changed in the elemental parts of the source code, this should be a rather simple addition.

Abbreviations

API	Application Programming Interface
ASHA	American Speech-Language-Hearing Association
CAPD	Central Auditory Processing Disorder
DAW	Digital Audio Workstation
DSP	Digital Signal Processing
GUI	Graphical User Interface
HRTF	Head Related Transfer Function
IDE	Integrated Development Environment
LiSN	Listening in Spatialized Noise
SDK	Software Development Kit
SNR	Signal to Noise Ratio
SPD	Spatial Processing Disorder
SRT	Speech Reception Threshold
UI	User Interface
VR	Virtual Reality

List of Figures

Sources

[1] *Development and Evaluation of the LiSN & Learn Auditory Training Software for Deficit-Specific Remediation of Binaural Processing Deficits in Children: Preliminary Findings*, Sharon Camerion, Harvey Dillon, Date: 2011

[2] *Correlating performance on the Listening in Spatialized Noise – Sentences Test (LiSN-S) with the Listening in Spatialized Noise – Universal Test (LiSN-U)*, Kiri Mealingsa, Sharon Camerona and Harvey Dillon, Published: April 2020, Date: 2019

[3] The cocktail-party problem revisited: early processing and selection of multi-talker speech, Adelbert W. Bronkhosrt, Date: April 2015

[4] *Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners*, Jens Cubick, Jörg M Buchholz, Virginia Best, Mathieu Lavandier, Torsten Dau, Date: 2016, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6246072/`

[5] *Central Auditory Processing Disorder*, American Speech-Language-Hearing Association (ASHA), `https://www.asha.org/Practice-Portal/Clinical-Topics/Central-Auditory-Processing-Disorder/`