Sound localization

Sound localization in audio-based games for visually impaired children

R. Duba
B.W. Kootte

Delft University of Technology

TU Delft Delft University of Technology

Challenge the future

SOUND LOCALIZATION

SOUND LOCALIZATION IN AUDIO-BASED GAMES FOR VISUALLY IMPAIRED CHILDREN

by

R. Duba
B.W. Kootte

in partial fulfillment of the requirements for the degree of

Bachelor of Science
in Electrical Engineering

at Delft University of Technology,
to be defended on Monday June 29, 2015

Student numbers:	4147472 (R. Duba)
	4174283 (B.W. Kootte)
Supervisors:	dr. ir. R.C. Hendriks
	S. Khademi, M.Sc.
Thesis committee:	dr. ir. M.A.P. Pertijs
	dr. ir. R.C. Hendriks
	dr. J. Hoekstra

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

ABSTRACT

This thesis describes the design of a sound localization algorithm in audio-based games for visually impaired children.

An algorithm was developed to allow real-time audio playback and manipulation, using overlapping audio blocks multiplied by a window function. The audio signal was played through headphones.

Multiple sound localization cues are evaluated. Based on these cues, two basic sound localization algorithms are implemented. The first uses only binaural cues. The second expands this with spectral cues by using the head-related impulse response (HRIR). The HRIR was taken from a database, interpolated to obtain optimal resolution and truncated to minimize memory usage. Both localization algorithms work well for lateral localization, but front-back confusions using the HRIR are still common.

The signal received from a sound source changes with the distance to the sound source. Both the distance attenuation and propagation delay are implemented.

As an alternative means of resolving front-back ambiguities, the use of a head tracker was investigated. Experiments with a webcam based head tracker show that a head tracker is a powerful tool to resolve front-back confusions. However, the latency of webcam based head trackers is too high to be of practical use.

PREFACE

This thesis describes the design of a sound localization algorithm for use in an audio-based game for visually impaired children. It is written as part of the 'Bachelor Graduation Project' of Electrical Engineering at the Delft University of Technology.

The goal of the project was to design an audio-based game for visually impaired children. The motivation for this project is that most games on the market are based on visual feedback, inhibiting visually impaired children to enjoy these games. In this project we developed a game that can be enjoyed by both sighted and visually impaired children.

The project was carried out by six people divided into three subgroups. Each subgroup developed one subsystem of the final game and wrote their own thesis. The first group, consisting of Marc Bisschop and Judith Noortman, developed a method for embedding accurate sound reflections into the game. The second group, the authors of this thesis, developed the sound localization and audio playback algorithms. The last group, consisting of Taco Coppoolse and Wirdmer van Dam, combined the work from the first two groups and expanded it into an enjoyable game.

We would like to thank Marc, Judith, Taco and Wirdmer for their contribution to this project. It was a pleasure to work with them. We also want to thank our supervisors, Richard Hendriks and Seyran Khademi, for their support and guidance throughout the project.

In the span of ten weeks, we implemented different localization cues, as well as different physical effects important for modelling sound propagation. It has been an exciting journey that has taken us on some detours outside our Electrical Engineering field, into the fields of Computer Science, Physics, Acoustics and Psychology. We enjoyed working on this project and hope that you will enjoy reading this thesis as well.

Remco Duba
Bart Kootte
Delft, June 2015

Contents

1 INTRODUCTION

1.1. BACKGROUND

Children all around the world enjoy playing computer games. It provides them with an adventurous, exciting and possibly instructive way to pass time. Yet, visually impaired children do not have this privilege. Almost all video games rely on visual feedback to give the user information about the location of various objects in the game. Although almost all games provide auditory feedback, it is rarely designed such that all necessary localization cues are available in the audio signal. Hence, visual feedback is required to allow users to play the game.

To allow visually impaired children to enjoy computer games, the game needs to be designed such that all necessary localization cues are present in the audio signal. Another option is to provide localization cues with haptic feedback, but this is outside the scope of this project. The most obvious way to encode localization cues in the audio signal is to use the same methods that the human auditory system uses in real life to localize sound sources. This approach requires the smallest learning curve from the user.

1.2. PROJECT

In this project, a prototype game is developed in which all localization cues are present in both visual and auditory feedback. This allows the game to be enjoyed by both sighted and visually impaired children. The game was developed by six people, divided into three subgroups. The first subgroup developed an algorithm to simulate sound reflections in the game. The second subgroup was responsible for the gameplay and user interface. The third subgroup developed the sound localization algorithm. The work of the third subgroup is discussed in this thesis.

1.3. THESIS STRUCTURE

Chapter 2 describes the problem analysis. First, all relevant boundary systems are described. Next, the requirements for the sound localization algorithm are listed and motivated.

Chapter 3 contains research related to sound localization. It covers the physics of sound propagation, auditory perception and the wide range of sound localization cues utilized by the human auditory system.

Chapter 4 describes the design process. It motivates the design choices and briefly discusses the implementation.

Chapter 5 describes several experiments designed to test whether the design satisfies the requirements. This chapter presents the results of these experiments to the reader.

Chapter 6 discusses the results presented in Chapter 5. It is discussed whether the requirements are satisfied and how unexpected results can be explained.

Chapter 7 summarizes the conclusions in this thesis and gives recommendations for future work.

At the end of this thesis, a glossary is included, listing all the uncommon terms and abbreviations used in this thesis.

2

PROBLEM ANALYSIS

The goal of this project is to design and write the software for an audio-based game for visually impaired children. Because visually impaired children have little or no visual feedback, the game is purely audio based. Haptic feedback is outside the scope of this project. In order to facilitate navigation for the user, the same methods of sound localization will be used as the average user uses in real life. This thesis focuses on the sound localization aspect of the game.

2.1. ANALYSIS OF BOUNDARY SYSTEMS

This section describes the boundary of the designed sound localization subsystem, as discussed in Chapter 4, with other subsystems. A system level block diagram is shown in Fig. 2.1. We will now discuss the interface between the sound localization subsystem and each of the blocks in the diagram in more detail.

Audio output device The entire interface to the audio output device is part of the sound localization subsystem. The motivation for this will become clear in Section 4.4.

Audio file input Before sound can be manipulated, it needs to be loaded from a file. This is also part of the sound localization subsystem.

Gameplay positions The gameplay subsystem facilitates important gameplay elements such as movement, collision detection, interaction with virtual entities and storyline. Its implementation is discussed in [1]. Coordinates of all entities in the game are defined and manipulated by the gameplay subsystem. Because sound localization is based on the position of all sound sources relative to the player position, these coordinates should be passed to the sound localization subsystem.

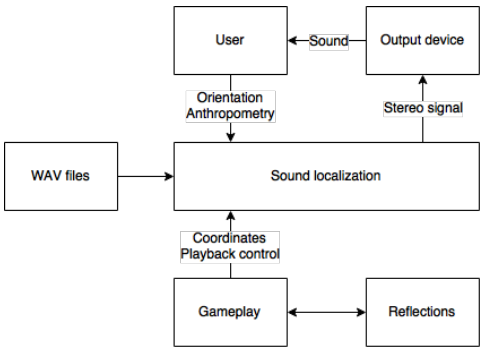

Figure 2.1: System level block diagram of the game showing the interface between sound localization and other subsystems.

Gameplay audio playback API Where and when each sound source starts and stops playing is controlled by the gameplay subsystem [1]. For example, a footstep sound is only playing when the player is moving, or a different background music is played in different rooms. Therefore, a well defined application programming interface (API) should be present in the sound localization subsystem that allows different sounds to be loaded, started and stopped by the gameplay subsystem.

Reflections Sound reflections are implemented in order to create an impression of the virtual space and aid distance perception [2]. This is done by the reflection subsystem. It translates all the reflections to attenuated virtual sources. These virtual sources can have different coordinates than the original sound source. The gameplay subsystem is responsible for creating these virtual sound sources, because it is already responsible for the creation of each of the original sound sources, as explained in the previous section. Even though the sound localization subsystem has no direct boundary with the reflection subsystem, the audio API should provide support for multiple copies of the same sound source.

The user The final boundary system is the user itself. Sound localization may depend on user orientation or anthropometric measurements. The measurement of any of these quantities, if required to satisfy the requirements in Section 2.2, is part of the sound localization subsystem.

2.2. REQUIREMENTS

In the following section we derive all the requirements for the sound localization subsystem.

2.2.1. FUNCTIONAL REQUIREMENTS

The following functional requirements are set up for the sound localization:

1. Functional requirements

- **Req. 1.1** *The program should be able to load both mono and stereo sound from WAV files.*
- **Req. 1.2** *Audio playback should be performed in real time, with a maximum delay between user input and sound playback of* 100 ms.
- **Req. 1.3** *Audio playback should be possible. The playback should be smooth, i.e. it should not contain any audible clicks or stuttering.*
- **Req. 1.4** *Sound localization should be possible in both azimuth and elevation.*
- **Req. 1.5** *The user should be able to distinguish between a sound source directly in front and directly behind the player.*
- **Req. 1.6** *The user should be able to estimate the relative change in distance of a moving, constant amplitude sound source.*
- **Req. 1.7** *Localization performance should be independent of the users position in the real room.*

Audio input and output are described by requirements 1.1 and 1.3. In order to create a good user experience, requirement 1.3 ensures smooth audio playback. Requirement 1.2 is also important for a good user experience.

To allow efficient navigation through the virtual space, auditory cues should contain the information to unambiguously map any auditory signal to a point in space relative to the player. This is captured in requirements 1.4 to 1.6. An important aspect of the game is accessibility for visually impaired people. Therefore, equipment dependent on the user's position in the real room is undesired. Requirement 1.7 ensures this is avoided.

2.2.2. DEVELOPMENT REQUIREMENTS

For developing the sound localization algorithms, the following requirements are set up:

2. Development requirements

- **Req. 2.1** *The system should be developed in a time span of 10 weeks.*
- **Req. 2.2** *Only development tools available without additional cost should be used.*
- **Req. 2.3** *A well defined API should be defined for use by the gameplay subsystem. It should be capable of loading files, starting and stopping playback, creating copies of the same sound source and mixing different sound sources.*

Requirement 2.1 is based on the time span of the Bachelor Graduation Project. To avoid the need for budget and the risk of impeding any follow up projects, requirement 2.2 was set up. The sound localization described in this thesis is only a subsystem of the entire game. Requirement 2.3 ensures easy integration with the gameplay subsystem. The copies of the same sound source are required to support reflections.

2.2.3. HARDWARE REQUIREMENTS

The following requirements are set up regarding the hardware that the user requires to be able to play the game:

3. Hardware requirements

Req. 3.1 *Dependence on available hardware should be limited to a minimum.*
Req. 3.2 *CPU usage should be limited to* 80 % *on a* 2.3 GHz *Intel Core i5.*
Req. 3.3 *Memory usage should be limited to* 500 MB.

Requirement 3.1 keeps the game accessible for a wide audience. When specialized hardware is required, this accessibility is lost. Requirements 3.2 and 3.3 make sure that the game is playable on the average PC. These requirements are based on the specifications of the slowest computer used for developing the game.

3

RELATED RESEARCH

This chapter contains all the research to lay the foundation for the system described in Chapter 2. The first section describes all the acoustic properties of sound, the next section discusses the structure of the human hearing system and the last section explains the sound localization of the human brain. The acoustic properties of sound together with the structure of the human hearing system are the most important aspects of human sound perception.

3.1. SOUND

This section describes some of the acoustic properties of sound waves. The localization of sound depends on its acoustic properties. Understanding the physics of sound waves will provide a better understanding of the concepts of sound localization. Therefore, we will start with a short introduction of sound waves and then discuss what happens when sound travels through the air or when it reaches an object, for instance the head.

3.1.1. SOUND WAVES

Sound can be described as the vibration of molecules in a medium. It is a longitudinal wave, propagating via a series of compressions in the medium. The medium used in this project is air. Compression of air increases the pressure and rarefaction decreases the pressure, see Fig. 3.1. The wavelength λ is the distance between two wave crests and it depends on the speed of sound c and the frequency f. The wavelength is given in Eq. (3.1).

$$\lambda = \frac{c}{f} \tag{3.1}$$

When a sound wave reaches an obstacle, it can be reflected from the obstacle or diffracted around the

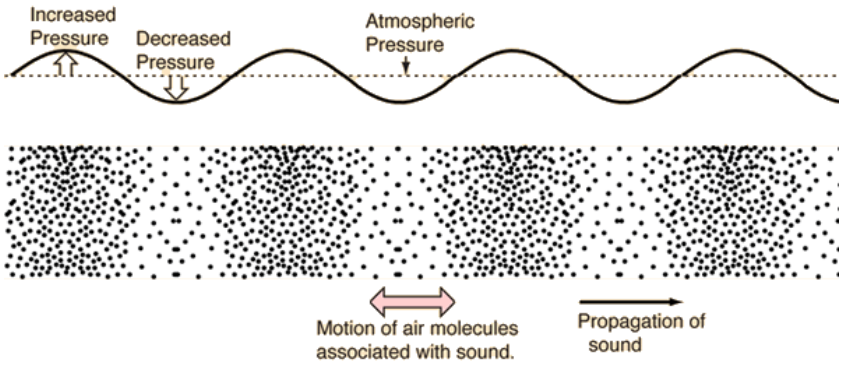

Figure 3.1: A propagating sound wave in air.
Source: http://hyperphysics.phy-astr.gsu.edu/hbase/sound/tralon.html, accessed on 19-6-2015.

obstacle. Reflections are described by the law of reflection, i.e. the angle of incidence equals the angle of reflection [3, p. 515]. Scattering occurs when an object is around two thirds of the wavelength [4, p. 102].

When two sound waves are superimposed, interference can be observed. This can be either constructive interference, when the amplitudes add, or deconstructive interference, when the amplitudes cancel.

Sound pressure is expressed in dB. A pressure of 20 µPa is the threshold for human hearing and is the baseline for auditory pressure defined as 0 dB SPL (sound pressure level) [5, p. 19].

3.1.2. DISTANCE

The effect of the distance from the source to the receiver on the sound wave is threefold:

- Propagation delay
- Sound attenuation
- Sound absorption

The speed of sound is finite, therefore, when a source emits sound, it will not instantly be heard by the receiver. The speed of sound in air is given by Eq. (3.2) [4, p. 6]. It depends on the composition and the temperature of the air.

$$v_{gas} = \sqrt{\frac{\gamma RT}{M}} \tag{3.2}$$

γ = adiabatic index (1.4)
R = gas constant (8.31 J K^{-1} mol^{-1})
T = temperature (K)
M = molecular weight of the gas (0.029 kg mol^{-1})

As the sound waves propagates through the air, the intensity is spread across the surface area of a sphere [6, p . 36]. Most natural sound sources are isotropic, i.e. the intensity is spread equally across the surface of a sphere. The resulting sound intensity at the receiver for free field given by Eq. (3.3). The drop in sound intensity equals 6 dB per doubling of the distance.

$$I = \frac{QW}{4\pi r^2} \tag{3.3}$$

Q = directivity of the source (-)
W = intensity of the source (W)
r = distance to the source (m)

The third effect is that the energy at high frequencies is absorbed by the atmosphere. This is due to heat conduction, internal friction, slowness of energy exchange between the internal degrees of freedom in a gas molecule and the relaxation of air molecules [7–9].

3.2. AUDITORY PERCEPTION

The anatomy of the human hearing system consists of three sections: the outer ear, the middle ear and the inner ear [4, p. 66]. The anatomy of the human hearing system is shown in Fig. 3.2.

The main function of the outer ear is to help sound localization with the acoustic effect of the pinna (the external flap of tissue) and the concha (the hollow at the entrance of the ear canal). They are particularly useful in determining whether a sound source is in front or behind and to lesser extent as to whether it is above or below [4, p. 66]. This is explained in more detail in Section 3.3.2.

The function of the middle ear is twofold. Firstly, the tympanic membrane converts acoustic pressure variations from the outer ear to mechanical vibrations in the middle ear [4, p. 67]. These mechanical vibrations are passed to the inner ear by three bones commonly known as the hammer, anvil and stirrup. Secondly, the middle ear provides protection from loud sounds.

The inner ear consist of structure known as the cochlea. Its function is to convert the mechanical vibrations in the middle ear into nerve firings. These nerve firings are then processed by the brain [4, p. 71].

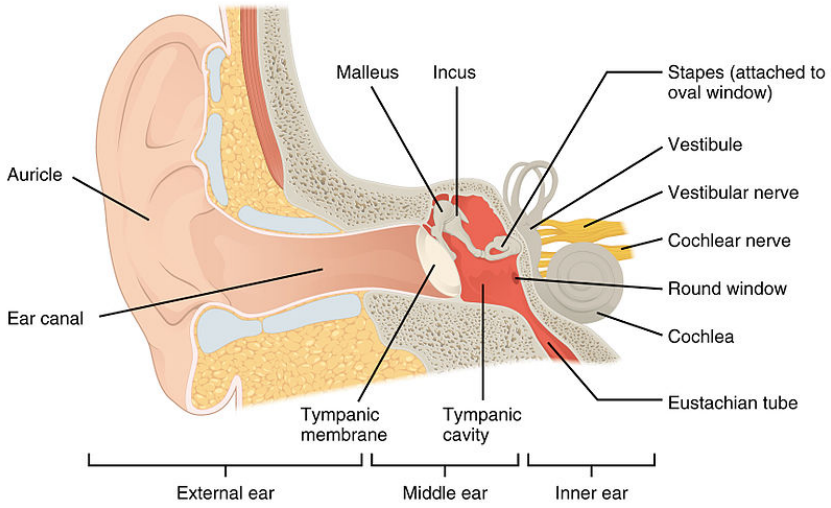

Figure 3.2: The anatomy of the human ear.
Source: Structures of the Ear, openstax cnx, accessed on 11-6-2015,
http://cnx.org/contents/14fb4ad7-39a1-4eee-ab6e-3ef2482e3e22\spacefactor\@m7.1:92/Anatomy_&_Physiology

3.3. SOUND LOCALIZATION

The brain uses a variety of cues for sound localization. Different cues cover different parts of the frequency range or different ranges of coordinates [4, p. 102] [10].

First, cues that use the difference between the sounds received by the left and right ear, known as binaural cues, are discussed in Section 3.3.1. Monaural cues based on spectral properties are discussed in Section 3.3.2. Next, the effect of head movements is discussed in Section 3.3.3. The typical spatial resolution that a human can achieve is discussed in Section 3.3.4. The influence of reflections is discussed in Section 3.3.5. Finally, the various cues used for distance perception are discussed in Section 3.3.6.

3.3.1. BINAURAL CUES

The most important cues for sound localization in the horizontal plane are the interaural intensity difference (IID) and interaural time difference (ITD). Some researchers prefer the term interaural level difference (ILD) over IID and the term interaural phase difference (IPD) over ITD. Using IPD is common when only sine functions are considered and ITD is common for broad band stimuli. The convention used in this thesis is IID and ITD, because the stimuli considered are sound sources with a wide range of frequencies.

The human ears are located at two different points in space. Therefore sound will have different arrival times for the left and right ear. The difference in arrival times is called the interaural time difference (ITD). Sound originating from the midsagittal plane (the plane perpendicular to the line connecting the ears) will result in zero ITD.

The time difference translates to a phase difference for sine waves. The phase difference will be small for frequencies below 1600 Hz, but when the frequency is increased above 1600 Hz, the phase difference can exceed one period. At 1600 Hz, the phase difference between two sine waves is exactly one period. Because there is no phase difference, there ITD will not provide any localization cue. This is why the ITD will only give accurate localization for frequencies below 1600 Hz.

The ITD is a function of the radius of the head and is given in Eq. (3.4) [4, p. 97-99]. Figure 3.3 illustrates this equation. The term $r \cdot \sin(\theta)$ in Eq. (3.4) is the distance that the sound on the left side in Fig. 3.3 has to travel before reaching the head, when the sound on the right already has reached the ear. The factor $r\theta$ is the extra path around the head. The delay is underestimated without this.

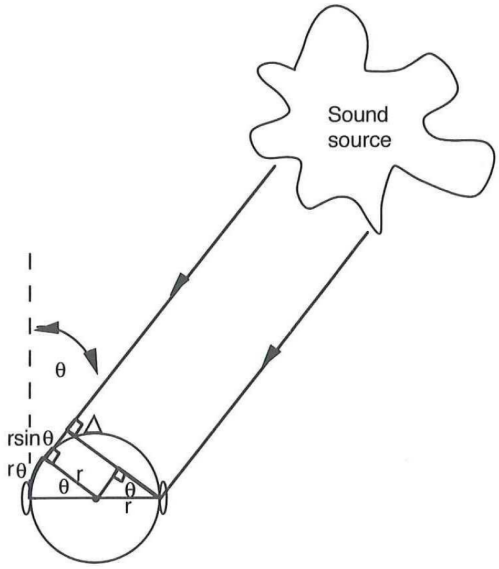

Figure 3.3: A model for the interaural time difference.
Source: [4, p. 99]

$$\text{ITD} = r\frac{\theta + \sin(\theta)}{c} \tag{3.4}$$

r = half the distance between the ears (m)
θ = angle with the midsagittal plane (rad)
c = sound velocity, (m s^{-1})

The second tool for binaural localization is the interaural intensity difference (IID). The shadowing of the head leads to a pressure difference between the two ears [4, p. 101][10].

The pressure difference depends on both the location and frequency of the sound. This is due to the shadowing of the head and the diffraction of the sound waves when the wavelength is smaller than the diameter of the head. The wavelength decreases with the frequency, according to Eq. (3.1). Therefore, low frequencies are less diffracted by the head than high frequencies. At frequencies above 637 Hz, the sound waves starts to diffract around the head [4, p. 102]. Consequently, lower frequencies have a lower IID than higher frequencies, hence the brain uses this tool for localization at higher frequencies. The level difference can be up to 35 dB for high frequencies and this cue is mainly used above 3000 Hz.

The duplex theorem states that the brain uses ITD at lower frequencies and IID at higher frequencies for sound localization [11]. Sound localization performance is worse when both IID and ITD cannot be used effectively. This results in decreased localization accuracy in the range from 1000 Hz to 3000 Hz.

For wide band stimuli, ITD is also used at higher frequencies, as opposed to only IID for pure tone stimuli as described above [12]. This is because the ITD is two fold, both the onset time difference and the ongoing time difference are used. The onset is the initial arrival time at the ears. This is what the brain uses for low frequencies. The ongoing time difference is continuous and is based on the the envelope of the signal, instead of the instantaneous amplitude. This is important for both low and high frequencies [13]. It is suggested that the brain uses ITD with wide-band stimuli mainly to predict all possible locations and that it uses IID and other cues to resolve confusion [11].

Although ITD and IID are the most important localization cues, they are not sufficient. Using only the ITD and IID tends to expand the size of the sound image, puts the images inside the head or produces images in the back instead of the front [14]. To resolve these issues, the human hearing system uses at least two additional cues: monaural cues and head movements.

3.3.2. MONAURAL CUES

Monaural cues refer to the cues that can be perceived using only one ear, i.e. they do not rely in interaural differences. For sound sources on the midsagittal plane (the plane perpendicular to the line between the two

ears), the signals received by the left and right ear are identical [15, p. 95][10]. To distinguish between front and back and between different elevations, acoustical filter effects of the human anatomy are used.

Most importantly, the directional filtering of the pinnae [4, p. 102/103] [15, p. 98] [10, 16, 17], but also the filter effects of the head and torso are used. The received sounds create reflections due to complex shape of the pinnae. The constructive and destructive interferences of these delayed reflections result in a comb filter [4, p. 103] [10]. The auditory system is able to recognize the different spectral properties and resolve front-back ambiguity or deduce the elevation of the source [4, p. 103] [16, 17]. This type of localization is most effective for frequencies in between 4kHz and 16kHz [10, 17].

The research in [17] has shown that the filtering of the pinnae can be approximated by simple low-pass, high-pass, band-pass and band-stop filters. A more accurate approach is to use an experimentally determined transfer function that incorporates the acoustical transfer of the head, pinnae, shoulders and torso. Applying this transfer function to an audio source can influence the perceived elevation and resolve front-back ambiguity [10, 17].

This transfer function is called the HRTF (head related transfer function). Its Fourier transform is called the HRIR (head related impulse response). The HRIR can be convoluted with the sound to obtain a signal that the brain perceives as coming from a certain direction. The IID and ITD are incorporated into the HRIR. They are present when the HRIR from the left ear together with the HRIR from the right ear are used.

HRIRs are highly individual, because the shape of the ear differs per individual. Even the two ears from the same subject can differ vastly. Personalized HRIRs have all the spectral cues for that person to accurately localize sound, but the HRIR of the person is difficult to measure. A non-individualised HRIR has the spectral cues for sound localization, but it is not optimal. The ITD and IID cues are preserved but not all of the spectral cues necessary for localization in the midsagittal plane are preserved. In addition, there are more front back confusions without personalized HRIR [18]. However, people can be trained for using a non-individualized HRIR. Subjects in the study from [18] who were trained with a HRIR that is not their own, performed better at localization and experienced less front back confusions. Non-human HRIRs based on a spherical model are only accurate for low elevations and low frequencies. Sounds like footsteps can be located using these HRIRs [19]. Individual HRIR measurements for each user are very time consuming and difficult to do due to the expensive instruments involved. There are free databases available online, ranging from 10 to 100 subjects and with variation in accuracy and number of sample points [20].

3.3.3. HEAD MOVEMENTS

In Section 3.3.1, we stated that binaural cues alone cannot resolve front-back ambiguity. Besides the cues described in Section 3.3.2, head movements are used as an important cue for resolving ambiguities [4, p. 103][15, p. 43] [10].

When binaural cues are ambiguous, we move our head in the direction of the auditory event, which is likely to coincide with the location of the source [4, p. 103][10]. The change in IID, ITD and spectral properties can be used to obtain more accurate localization [4, p. 103][10].

Some research suggests that this means of localization is relatively weak [10]. However, confusion may result when the information obtained with head movements conflicts with other localization information. For example, when using headphones, the location of the auditory event does not change when one moves his head, leading to the illusion that the sound source is in the head (internalization). Therefore, head movements play a significant role in the perceived internalization of the sound source [4, p. 103][21].

When using headphones, the head movement cue can be imitated using a head tracker. This device measures the yaw, roll and pitch angles of the head. Based on these measurements, audio signals can be adjusted in real-time to maintain a stable perceived source location [10]. The latency of such a system should not exceed 80 ms in order not to degrade localization performance [22].

3.3.4. SPATIAL RESOLUTION

The monaural and binaural cues combined with head movements makes localization possible for both the azimuthal and elevation angle. The resolution for localization is not infinite. When two sound events are close together, they will tend to be heard as originating from the same position. The minimum angle for which two events are separated is called the minimum audible angle (MAA). This angle depends on the position of the sound event, i.e. the MAA is smaller for sounds in front of the head than for events from the sides. It also depends on the frequency of the sound. The MAA has a minimum for frequencies between 250 Hz and 1000 Hz and for frequencies between 3000 Hz and 6000 Hz [23]. The MAA for wideband stimuli is between 1° and 2° in both elevation and azimuth for sound events in front of the head and increases to 10° for sounds to

the side of the head [24].

3.3.5. REFLECTIONS

Reflections are not a part of the sound localization subsystem, as defined in Section 2.1. Reflections are implemented in the reflection subsystem, described in [2]. However, reflections have a large influence on localization performance. Therefore, we will briefly discuss this influence here.

Reflections are delayed signals from a sound source that are reflected from the surroundings of the source. The collection of reflections still audible when the direct sound has ended is known as reverberation.

When the direct sound and reflections arrive at the head, only the first sound event determines the direction of the auditory event. This is known as the precedence effect [25]. It is suggested that the precedence effect is a way to resolve errors in localization [26]. But, despite the precedence effect, early lateral reflections tend to worsen localization [27], while early reflections coming from the same direction as the source improve localization.

The influence on binaural localization becomes larger for earlier reflections. When the delay of the reflection is less then 630 µs, the sound is heard as fused and will be localized between the direction of the source and the direction of the reflection. The auditory event will shift towards the source when the delay is increased to 1 ms [15, p. 222]. When the delay is between 1 ms and 5 ms, the auditory event remains at the source [28]. When the delay is increased beyond 5 ms, the auditory event will separate into two auditory events. The delay where this occurs is known as the echo threshold. The echo threshold depends on the spectrum and the level of the signal [15, p. 226/227] [28].

3.3.6. DISTANCE PERCEPTION

Distance perception can be realized by using different cues, all of which are directly related to the physics of sound propagation. The sound level decreases with distance and the difference in expected sound level and sound level, is a cue for distance [29, 30]. This cue requires some training for unfamiliar sounds.

Another cue is the energy density of the frequency spectrum. If the sounds are low-pass filtered, they are received as further away [31].

Another important cue is the direct to reverberant sound ratio [32, 33]. Sounds in reverberant environments are more often perceived as coming from outside the head (externalization), while sounds in anechoic environments are more often perceived as coming from inside the head (internalization) [21].

3.4. EFFECT OF VISUALLY IMPAIRMENT

In this section we will discuss the differences in sound localization performance and approach between visually impaired and sighted people.

It is well known that early-blind people are able to localize sound better than sighted people [34, 35]. People with residual vision are found to perform worse at sound localization than blind and sighted people [35]. The performance of early-blind people in other auditory domains differs. For example, they have a superior ability to recognize pitch changes [36], but perform worse at frequency discrimination [34].

Sighted people often generate mental imagery as a tool to visualize a spatial representation. One would think that people who have never experienced external visual stimulation are not able to generate mental imagery. However, there is evidence that early blinds are able to generate mental imagery [34, 37]. The relationship between mental imagery and visual perception is not straightforward and it is believed that mental imagery is the product of different information sources [34].

Spatial cognition allows us to orient and navigate in our environment. Both visual, auditory and haptic feedback is used for spatial cognition. When only auditory or haptic feedback is used for spatial cognition, visual feedback is still needed for calibration [34, 35, 38]. This calibration is required to get an overview of the space. While blind people must rely only on auditory and haptic feedback, their spatial knowledge is similar to or better than blindfolded sighted people [34, 35, 38]. Although the performance is similar, the approach is different. Blind people generate a route-like representation of the environment, where as sighted people create a survey-like representation [34, 39]. One of the consequences of the route-like approach is difficulty in estimating Euclidean distances [40].

Another limitation of auditory perception is that it only allows the sequential processing of information [34]. This leads to difficulty in processing simultaneous cues. Visual perception on the other hand, allows the simultaneous perception of different objects.

4

DESIGN PROCESS

In this chapter, we will discuss the implementation of an audio engine capable of generating real-time binaural sound. This implementation is based on the theory presented in Chapter 3. First an overview of the design is given in Section 4.1. Next, several important design choices are discussed in Sections 4.2 and 4.3. The real-time playback is discussed in Section 4.4. Then the implementation of several localization cues is described in Sections 4.5 to 4.7. Finally, a suitable audio source is discussed in Section 4.8.

4.1. FUNCTIONAL OVERVIEW

In this section we will explain the structure of the audio engine and how this structure follows from the requirements in Section 2.2. A block diagram of the audio engine is shown in Fig. 4.1. Each block in the diagram corresponds to a software class or module.

First, a WAV file is loaded from the file system and stored in a track, shown by the WAV and track blocks in Fig. 4.1. This satisfies requirement 1.1. The track then offers various methods for manipulation by the gameplay subsystem, such as resetting and seeking, as described by requirement 2.3.

When considering reflections, multiple copies of the same track should be created. Each of these copies, or channels, has a different location and attenuation. These copies are described in requirement 2.3. All signals up until now are mono signals. In order to create stereo signals, each of the channels is localized, shown by the localize blocks in Fig. 4.1. The localization module encodes localization cues in the intensity and phase difference between the left and right signal. The cues are also encoded in the frequency spectrum of the signal. These mechanisms are described in detail in Section 4.6.

To accomplish real-time processing (requirement 1.2), audio is processed in small blocks. The term block refers here to a sequence of samples and should not to be confused with a block in the block diagram. The small audio blocks are retrieved sequentially from the track in real-time, localized with the coordinates at that

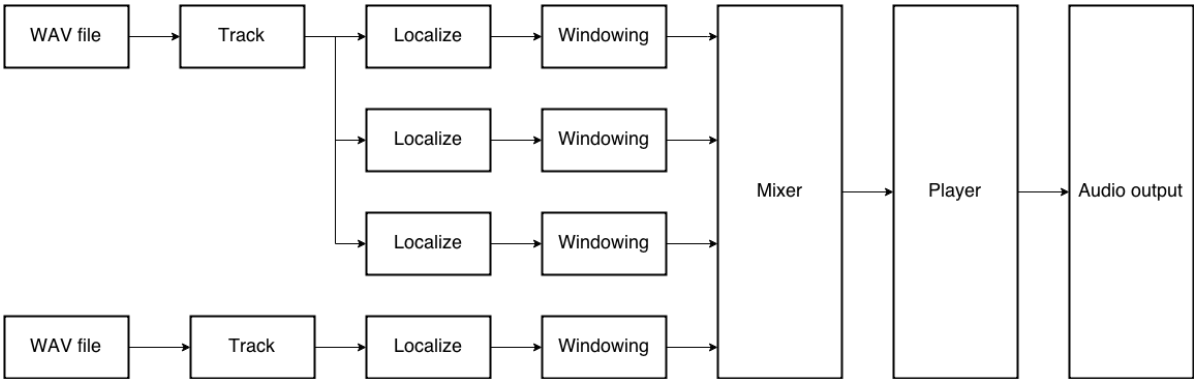

Figure 4.1: Block diagram of the audio engine. WAV files are loaded into a track. Multiple copies of each track (channels) are localized and windowed before being mixed together. The mixed signal is streamed to the audio output by the player.

time. When walking away from a sound source, the propagation delay of each block increases. This results in small gaps in the resulting sound vector when the blocks are played sequentially. This can be fixed by applying a window to the audio signal, shown in the windowing block in Fig. 4.1. This accomplishes smooth audio playback (requirement 1.3). The windowing is explained in Section 4.4.

Next, all the windowed channels have to be merged into a single stereo sound signal. This is done by the mixer, shown in Fig. 4.1 by the mixer block. Furthermore, the mixer handles the increased block size due to convolutions and delays applied by the localization algorithm.

Finally, each stereo block produced by the mixer is played sequentially by the player, show by the player block in Fig. 4.1. The player, implemented with the PyAudio python module [41], streams the audio to the audio output device. A discussion of the choice of audio output device is found in Section 4.3.

4.2. PROGRAMMING LANGUAGE

The programming language used can have a huge influence on the final product. In this section we will motivate our choice of programming language.

The development of the sound localization subsystem is focused on the signal processing required for sound localization. Likewise, the development of the reflection subsystem is focused on signal processing and linear algebra [2]. The project only has a short time span (requirement 2.1), therefore a programming language with available scientific computing libraries is preferred. For scientific computing, there are two commonly used programming environments:

- The MATLAB language and environment [42].
- The Python [43] programming language in combination with the SciPy [44] ecosystem.

MATLAB is numerical computing environment. It consists of a programming language, IDE (Integrated Development Environment) and a large library for data processing. MATLAB is commercial software, but is available for TU Delft students.

Python is a general purpose programming language. It contains a standard library containing many general purpose programming functions, but no scientific computing functions. SciPy is commonly used to obtain scientific computing functionality in Python. SciPy is an ecosystem consisting various Python modules (Numpy, SciPy library, Matplotlib). With SciPy, Python has about the same scientific computing capabilities as the MATLAB programming language. Python is free and open source, making it available for everyone.

Both development environments are available without additional cost, satisfying requirement 2.2. We have chosen to use Python because of two main reasons. First, Python contains more general purpose programming functionality than MATLAB, such as multithreading support. Especially for the gameplay subsystem, a general purpose programming language is more suitable. Second, Python has a larger ecosystem of third party modules. Such modules include dedicated packages for game development [45], GUI design and audio playback [41].

4.3. AUDIO OUTPUT DEVICE

The audio signal from the PC has to be transformed into a actual sound. The audio output device is responsible for this conversion. We have considered the following options for the audio output device:

- Headphones
- Loudspeakers

Sound coming from loudspeakers is convoluted with both the head-related impulse response and the room impulse response before reaching the tympanic membrane. The advantage of the convolution with the real room impulse response is that sound is perceived as coming from the outside the head [21]. It is also experienced as more dynamic [21].

However, both the left and right ear receive the signal emitted from all loudspeakers. This is known as cross-talk and complicates localization. This can be solved using cross-talk canceled stereo [46], but this makes localization dependent on the user's position in the room, compromising requirement 1.7. Furthermore cross-talk cancellation affects the spectral properties of the sound.

Sound played through headphones are directly in ear and undistorted by the room impulse response as well the head related transfer function of the user. But there are downsides to using headphones, mostly the

lack of bone conducted sound and the induced resonances, due to the blocking of the ear canal. This possibly degrades localization performance.

Because of the arguments given above, we decided to use headphones as the audio output device. However, the transfer function of the headphones contains deep spectral notches similar to the ones in HRTFs, possibly compromising elevation cues [47]. The distortion of the signal can be measured and compensated, but this is different for each headphone. We will not compensate this distortion, because each user will use different headphones.

4.4. REAL-TIME AUDIO PLAYBACK

One of the most important aspects of a typical video game is the fact that audio and visual feedback change in real-time as a response to user input. Therefore, it should be possible to perform real-time manipulations on the audio signal. This is also described by requirement 1.2. This can be realized by dividing the input signal into blocks of a predefined size. Each of these blocks is processed in real-time and sequentially streamed to the audio output device. However, there are some problems with this approach.

The first problem is that the block size k is increased during processing. Convolution with an impulse response is a typical operation that results in an increased block size. This problem can be solved in the mixer by adding all the different channels to a single buffer of size $p > k$. The first k samples of the buffer form the resulting block. Next, the buffer is shifted such that sample k becomes sample 0. This way, any samples exceeding the block size are added to the next block.

Secondly, the audio content resulting from two consecutive blocks might not connect perfectly in the time domain. This happens when the player is moving relative to the audio source. In this case, the propagation delay of consecutive blocks differs, creating empty gaps or overlap in the resulting audio signal. This effect is undesired, because it will result in audible clicks in the output signal, compromising requirement 1.3. Specifically for propagation delay, this effect can be seen as a discrete manifestation of the Doppler effect. The Doppler effect describes the frequency shift of a wave when the observer is moving relative to the wave source [3, p. 518].

A physically accurate solution to the problem is to resample the audio signal to the desired length. Not only does this realize the frequency shift, but it also ensures that the blocks perfectly connect. In order to satisfy requirement 1.2, blocks must be processed at a frequency of at least 10 Hz. For a single sound source in a cubic room with second order reflections, this results in 37 stereo audio channels [2]. The left and right signals have a different propagation delay, resulting in $10 \times 37 \times 2 = 740$ resampling operations per second. Resampling is a fairly CPU intensive operation and in order to satisfy requirement 3.2, we decided not to choose this option.

As a faster alternative, we chose to use overlapping blocks multiplied with a window function. All blocks fetched from the audio track are l samples larger than the block size k in order to ensure that two consecutive blocks are always overlapping. After processing, block i has a length of $m_i \geq k + l$ with a propagation delay d_i. The propagation delay d_i is not included in the audio signal m_i output by the processing algorithm, but is output as a separate variable.

The constraint on l, in order to ensure blocks are always overlapping, depends on the block size k, processed block length m_i and the maximum velocity of the player relative to the audio source v_{max}. For a constant m_i, this is given by Eq. (4.1), where c is the speed of sound. For $v_{max} = 2\,\mathrm{m\,s^{-1}}$, $c = 344\,\mathrm{m\,s^{-1}}$ and $k = 4096$, this gives $24 < l < 4072$.

$$\frac{v_{max}}{c} k < l < \left(1 - \frac{v_{max}}{c}\right) k \tag{4.1}$$

The overlap between two consecutive blocks is illustrated in Fig. 4.2. From this figure, we can deduce that the overlap of the audio signals is given by Eq. (4.2):

$$c_{1,2} = m_1 - k + d_1 - d_2 \tag{4.2}$$

Or more generally, for the overlap between block i and $i+1$ by Eq. (4.3):

$$c_{i,i+1} = m_i - k + d_i - d_{i+1} \tag{4.3}$$

Next we multiply the block with a window function. A window function is a function that is zero valued outside a specified interval. Perfect signal reconstruction is possible when two partially overlapping window functions sum up to one. A simple window function that satisfies this criteria is given by Eq. (4.4). We have

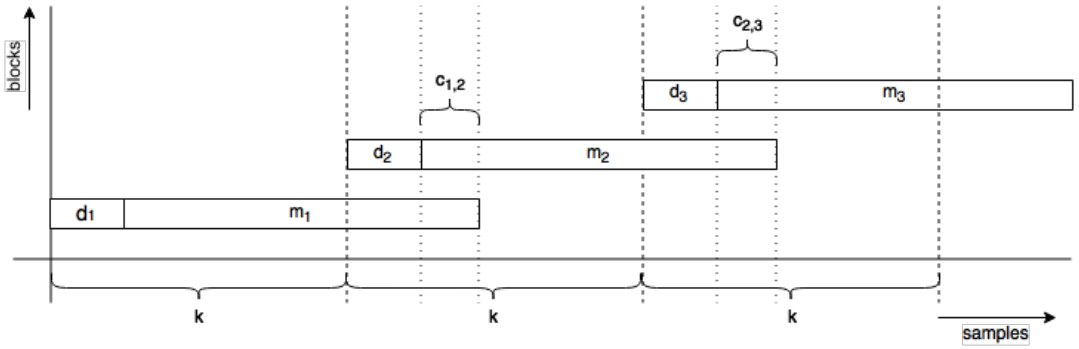

Figure 4.2: Shows the overlap between consecutive audio blocks. k denotes the block size, n_i the propagation delay of block i, m_i the audio size of block i after processing and $c_{i,i+1}$ the overlap between the blocks i and $i+1$.

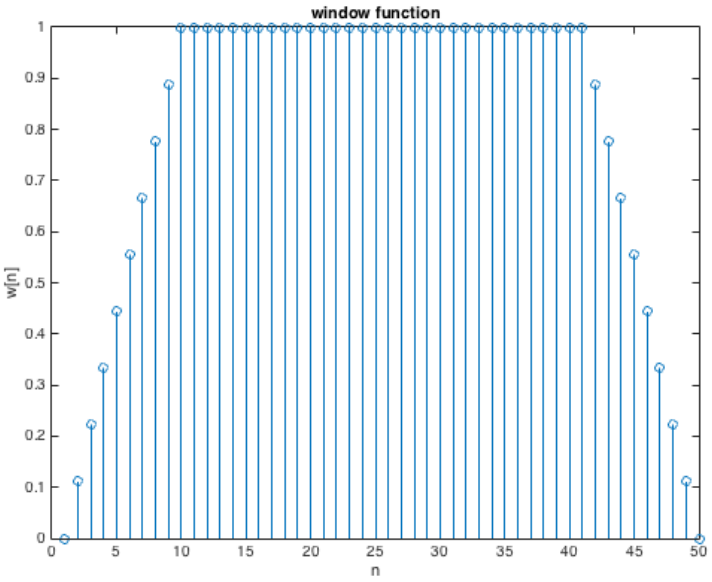

Figure 4.3: A plot of the window function $w[n]$ described by Eq. (4.4), for $k = 50$ and $c_{i-1,i} = c_{i,i+1} = 10$.

chosen to use this window for its simplicity. However, different window functions, such as a Hanning window, might have better spectral properties.

$$w[n] = \begin{cases} \frac{1}{c_{i-1,i}} n & \text{if } n < c_{i-1,i} \\ 1 & \text{if } c_{i-1,i} < n < c_{i,i+1} \\ \frac{m_i}{c_{i,i+1}} - \frac{1}{c_{i,i+1}} n & \text{if } n > c_{i,i+1} \end{cases} \tag{4.4}$$

4.5. DISTANCE

As explained in Section 3.1.2, distance results in a propagation delay, sound attenuation and sound absorption.

The speed of sound depends on the air composition and temperature, as given in Eq. (3.2). Those variables were set at the outdoor air and a temperature of 20 °C. The resulting speed of sound is $344\,\mathrm{m\,s^{-1}}$.

This finite speed results in a propagation delay, the time that the signal takes to arrive at the ear. Equation (4.5) gives the number of samples propagation delay n that corresponds to the propagation delay.

$$n = F_s \frac{r}{c} \tag{4.5}$$

n = propagation delay (samples)
F_s = sample frequency (44.1 kHz)
r = distance between source and player (m)
c = speed of sound (344 m s^{-1})

The intensity decreases with the distance traveled, because the intensity of a isotropic source is spread across the surface of a sphere. The intensity at the player is given in Eq. (3.3). The Q, W and 4π are all constant and do not change over time or distance. The only relevant variable is r, the distance between the source and the player. The intensity at the player is thus a constant multiplied by r^{-2}.

The last effect of the distance between the source and the player is the sound absorption for high frequencies due to the interaction with air molecules. This effect is only noticeable at high frequencies and large distances. Because the rooms in the game are small (less then 10 m), we decided not to implement sound absorption in the game.

4.6. SOUND LOCALIZATION

In this section we propose an implementation of the techniques discussed in Section 3.3. This implementation will be used in the audio engine of a game for visually impaired children.

The brain uses multiple localization cues to localize a sound at a point in space, but the ITD and IID are the most commonly used cues. For more accurate localization, the head-related impulse response (HRIR) is used. This can be implemented with a convolution, which takes relatively long to compute. However, the IID and ITD are significantly faster to compute, because they are based on simple formulas. Furthermore, the IID and ITD are less dependant on the user's anthropometric measurements.

We implemented both the HRIR and the binaural cues (IID and ITD), because they both have their own strengths and weaknesses. The precedence effect, described in Section 3.3.5, states that only the direct sound is used for localization. There are also many more reflections than direct sound sources. Therefore we chose to apply the ITD and IID to reflections and the HRIR to the direct sound. This is a trade-off between localization accuracy and the CPU-usage, described by requirements 3.2 and 1.4.

4.6.1. INTERAURAL TIME DIFFERENCE

The interaural time difference (ITD) describes the difference in arrival time between the two ears, as explained in Section 3.3.1. Equation (4.6) gives the delay in samples for the left (−) and right (+) signals. It is a function of the angle of incidence and half the distance between the two ears (typically 11.5 cm [4, p. 98]). The delay is passed to the windowing module described in Section 4.1.

$$\mathrm{n_{ITD}} = \pm F_s r \frac{\theta + \sin(\theta)}{2c} \tag{4.6}$$

$\mathrm{n_{ITD}}$ = delay for left (−) and right (+) signals (samples)
F_s = sample frequency (44.1 kHz)
r = half the distance between the ears (11.5 cm)
θ = angle with the midsagittal plane (rad)
c = sound velocity (344 m s^{-1})

4.6.2. INTERAURAL INTENSITY DIFFERENCE

The interaural intensity difference (IID), as explained in Section 3.3.1, depends on the frequency of the sound and the angle of incidence. In this section we propose an implementation of the IID cue. We have chosen to neglect the frequency dependence in our implementation. This made it easier to implement (requirement 2.1) and made it less CPU consuming (requirement 3.2). We have chosen a frequency of 4 kHz. The reason for this was that the human ear is most sensitive in this frequency range [48, p. 71]. Furthermore, this frequency is above the 3 kHz threshold, where the IID is used by the brain, as stated in Section 3.3.1. The intensity as a function of the angle of incidence is shown in Fig. 4.4. The IID is implemented by approximating the function in Fig. 4.4. This approximate function is given in Equation (4.7), showing the intensity difference between the left and right ear. The arctangent approximates the curve shown in Fig. 4.4. The $\frac{\pi}{4}$ and $\frac{2}{\pi}$ constants are experimentally determined to obtain an approximately fitting curve.

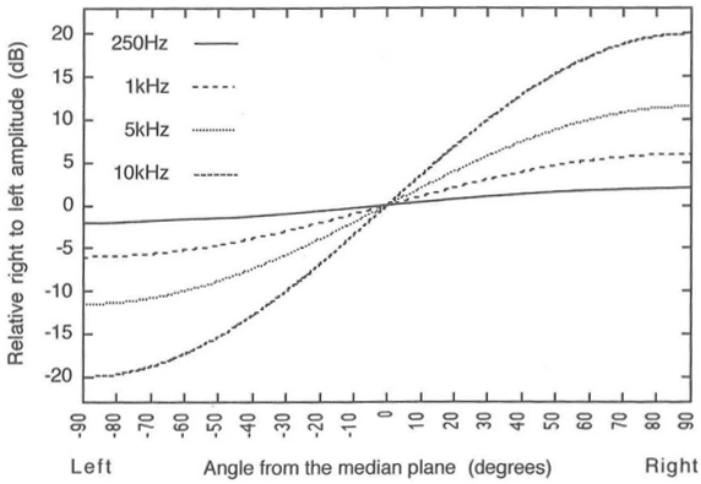

Figure 4.4: The interaural intensity difference for different frequencies
Source: [4, p. 102]

$$\text{IID} = 10\,\text{dB} \cdot \frac{2}{\pi} \cdot \arctan\left(\frac{\theta}{\frac{\pi}{4}}\right) \tag{4.7}$$

θ = the angle of incidence on the azimuth (rad)
IID = the intensity gain (left ear) and drop (right ear) (dB)

4.6.3. HEAD-RELATED IMPULSE RESPONSE

Localization should be possible in azimuth and elevation (requirement 1.4) and users should be able to distinguish between sound sources in front and behind the player (requirement 1.5). ITD and IID only give auditory cues for azimuth angles between −90° and 90° and all sounds originating from the midsagittal plane are identical. The head-related impulse response (HRIR) contains spectral cues to distinguish between all directions. The HRIR can be obtained by either measuring it or using an available measurement from an database. To limit the amount measurements required to create the HRIR, the data can be interpolated such that neighbouring coordinates are less than one degree apart. This is smaller than the minimum audible angle (MAA). However, HRIRs can become very large. In order to satisfy requirement 3.3, which states that memory usage should be limited, the HRIRs are truncated such that auditory cues are largely preserved, but memory usage is decreased.

MEASUREMENT

Measuring the head-related impulse response is difficult for various reasons. First, the instruments required are expensive. Second, the HRIR must be measured for each angle, making it a time consuming process. Furthermore, the HRIR is highly individualized, requiring different HRIR measurements for each user. Because of these reasons, we chose to use a HRIR from an existing database. There are only a small number of databases online and not all are well documented. Most contain only a small number of subjects with sparsely sampled data points in space. The database of CIPIC was chosen because of its extensive documentation, the number of subjects and the included anthropometric measurements.

This database contains HRIR measurements for 45 subjects at 1250 positions [49]. The HRIRs in the CIPIC database are measured in an anechoic chamber with the subject at the center of a 1 m radius circle with the axis aligned with the subject's ear axis. Loudspeakers are mounted on the circle and transmit a signal. The signal is recorded with microphones inside both ears. The output was digitized at 44.1 kHz with 16 bit resolution. Each sample is around 4.5 s long, containing 200 time-samples. The positions are measured for 25 azimuth angles and 50 elevation angles.

The CIPIC database also contains 27 different anthropometric variables for each subject, 10 for the ear shape and 17 for the shape of the head and torso. The HRIR is a function of these variables.

Figure 4.5 shows the HRIR and HRTF for the left and right ear at −80° azimuth (left) and −45° elevation (below) for a subject from the CIPIC database. The difference in arrival time for the first peak of the left and

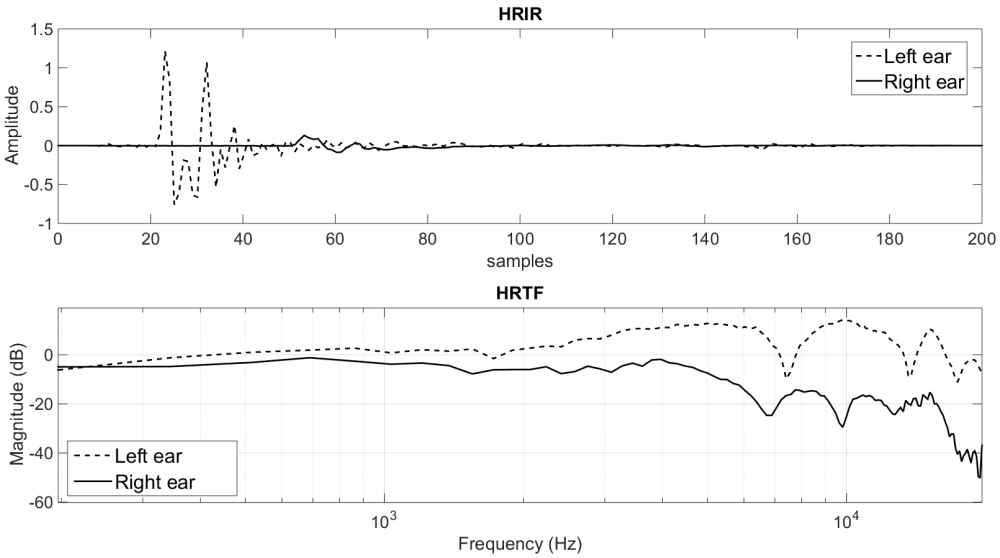

Figure 4.5: The HRIR and HRTF at −80° azimuth (left) and −45° elevation (below).

the right ear is the ITD. The IID is the magnitude difference between the left and right ear. This difference is very small for frequencies below 2 kHz. The IID and ITD show the characteristics that are described in Chapter 3.

INTERPOLATION

The database from CIPIC contains a total of 1250 positions, 50 for the elevation angle and 25 for the azimuth angle. As explained in Section 3.3.4, the MAA in the front is around 1° for azimuth and 2° for elevation. This MAA cannot be reached with the available HRIR. To obtain optimal spatial resolution, the HRIR can be interpolated. This will result in a HRIR where the angle between two adjacent impulse responses is smaller than the MAA.

The HRIR contains all auditory cues required for localization, including the ITD. The ITD causes a time shift between two adjacent impulse responses. Before the HRIR can be interpolated, it must first be aligned to account for this time shift. The time shift, or ITD, is reapplied after interpolation.

There are two ways to align the HRIRs. The first is peak detection, which is used by the researchers from CIPIC. The second method is more common in literature and uses the cross correlation function [50]. Peak detection is easier, because only the first peak higher than a predefined threshold has to be found and the HRIR can be aligned based on that location. The cross correlation method searches for the time shift that, when applied to the a HRIR, results in the highest correlation with the adjacent impulse response.

The alignment in this thesis was done with peak detection, because of its easier implementation. After peak detection, all the HRIRs were shifted until the first peaks were all aligned. The alignment also gets rid of the propagation delay, which was present in the original CIPIC data. This allows for more effective windowing, as described in Section 4.4. The ITD is cue is missing from the HRIR after alignment. Because the ITD is an important cue for localization, it is reapplied using the methods described in Section 4.6.1.

The next step is to interpolate the HRIR. Interpolation is a method to construct new data points within the range of known data points. This is done by analysing neighbour values and using appropriate curve fitting to find the value of the intermediate data points. This is commonly done, because the data is scarcely sampled and more data points need to be known. While there are multiple ways to interpolate data, the common methods for HRIR are:

- Linear interpolation
- Spline interpolation
- Sine interpolation
- Sinc interpolation

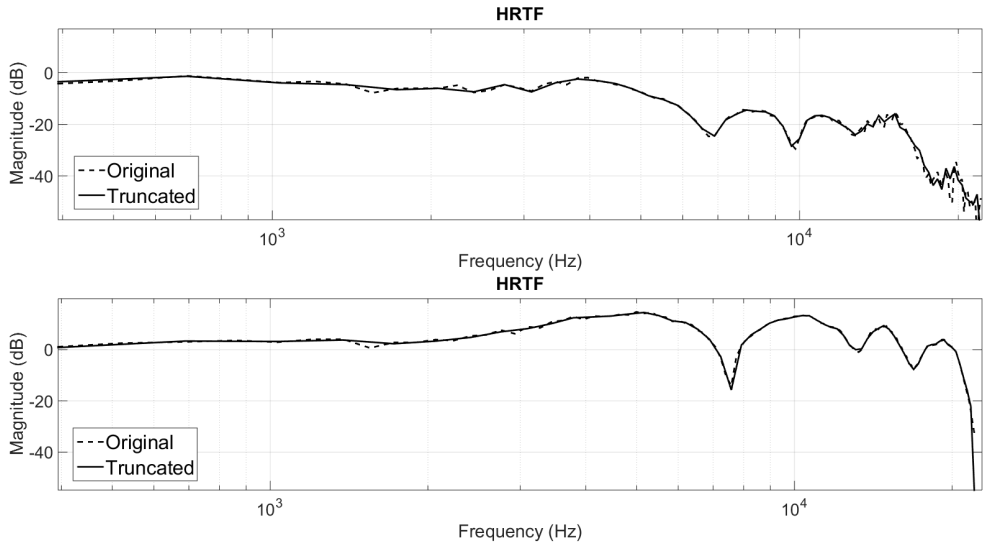

Figure 4.6: The original and truncated HRTFs at (−80° −45°) (above) and(−80° −45°) (below).

The simplest method is linear interpolation. The value of the two nearest sources and the distance to those sources are used to calculate the value on the new position using Eq. (4.8). The interpolant is a straight line between two points and the intermediate values lie on that line. The derivative of the interpolated function is not continuous at the original sample points. However, out of the methods given above, linear interpolation results in the smallest error [51, 52].

$$y_2 = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0} \tag{4.8}$$

TRUNCATING

A single interpolated HRIR can become very large. The size of a 16 bit HRIR, specified at every 1° interval in both azimuth and elevation with an impulse response length of 200 samples is $16 \times 360 \times 180 \times 200 = 24.7$ MB. Because the HRIR size might compromise CPU or memory requirements (requirements 3.2 and 3.3), we decided to truncate the HRIR. The first 50 samples of the impulse response following the first peak contain most of the information. The other 150 samples do not contain much spectral information. We decided to truncate the impulse response from 200 samples to 100 samples. Figure 4.6 shows the Fourier transform of both the original HRIR from the database and the truncated HRIR. As seen in the figures, most spectral cues are preserved after truncating. The cues are better preserved for the ear closest to the direction of the sound.

PERSONALIZED HRIR

A non-personalized HRIR can be used for good localization performance on the horizontal plane, but the performance for the midsagittal plane can be dramatically worse [53]. This is because the filter effect of the pinnae is stronger for sound in the midsagittal plane. This problem can be overcome by using a personalized HRIR, but measuring HRIRs is not a practical option. An easier way is to try to match the ear shape to the closest person in the database. This can be done by measuring the user anthropometric variables manually or with image processing. In the database there are 26 different variables, but localization is improved when the matching is based on only 8 variables [54]. As an alternative approach, these 8 variables can be used to compute the HRIR. Although both these options are outside the scope of this project, they are interesting for future research.

4.7. HEAD TRACKING

In this section we will discuss the design of a system that imitates the head movement localization cue described in Section 3.3.3.

The localized sound created using the techniques described in Section 4.6 proved to be successful in localizing sounds laterally, but offers moderate performance in resolving front-back ambiguity. Also sounds

often appear to originate from inside the head. A more in depth evaluation of the performance of these techniques can be found in Chapter 5. In order to still satisfy requirement 1.5, we decided to implemented a head tracker as an additional method to resolve front-back ambiguity. Furthermore, we expect that increased externalization would increase the level of realism experienced by the user.

Such a system typically consists of the following:

- A sensor measuring the yaw, roll and pitch angles of the users head.
- Signal processing, processing the raw sensor input into yaw, roll and pitch angles.
- An algorithm that modifies the apparent location of the auditory event in real-time, using the techniques described by Section 4.6.

In the following paragraphs, we will focus on the first two components, the sensor and the signal processing, hereafter referred to as the *head tracker*.

All head trackers encountered for our research can be categorized as following:

- Infrared based head trackers.
- Face recognition based head trackers.
- Inertial head trackers.

Infrared based trackers use multiple infrared LEDs attached to the users head. A stationary receiver, for example a webcam, recognizes and tracks the LEDs. All the angles can be calculated based on the LED's position on the screen. Examples of systems that use this approach are TrackIR [55] and FreeTrack [56].

Similar to infrared based trackers are face recognition based trackers. These trackers do not rely on infrared LEDs, but use pattern recognition techniques to perform head tracking. Opentrack, which contains the OpenCV based Human Face Tracker, is an example of head tracking software that uses this technique [57].

The third option is the inertial head tracker. These trackers contain an accelerometer and gyroscope, and often a magnetometer to eliminate drift. Examples in this category include EDTracker [58].

Requirement 3.1 states that no special hardware should be required for the game. Therefore, special infrared LEDs and inertial head trackers are not preferred. This leaves the face recognition based head tracker, as it requires only a webcam. A webcam is integrated into most modern laptops, satisfying requirement 3.1. However, this approach has two major drawbacks. These drawbacks also apply to the infrared based head tracker. The first drawback is the dependence on the user's position and orientation. The user's head should always be in the webcam's field of view. This dependence partially compromises requirement 1.7. Second, webcams typically have a latency in the order of magnitude of 100 ms. Therefore, requirement 1.2, which states that the latency must be limited to 100 ms, might not be satisfied.

Because of the limited development time (requirement 2.1), we have chosen for the option that required the least additional hardware. This is the Opentrack head tracking software, using Human Face Tracker (ht tracker). It is a face recognition based tracker, requiring only a webcam. Opentrack handles all webcam interfacing, image processing and filtering. The yaw, roll and pitch angles are written to a virtual joystick device, emulated with VJoy Virtual Joystick Driver [59]. The virtual joystick is read in the game using the PyGame python module [45].

4.8. AUDIO SOURCE

An important aspect for localization is the audio clip which will be played. First we will explain what the requirements are for the audio source based on the theory in Chapter 3. Next we will give an example of a suitable audio source.

The user should be able to localize the sound in both azimuth and elevation (requirement 1.4). As shown in Chapter 3, a combination of different localization cues is required to achieve this goal. The different localization cues are effective at a limited range of frequencies, hence the audio source should contain a wide range of frequencies. The MAA is small for frequencies between 250 Hz and 1000 Hz when ITD cues are most effective and for frequencies between 3000 Hz and 6000 Hz when IID cues are most effective. Localization in the midsagittal plane requires the use of spectral cues, requiring frequencies in between 4000 Hz and 16000 Hz.

The duration of the sound should exceed 250 ms to achieve accurate localization [15, p. 95]. When head movements are taken into consideration, the duration should exceed 350 ms in order for head movements to have a possible influence on the localization accuracy [15, p. 95].

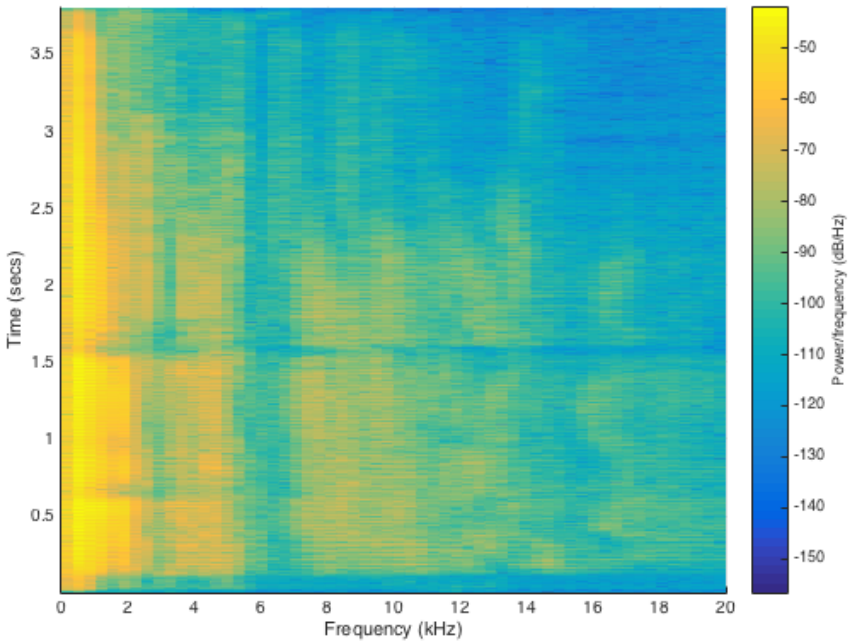

Figure 4.7: A spectrogram of an audio clip containing the sound a singing women.

The choice for the audio clip for the game is part of the gameplay subsystem and is not described in this thesis. More information can be found in [1]. However, we will describe the audio clip used for testing the sound localization and obtaining the results in Chapter 5. The chosen audio clip contains a singing women and has a duration of 3.8 s, well above the 350 ms mark. A spectrogram is shown in Fig. 4.7. It can be seen that it has sufficient spectral content in the frequency ranges discussed above, making it a suitable audio source.

5

RESULTS

This chapter contains results of experiments conducted to verify whether the implementations described in Chapter 4 satisfy the requirements in Section 2.2.

5.1. REAL-TIME AUDIO PLAYBACK

Two experiments were conducted to test the audio playback engine.

In the first experiment, a stationary sound source was used. The transmitted audio signal was localized, windowed, mixed and fed to the player, as described in Section 4.1. All the blocks passed to the player are concatenated to obtain the complete audio vector. This audio vector was compared with the audio vector created by localizing the source signal as a whole. Both audio vectors are identical.

The second experiment was conducted to check whether requirement 1.3 was satisfied. Requirement 1.3 states that the sound should not contain any audible stuttering or clicks. This was tested by playing the audio signal, resulting from the player moving relative to the sound source, and asking six test subjects whether they think it contains any stuttering, clicks or other artifacts. None of them noticed anything.

5.2. DISTANCE

The distance implementation in Section 4.5 is comprised of two operations. The first is the propagation delay. When the sound localization subsystem was integrated with the other subsystems (gameplay and reflections), reflections proved to work as expected [2]. As the propagation delay is important for correct reflections, we can conclude that the propagation delay is working as expected. The second operation is distance attenuation. The decrease in amplitude when the player moves away from the source is clearly noticeable.

5.3. SOUND LOCALIZATION

The implementation of the sound localization engine described in Chapter 4 was tested with multiple test subjects. This was done by randomly selecting a sound source location from a set of predefined coordinates. The sound described in Section 4.8 was localized using the methods described in Section 4.6 and output to a headphone. The test subject was presented with a multiple choice option of the source position. This process was repeated 20 times per test subject, for 7 different subjects. The test has been performed for the implementations listed below, in that order.

- Only IID, as described in Section 4.6.2.
- Only ITD, as described in Section 4.6.1.
- IID and ITD combined, as described in Sections 4.6.1 and 4.6.2.
- HRIR from CIPIC database, as described in Section 4.6.3.
- Processed HRIR. The HRIR is aligned, interpolated and truncated, as described in Section 4.6.3. The ITD is added as described in Section 4.6.1.

The options available to the test subject for the ITD, IID and combined ITD and IID test are:

- Front
- Left, 30° from midsagittal plane
- Left, 90° from midsagittal plane
- Right, 30° from midsagittal plane
- Right, 90° from midsagittal plane

The options available to the test subject for the HRIR tests are:

- Front
- Back
- Left front, 30° from midsagittal plane
- Left back, 30° from midsagittal plane
- Left, 90° from midsagittal plane
- Right front, 30° from midsagittal plane
- Right back, 30° from midsagittal plane
- Right, 90° from midsagittal plane

The results are summarized in Table 5.1. The table lists the percentage of the 20 questions that were correctly answered by the test subject. The number in parenthesis gives the percentage of correct answers when front-back confusions are not taken into account. This allows a fair comparison between the IID/ITD tests and HRIR tests.

Table 5.1: Percentage of locations correctly answered for different test subjects and localization techniques. The number in parenthesis denotes the percentage of correct answers when the distinction between front and back is not taken into account.

Test subject	IID (%)	ITD (%)	IID + ITD (%)	HRIR from database (%)	Processed HRIR (%)
A	85	65	65	40 (60)	65 (80)
B	70	45	70	60 (80)	65 (85)
C	70	85	85	65 (85)	55 (90)
D	80	80	80	45 (65)	60 (80)
E	100	85	65	70 (95)	50 (90)
F	85	55	85	55 (100)	55 (85)
G	75	75	100	35 (75)	60 (90)
Mean	81	70	79	53 (80)	59 (86)

First, we see that all localization types have a success higher than 70% for lateral localization, i.e. when neglecting front-back confusions. Second, we see that IID outperforms ITD, with the combined IID and ITD resulting in performance similar to only IID. When comparing the lateral localization performance of the HRIR with the combined IID and ITD, we see that the addition of spectral cues results in similar or slightly better performance. Furthermore, subjects were better able to estimate the location with the processed HRIR in comparison with the original HRIR from the database. Lastly, we notice that front-back confusions are very common when using the HRIR.

Test subjects reported that the sound source appears very close, even inside the head.

5.4. HEAD TRACKER

Next we discuss the testing of implementation of the head tracker described in Section 4.7. One of the main reasons for the implementation of the head tracker was to improve the ability to resolve front-back ambiguities, as stated in requirement 1.5. Therefore, the first experiment is to test whether requirement 1.5 is satisfied. This was done similarly to the experiment in Section 5.3. A sound source was randomly placed either directly in front or directly behind the player. The test subject then selects whether the sound source is perceived in front or behind the player. This is repeated 20 times without head tracker and 20 times with head tracker. The results are summarized in Table 5.2.

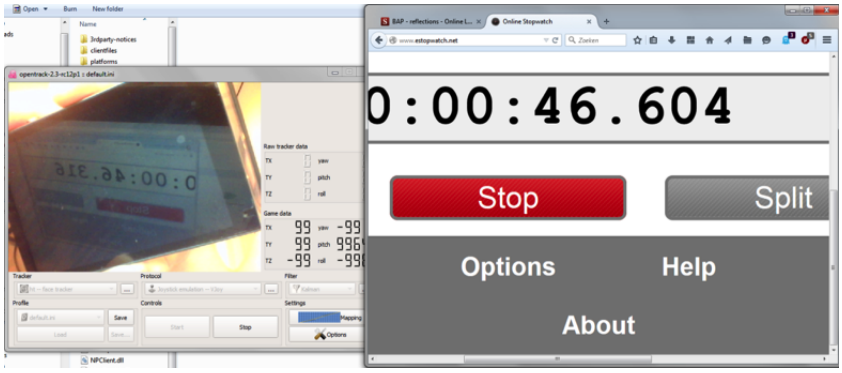

Figure 5.1: Screen capture illustrating the experimental setup used for measuring webcam latency. The webcam latency in this screen capture is 46.604 – 46.316 = 0.288 s.

Table 5.2: Percentage of locations correctly answered for different test subjects with and without head tracking.

Test person	Without head tracker (%)	With head tracker (%)
A	55	95
B	60	100
C	55	85
D	60	90
E	45	100
F	45	85
G	65	100
Mean	55	94

Table 5.2 shows that the number of front-back confusions is significantly decreased when using a head tracker. Without head tracker, the performance is only slightly better than the 50 % expected with random guessing. With head tracker, the number of correct localizations increases up to 100 % for some test subjects. However, test subjects reported a significant lag in the location of the auditory event in response to head movements. Also, head tracking was lost when the head was rotated more than $80 \pm 10°$ with respect to the midsagittal plane, when the test subject was wearing glasses or when when the head was moved rapidly.

In response to the reported lag, we conducted an additional experiment in order to verify whether requirement 1.2 is satisfied. Section 3.3.3 reports that latency larger than 80 ms is typically perceived as lag. The purpose of the experiment was to measure the latency from a head movement to the head tracking software Opentrack. Because it is expected that this would cause the most significant latency, the latency from transmitting the information from Opentrack to Python was not taken into account. The experiment was performed on two different laptops, both with build-in webcam. It was conducted as following. A stopwatch application with 1 ms resolution was opened on the screen, with Opentrack, showing the video feed from the webcam, next to it. A mirror placed in front of the webcam results then in two different times on the screen. The first is the stopwatch time, the second is the stopwatch time reflected by the mirror and captured by the webcam. The difference between the times equals the webcam latency. This is illustrated in Fig. 5.1. The results are summarized in Tables 5.3 and 5.4.

Table 5.3: Measured webcam latency of a MacBook Pro 13-inch (early 2011) running Windows 7.

Test	Latency (ms)
1	368
2	362
3	320
4	406
Mean	364

Table 5.4: Measured webcam latency of a HP EliteBook 8560W running Windows 7.

Test	Latency (ms)
1	288
2	174
3	348
4	290
5	348
Mean	290

The measurements in Tables 5.3 and 5.4 show that the webcam latency of the two laptops is 364 ms and 290 ms.

5.5. RESOURCE USAGE

Requirements 3.2 and 3.3 limit the CPU usage to 80 % on a 2.3 GHz Intel Core i5 and the memory usage to 500 MB. This concerns only the sound localization subsystem. The integrated game, consisting of the sound localization, reflection and gameplay subsystems, uses 240 MB of memory and 43 % CPU on a 2.3 GHz Intel Core i5. This is the most accurate scenario for measuring the usage of computational resources. The sound localization subsystem described in this thesis uses less resources than the integrated game.

6

Discussion

In the discussion, we will interpret the results presented in Chapter 5 and discuss whether these results satisfy the requirements given in Section 2.2.

6.1. Real-time audio playback

In Section 5.1, it was reported that the real-time audio playback is both working reliably and without audible artifacts, satisfying requirement 1.3. Furthermore, the configurable block size described in Section 4.4 allows the latency to be controlled such that the audio playback satisfies requirement 1.2.

6.2. Distance

Section 5.2 states that the decrease in sound amplitude when the player moves away from the source is clearly noticeable. Therefore, we can conclude that distance perception is functional, satisfying requirement 1.6. However, a quantitative test was not conducted.

Distance perception in the human auditory system is mainly based on two cues. The first is intensity attenuation, whose implementation is described in Section 4.5. The second is the direct to reverberant sound ratio, described in Section 3.3.6. As this requires the implementation of reflections, it was not considered in this thesis. Therefore, the accuracy of the distance perception described in this thesis is limited. The influence of the lack of reflections is supported by the fact that test subjects reported an increased sense of internalization in Section 5.3. Lack of reverberant sound is a possible cause of internalization, as described in Section 3.3.6.

6.3. Sound localization

Although perception of the elevation of sound was part of the initial requirements (requirement 1.4), no experiment was conducted to test this quantitatively. There are several reasons for this. First, the importance of elevation in the integrated game was small. Second, expectations were very low based on the difficulty in resolving front-back ambiguities and the degradation of elevation cues caused by uncompensated headphones (Section 4.3). However, elevation cues are incorporated into the HRIR, therefore some amount elevation perception is expected.

In contrast the elevation cues, lateral localization was extensively tested. The results in Section 5.3 show that lateral localization performs well ($> 70\,\%$ correct) for all tested localization techniques. Localization based on only IID performs better than only ITD. This might be explained by the fact that the two cues contradict each other, confusing the test subject. The audio source signal contains frequencies in both the range where ITD cues are effective and the range were IID cues are effective, as explained in Section 4.8. Yet, the IID cue might be dominant. This also explains the similar performance when ITD is added to the IID cue. Only IID results in $81\,\%$ correct localization, IID and ITD combined results in $79\,\%$ correct localization.

Furthermore, the results in Section 5.3 show that the lateral localization performance when using the HRIR is similar to when only binaural cues are used, i.e. the combined IID and ITD. This is expected, as the HRIR contains both these binaural cues. Surprisingly, the processed HRIR performs better than the original HRIR from the CIPIC database. Interpolation should not influence the performance in the experiment in

Section 5.3, as the different locations are at least 30° apart. Truncation should only worsen localization performance. A possible explanation is the fact that test subjects were first subjected to original HRIR experiment and then to the processed HRIR experiment. This gives them the possibility to learn and adapt, performing better with the processed HRIR. Nevertheless, we can conclude that the performance loss from truncation is small.

The ability to resolve front-back ambiguities was tested in both Sections 5.3 and 5.4. It can be concluded that it is very difficult to resolve front-back ambiguities by spectral cues alone. Table 5.1 shows that localization performance decreases drastically when front-back confusions are taken into account. In fact, Table 5.2 shows that front and back can only correctly distinguished in 55 % of the cases. This is close to the 50 % obtained by random guessing. One possible explanation is that a non-personalized HRIR is used. Section 3.3.2 states that this increases the amount of front-back confusions. Another explanation is that the spectral notches in the transfer function of the headphone degrade spectral cues.

Furthermore, criticism on the experiment conducted in Section 5.3 is that test subjects were able to see immediately whether their answer is correct. This gives the subjects additional information, biasing test results. Also test subjects complained about the non-intuitive interface, causing them to occasionally select the wrong answer by accident.

6.4. HEAD TRACKER

The results in Section 5.4 show a significant improvement in the ability to resolve front-back ambiguity, satisfying requirement 1.5. However, the webcam based head tracker proved to be unreliable when the user moves the head rapidly, moves the head more than 80 ± 10° or wears glasses.

Yet, this is not the largest problem with the webcam based head tracker. The latency measured for two laptops in Section 5.4 equals 364 ms and 290 ms. This is significantly higher than the 80 ms threshold described in Section 3.3.3 and the 100 ms maximum in requirement 1.2. Therefore, we can conclude that requirement 1.2 cannot be met with a webcam based head tracker on a typical laptop. The only suitable alternative among the options listed in Section 4.7, is the inertial head tracker. However, this would compromise requirement 3.1. This shows the problem when utilizing head movements for sound localization. It is either cheap and accessible, but too slow, or it is done properly, but requires special hardware. A solution requires a compromise between requirements 3.1, 1.2 and 1.5.

A possible consequence of a lower latency is, apart from a better user experience, improved externalization. Section 5.3 states that test subjects reported internalization, i.e. that the sounds where perceived as coming from inside the head. Section 3.3.3 explains that that a good head tracker decreases the level of internalization.

6.5. RESOURCE USAGE

The measurements in Section 5.5 report that the integrated game uses 240 MB of memory and 43 % CPU on a 2.3 GHz Intel Core i5. The sound localization subsystem described in this thesis uses less resources, satisfying requirements 3.2 and 3.3.

Even though the requirements are met, concern exists about the CPU usage when the game is expanded and more complex rooms are implemented. As the shape of the room becomes more complex, the amount of reflections increase drastically [2]. As each reflection is localized, as described in Section 4.1, the CPU time used for localization increases drastically as well.

7

CONCLUSION

This thesis described the implementation of a sound localization algorithm for use in an audio-based game for visually impaired children. In this game, the user should be able to localize the sound source based on localization cues in the audio signal. An algorithm was developed to allow for the real-time playback of localized audio. Next, different localization cues were implemented. First, only binaural localization cues were implemented, but this was later expanded with spectral cues by using the head-related impulse response (HRIR). Finally, a head tracker was implemented to study the influence of head movements on localization performance and its feasibility for integration into audio-based games.

7.1. SUMMARY OF CONCLUSIONS

The main conclusions found in this thesis are:

- Real-time audio playback can be reliably implemented with the overlap and window technique described in Section 4.4. The sound did not contain any audible clicks or stuttering and the original sound can be recovered from a sequence of windowed audio blocks.
- Binaural cues (ITD and IID) work well for lateral localization. Similar performance is achieved when the HRIR is used. The HRIR can be effectively interpolated and truncated, without measurable loss in performance. The usage of HRIRs makes it theoretically possible for the user to distinguish between front and back as well as to perceive elevation. However, front-back confusions were common among test subjects. Possible causes are degradation of spectral cues by the headphones or the use of a non-personalized HRIR. The effectiveness of elevation perception was not tested.
- Front-back confusions were nearly absent when a head tracker was used. However, the latency of the head tracker is too high and the lag is noticeable by the user. The latency was mainly caused by the webcam.
- The use of headphones is a suitable option for the playback of binaural sound. However, the influence of the transfer function of the headphone has possibly led to an increase in front-back confusions. Compensation of this effect is not feasible, as each user uses different headphones.
- The effect of distance was implemented as both attenuation and propagation delay. A change in attenuation creates a noticeable change in perceived distance. Also the propagation delay works as expected. However, users experienced an increased sense of internalization. This can be caused by the lack of reflections, as the direct to reverberant sound ratio is an important cue for distance perception.
- Both the CPU usage and memory usage were less then required. Therefore, the game can be played on most computers. The game will therefore be accessible by visually impaired children. They will be able to play the game on their own computer without the need of additional hardware.
- The localization tools were used in a game for visually impaired children. The results are described in [1]. The localization subsystem worked well in conjunction with the other subsystems. Many users were able to play the game and found it enjoyable.

7.2. RECOMMENDATIONS

During this project, several potential improvements were found, but there was no more time available to investigate these improvements. We list these improvements as recommendations for future work:

- If personalized HRIRs are used, localization accuracy can be significantly improved. Because HRIRs are difficult to measure, an intermediate step would be to select a non-personalized HRIR from a large database, based on the anthropometric measurements of the user. Only a few anthropometric measurements have to be made, using less time and equipment. However, localization will not be as accurate as when the user's own HRIR is used.
- When the player moves relative to the audio source, the propagation delay changes, resulting in undesired artifacts in the resulting audio signal. This is solved by using overlapping blocks multiplied with a window function. Now a simple function is used, but another window function might have better spectral properties. An alternative option would to resample the blocks. This would require more CPU usage, but results in a physically more accurate audio signal.
- An inertial head tracker can be used to reduce the head tracker latency. It will also allow the user to fully rotate the head. This will result in a more powerful way to resolve front-back ambiguities. It might also improve the sense of externalisation.

A

ETHICAL ASPECTS

As part of the course 'Ethics and Technology', which in turn is part of the 'Bachelor Graduation project', we wrote an analysis of the ethical aspects of our product. In this section, the ethical issues are addressed, based on the assumption that the research done in this thesis will lead to a product. The product will be sold by a company, founded by the authors of this thesis.

The ethical aspects are divided into three subsections: the product, its implications and the company. Appendix A.1 describes the product that the company will distribute. Appendix A.2 describes the ethical aspects of the company policy. Appendix A.3 describes the impact on society when this product is brought on the market and its ethical implications.

A.1. THE PRODUCT

The final product would be an audio engine for games which produces auditory feedback for the player. The audio engine is based on the head-related impulse response of the user. This creates a level of realism not previously encountered in games, allowing all necessary localization cues to be included in the auditory feedback. This makes the game accessible for visually impaired people, but it can also be used to improve the level of realism in conventional visual games.

A.2. THE COMPANY

The company will be small consisting of mostly software developers. Because the company is only small, there won't be much diversity of employees. The deontological ethics say that the universal rule should be followed and an universal rule for hiring is gender equality. The hiring policy could be to hire woman, but it will be more focused on the best possible employee then on the gender of the employee. If woman are hired based on their gender and not their skills, then there is no gender equality.

The ecological footprint of the company will be minimal. The product is software based and therefore there is no production of physical goods. Only the consumption of energy for the use of computers will leave an ecological footprint. It is possible to counter this using green energy. Because the energy bill will be one of the minor expenses, this is a serious option to consider.

The company will be based in a developed country, possibly the USA or the Netherlands. The company is based in a developed country because there are the most resources available and the biggest costumers (gaming industry) are based in developed countries.

A.3. IMPLICATIONS

With the development of the audio engine, gamers will have a more realistic gaming experience. Especially visually impaired people will be able to play games, making them more equal to sighted gamers. Also, visually impaired people will be able to train their audio localization skills to withstand the challenges of the real world.

A downside of the technology is that the increased realism might increase the chance of game addiction. Addicted gamers will be more socially isolated. This will lead to negative health effects for the addicted gamers and it will negatively impact society as a whole.

However, casual gamers will have a game which they can enjoy. Visually impaired people will be able to play games with this technique, both for leisure and for training purposes. This will lead to better integration of visually impaired people in society. From the viewpoint of consequentialism, we can say that the happiness of the casual and visually impaired gamers outweigh the misery of the few.

GLOSSARY

Anthropometry The measurement of a person's physical dimensions.

API Application programming interface.

Auditory event The sound as perceived by the listener.

Binaural Involving two ears.

HRIR Head related impulse response, describing the filter effects of the pinnae, head and torso. Fourier transform of the HRTF.

HRTF Head related transfer function, describing the filter effects of the pinnae, head and torso. Fourier transform of the HRIR.

IID Interaural intensity difference.

Interaural Between the two ears.

ITD Interaural time difference.

MAA Minimum audible angle, the minimum angle between two sound sources for which the auditory events are distinguishable.

Midsagittal plane The plane perpendicular to the line between the two ears.

Monaural Involving one ear.

Player The virtual entity controlled by the user.

Reverberation The collection of reflected sounds audible after the direct sound has ended.

User The individual who is playing the game.

BIBLIOGRAPHY

[1] T. Coppoolse and W. van Dam, *Audio-based game for visually impaired children,* (2015).

[2] M. Bisschop and J. Noortman, *Sound reflections in an audio-based game for visually impaired children,* (2015).

[3] P. Tipler and G. Mosca, *Physics for Scientists and Engineers*, sixth edition ed. (Freeman, 2008).

[4] D. M. Howard and J. A. Angus, *Acoustics and Psychoacoustics* (Focal Press, 1996).

[5] W. A. Yost and D. W. Nielsen, *Fundamentals of Hearing An Introduction* (Holt, Rinehart and Winston, 1977).

[6] W. M. Hartmann, *Signals, sound, and sensation* (Amer Inst of Physics, 1997).

[7] K. F. Herzfeld and F. O. Rice, *Dispersion and absorption of high frequency sound waves,* Phys. Rev. **31**, 691 (1928).

[8] H. E. Bass and F. D. Shields, *Absorption of sound in air: High frequency measurements,* The Journal of the Acoustical Society of America **62** (1977).

[9] H. O. Kneser, *The interpretation of the anomalous sound absorption in air and oxygen in terms of molecular collisions,* The Journal of the Acoustical Society of America **5**, 122 (1933).

[10] J. Middlebrooks and D. Green, *Sound localization by human listeners,* Annual Review of Psychology **42**, 135 (1991).

[11] F. L. Wightman and D. J. Kistler, *The dominant role of lowfrequency interaural time differences in sound localization,* The Journal of the Acoustical Society of America **91** (1992).

[12] E. A. Macpherson and J. C. Middlebrooks, *Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited,* The Journal of the Acoustical Society of America **111** (2002).

[13] D. McFadden and E. G. Pasanen, *Lateralization at high frequencies based on interaural time differences,* The Journal of the Acoustical Society of America **59** (1976).

[14] W. M. Hartmann, *How we localize sound,* Physics today **52**, 24 (2008).

[15] J. Blauert, *Spatial Hearing* (The MIT Press, Cambridge, Massachusetts, 1983).

[16] R. A. Butler and K. Belendiuk, *Spectral cues utilized in the localization of sound in the median sagittal plane,* The Journal of the Acoustical Society of America **61** (1977).

[17] J. Hebrank and D. Wright, *Spectral cues used in the localization of sound sources on the median plane,* Journal of the Acoustical Society of America **56**, 1829 (1974).

[18] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, *Localization using nonindividualized head related transfer functions,* The Journal of the Acoustical Society of America **94**, 111 (1993).

[19] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, *Approximating the head-related transfer function using simple geometric models of the head and torso,* The Journal of the Acoustical Society of America **112** (2002).

[20] T. Carpentier, H. Bahu, M. Noistering, and O. Warusfel, *Measurement of a head-related transfer function database with high spatial resolution,* Tech. Rep. (forum acusticum, Paris, France, 2014).

[21] B. W.O., B. A.W., and A. M.A., *The contribution of head movement to the externalization and internalization of sounds,* (2013), 10.1371/journal.pone.0083068.

[22] D. S. Brungart, B. D. Simpson, and A. J. Kordik, *The Detectability of Headtracker Latency in Virtual Audio Displays, Eleventh Meeting of the International Conference on Auditory Display*, (2005).

[23] A. W. Mills, *On the minimum audible angle,* The Journal of the Acoustical Society of America **30** (1958).

[24] D. R. Perrott and K. Saberi, *Minimum audible angle thresholds for sources varying in both elevation and azimuth,* Journal of the Acoustical Society of America **87**, 1728 (1990).

[25] H. Wallach, E. B. Newman, and M. R. Rosenzweig, *A precedence effect in sound localization,* The Journal of the Acoustical Society of America **21** (1949).

[26] P. M. Zurek, *The precedence effect and its possible role in the avoidance of interaural ambiguities,* The Journal of the Acoustical Society of America **67** (1980).

[27] W. M. Hartmann, *Localization of sound in rooms,* The Journal of the Acoustical Society of America **74**, 1380 (1983).

[28] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, *The precedence effect,* The Journal of the Acoustical Society of America **106** (1999).

[29] B. Shinn-Cunningham, *Distance cues for virtual auditory space, Proceedings of the IEEE-PCM*, **2000**, 227 (2000).

[30] P. Zahorik, *Auditory display of sound source distance,* in *Proc. Int. Conf. on Auditory Display* (2002) pp. 326–332.

[31] D. Mershon and L. King, *Intensity and reverberation as factors in the auditory perception of egocentric distance,* Perception & Psychophysics **18**, 409 (1975).

[32] J. P. Maxfield, *Some physical factors affecting the illusion in sound motion pictures,* The Journal of the Acoustical Society of America **3** (1931).

[33] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, *Auditory distance perception in humans: A summary of past and present research,* Acta Acustica united with Acustica **91**, 409 (2005).

[34] Z. Cattaneo, T. Vecchi, C. Cornoldi, I. Mammarella, D. Bonino, E. Ricciardi, and P. Pietrini, *Imagery and spatial processes in blindness and visual impairment,* Neuroscience & Biobehavioral Reviews **32**, 1346 (2008).

[35] N. Lessard, M. Paré, F. Lepore, and M. Lassonde, *Early-blind human subjects localize sound sources better than sighted subjects,* Nature **395**, 278 (1998).

[36] F. Gougoux, F. Lepore, M. Lassonde, P. Voss, R. Zatorre, and P. Belin, *Pitch discrimination in the early blind,* Nature **430**, 309 (2004).

[37] A. Aleman, L. Van Lee, M. Mantione, I. Verkoijen, and E. De Haan, *Visual imagery without visual experience: Evidence from congenitally totally blind people,* NeuroReport **12**, 2601 (2001).

[38] M. Zwiers, A. Van Opstal, and J. Cruysberg, *Two-dimensional sound-localization behavior of early-blind humans,* Experimental Brain Research **140**, 206 (2001).

[39] M. Noordzij, S. Zuidhoek, and A. Postma, *The influence of visual experience on the ability to form spatial mental models based on route and survey descriptions,* Cognition **100**, 321 (2006).

[40] J. Rieser, J. Lockman, and H. Pick, *The role of visual experience in knowledge of spatial layout,* Perception & Psychophysics **28**, 185 (1980).

[41] *Pyaudio: Portaudio v19 python bindings,* `https://people.csail.mit.edu/hubert/pyaudio/`, accessed: 12-06-2015.

[42] *Matlab - the language of technical computing,* `http://nl.mathworks.com/products/matlab/`, accessed: 12-06-2015.

[43] *Python,* `https://www.python.org/`, accessed: 12-06-2015.

[44] *Scipy,* http://www.scipy.org/, accessed: 12-06-2015.

[45] *Pygame,* http://www.pygame.org/, accessed: 12-06-2015.

[46] E. Choueiri, *Optimal crosstalk cancellation for binaural audio with two loudspeakers,* Princeton University , 28 (2008).

[47] F. Wightman and D. Kistler, *Measurement and validation of human hrtfs for use in hearing research,* Acta Acustica united with Acustica **91**, 429 (2005-05-01).

[48] B. C. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, 2003).

[49] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, *The cipic hrtf database,* Workshop on Applications of Signal Processing to Audio and Electroacoustics , 99 (2001).

[50] J. Sodnik, R. Susnik, M. Stular, and S. Tomazic, *Spatial sound resolution of an interpolated hrir library,* Applied Acoustics **66**, 1219 (2005).

[51] T. Nishino, S. Mase, S. Kajita, K. Takeda, and F. Itakura, *Interpolating hrtf for auditory virtual reality,* The Journal of the Acoustical Society of America **100**, 2602 (1996).

[52] T. V. Sreenivas, V. C. Raykar, and R. Raman, *Head related impulse response interpolation for dynamic spatialization,* Texas Instruments DSPS fest-2 k, Banglore, India (2000).

[53] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. Davis, *Hrtf personalization using anthropometric measurements,* in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* (Ieee, 2003) pp. 157–160.

[54] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, *Head related transfer function personalization based on multiple regression analysis,* in *Computational Intelligence and Security, 2006 International Conference on*, Vol. 2 (IEEE, 2006) pp. 1829–1832.

[55] *Trackir premium head tracking for gaming,* http://www.naturalpoint.com/trackir/, accessed: 12-06-2015.

[56] *Freetrack optical head tracking software,* http://www.free-track.net/english/, accessed: 12-06-2015.

[57] *Opentrack head tracking software for ms windows, linux, and apple osx,* https://github.com/opentrack/opentrack, accessed: 12-06-2015.

[58] *Edtracker,* https://edtracker.org.uk/index.php, accessed: 12-06-2015.

[59] *Headsoft vjoy virtual joystick driver,* http://www.headsoft.com.au/index.php?category=vjoy, accessed: 12-06-2015.