University of Glasgow | School of Computing Science

Audio Augmented Reality: Real-Time Sound Replacement

Zhenyuan Shen

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

12/09/2024

Abstract

As augmented display technology develops, audio augmented reality (AAR) technology is also developing, which enriches the user experience by superimposing virtual sound layers on the real-world auditory environment so that users cannot distinguish between virtual and real sounds. This article studies the application of real-time sound replacement in AAR, taking everyday sounds (such as human voices) and converting them into other interesting sounds (such as dog barking or thunder). We use audio processing technology and deep learning algorithms (YamNet) to achieve efficient and real-time sound replacement. The system is able to replace sounds quickly and accuratel. This is done using the YAMNet model to ensure an immersive user experience. Therefore, this study shows that AAR technology has potential application value in entertainment, education, games or artistic creation. Through these findings, we aim to inspire new ideas and application avenues for AAR and promote its innovation and application in different fields.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: _____ Signature: _____

Acknowledgements

Firstly, I would like to thank my supervisor Stephen Brewster for his valuable guidance and unwavering support throughout the research process. I am also grateful to Jake Bhattacharyya for his significant assistance in designing and building the Unity models, and for exploring and discussing research questions with me. Lastly, I want to thank my family and friends for their understanding and support, which allowed me to concentrate on my academic pursuits.

Contents

Chapter 1: Introduction

Audio technology is rapidly being developing, allowing us to integrate virtual auditory layers in real-world environments. Following the development of virtual reality (VR) and augmented reality (AR), audio augmented reality (AAR) provides a new experience of interaction in which users are augmented with a virtual sound layer over the physical real-world environment in order to enhance their everyday lives. This acoustic transparency feature gives a user significant advantage over that provided by conventional headphones, as it not only enhances user safety but also introduces complex privacy considerations in maintaining environmental awareness. This, of course, adds to the safety of street users but also raises challenges toward privacy protection,which allows users to maintain environmental awareness,especially in urban settings where auditory cues are essential for navigation and interaction with surroundings[8]. However, although AAR technology has developed rapidly, many problems of real-time sound replacement still exist, such as the real-time performance, precision, and user experience, for instance, how to replace and perceive the direction of a virtual sound in a moment without the users' noticing.[4].

This research explores real-time sound replacement technology in audio augmented reality. It will use deep learning models (YAMNet, https://tfhub.dev/google/yamnet/1) and audio processing technology (binaural microphones for direction judgment and 3D sound effect playback) to classify target sounds and realize sound sources simultaneously.These microphones can effectively utilize head-related transfer functions (HRTFs) to simulate how sound is perceived from different directions, thus locating and converting certain everyday sounds into other interesting soundsrapidly [28][12](such as converting a human voice into a dog barking). Moreover, with the integration of deep learning models, this system enables real-time sound categorization and modification, hence, enabling substitution of everyday sounds with pre-defined auditory outputs quickly [13].In addition, we will verify the effectiveness of its model in Unity through experiments and discover the advantages or disadvantages of our experiments through different experimental comparisons and questionnaires, such as whether the sound conversion and direction are real-time. This study employs YAMNet, a deep learning model, alongside innovative audio processing tools such as binaural microphones, to navigate these challenges effectively. Recent work indicates that the current research and usability focus on optimization methodologies is necessary to ensure user experience-smooth and precise execution of sound conversions [10][1]. In this way, we will consider a series of optimization methods for these problems to meet users' needs as much as possible. Through rigorous experimental methods, this research aims to enhance user experience by swiftly converting everyday sounds into predefined auditory outputs, thus testing the real-time capabilities and user interaction dynamics of our AAR system[30].

Hence, in this paper we will introduce the survey of the AAR technology in the next chapter, stating the development of AAR today; Then we will introduce the implementation of the project including the code design and challenges in the third chapter; finally in the last chapter, we will discuss the test of the project and evaluation.

Chapter 2: Survey

2.1 Definition and Origin of AAR Technology

Augmented Reality (AR)[6] is a computerized system that superimposes virtual information based on computers onto the real world in a dynamic manner. It is capable of realizing the seamless integration of the natural world with the computer-created virtual world, hence the user cannot discern between it and real perception. The technology simulates human sensory perception for instance by vision, hearing, smell, and touch so that the sensory information may interact dynamically in space. Azuma summarized three well-accepted characteristics of AR in 1997 as follows[36]:

- Fusion of real and virtual environments
- Real-time interaction
- Three-dimensional Registration Support

But the AR that we will discuss is largely based around audio augmented reality (AAR). In the case of AAR, it is a kind of technology where the virtual audio features are added to the user's real-world audio, not much different from visual AR. AAR enhances auditory perception through amplification or attenuation of natural sounds or the addition of totally non-existing virtual sounds to enrich the auditory experience of human beings,for example, AAR application can zoom significant sound and weaken background noise to improve communication during the noisy environment[18]. The technology of AAR itself has its roots in the early 2000s, when it was primarily used for advanced communication systems in the military and aviation fields. However, since the increased popularity of smartphones and portable devices, AAR technology began to be commercialized and gradually came closer to people's lives, especially in entertainment and consumer electronics[39].

2.2 Current developments in AAR

2.2.1 Current Situation

In audio augmented reality (AAR), virtual audio content is integrated with the natural world to enhance the user's actual real-world environment and simulate the auditory experience of the real world.

With the emergence and improvement of 3D sound effects and recording technology, the experience of audio augmented reality will become more extensive. AAR technology adds virtual audio to real-world auditory environments to enhance the sense of immersion in the natural environment or make people feel sci-fi (sounds that should not be there appear). For example, work on mobile AAR systems emphasizes the design of auditory displays that provide location-based information[34].At the same time, technological advances in AAR have led to the development of sophisticated audio systems that utilize binaural recording techniques and adaptive filtering to create realistic listening environments[26]. This technology has been widely used in various fields, such as teleconferencing, barrier-free audio

systems, location-based audio games, pure audio games, collaboration tools, education and entertainment[37].

2.2.2 Acoustic Transparency

Compared with traditional headphones, a significant advantage of AAR technology is that it does not create barriers between personal and public listening space[17] (it can perfectly embed virtual audio in real life so that users are not aware of it). The emergence of some acoustically transparent headphones has connected AAR technology with acoustic transparency. These acoustically transparent headphones enable users to perceive the surrounding environment as if they were not wearing headphones, making it difficult to distinguish which sounds are natural and which are produced by the technology. Instead of widening the gap between people and reality like traditional headphones, people are given a sense of being outside the world[37].

However, technical issues are a major challenge for acoustic transparency in this field, and user privacy and security have become an essential part of the machine. That is, whether acoustic transparency will reveal personal information[21]. Research shows that users are very concerned about whether acoustically transparent headphones will leak their personal privacy and safety issues when wearing them. Generally speaking, experimenters will be more worried about privacy issues and believe in acoustic transparency, which can naturally perceive the environment and make it safer[21].

2.2.3 Applications of AAR Spatial Audio Technology

Existing research involves aspects such as sound quality perception, the ability to distinguish between real and virtual sounds, the creation of a sense of reality, and user adaptability to AAR wearable devices.

The application of AAR spatial audio and format technology in immersive theater provides new ways to create fictional spaces and parallel realities, which can engage audiences in a variety of ways and has the potential to promote more interaction, participation, and collaboration. AAR spatial audio technology enhanced sound quality perceived by users, in that they could effectively distinguish between adistinguish real and virtual sounds. For the listener to effectively discriminate between a real sound emitter from a virtual one, there is the so-called HRTFs enabled to do so. Some studies demonstrated that the greater the number of characteristics in the HRTF smoothing, the more this discrimination will be powerful[19].

Headphone theatre in immersive theatre relies on binaural sound technology, which enhances the audience's sense of immersion. However, many headphone theatre performances have poor immersion for viewers due to the isolation of audiences from each other and the failure to identify physical headphones as part of the performance[29]. Performances with headphones or can maximize the sound experience, but immersion is also an important factor. To overcome these limitations, some theatre groups have begun to use binaural recording devices in live performances.

For example, Anna by the National Theatre of the United Kingdom and Encounter by Complicité provide audiences with a more intimate and direct performance experience through headphones[27]. These works combine augmented sound with real-life interaction to create a new type of theatre experience that fully utilizes the advantages of AAR technology while retaining the interactivity and commonality of live performances.

In addition, Microsoft's HoloLens 2 is a mixed reality headset that captures real-world environmental information through advanced sensors and cameras, and then overlays high-resolution holographic images within the user's field of view. HoloLens 2 is equipped with spatial audio technology that can create 3D sound effects so that the sound seems to come from the actual spatial location of the virtual object[33]. This device can be used in architectural design. Designers can hear sounds from different directions through HoloLens 2, enhance their perception of spatial layout, and thus improve design efficiency and experience[38].

2.3 Concerns and Challenges of AAR

Despite the innovative progress of audio augmented reality (AAR) technology, there are still many problems. One of the important issues is user privacy and safety. As mentioned in the discussion of acoustic transparency, because AAR technology usually relies on continuous audio collection from the environment and outputs virtual audio, it may leak sensitive personal information without the user's explicit consent or without the user's awareness. In addition, the boundary between real sound and virtual sound is sometimes blurred under the influence of AAR technology, which may cause confusion or misinterpretation of auditory cues in critical situations. This may lead to some dangerous sounds being ignored as virtual sounds. For example, if a tiger is behind the user and roars, the user who is accustomed to the virtual sound may not perceive the danger. These challenges highlight the need to clearly stipulate user-centred design principles when developing AAR technology to protect user welfare and enhance the auditory experience without compromising user privacy and personal safety[37].

In addition, developing more affordable and accessible AAR technology products and promoting the application of AAR technology through policy support and technical training will be key factors in promoting the widespread application and development of AAR.For example, AAR applications can be effectively deployed on commonly used devices such as smartphones to promote wider access[5]. Continuous technical support and updates are essential to ensure the long-term effective operation of AAR technology, especially in a rapidly evolving technology environment, where timely updates and maintenance of the system are essential. Not only can this improve the user experience, but the security and stability of the continuous updates and iterations will be more perfect to ensure a positive user experience[37].

Chapter 3: Unity Scene Design and Implementation

3.1 Audio Event Classification

3.1.1 Define a Sound to be Captured

When it came to what sound to use as a trigger for our sound swap, "Speech" stood out as our final choice. The basic process of swapping sounds in theory is to detect whether the target sound is active in the environment, estimate the location of the sound, remove it in some way, and then play the replacement sound (dog barking) at the same location. Positioning can be the most difficult part, especially for objects in relative motion, so at first, we aimed for a relatively static or predictable sound as our target sound.

We considered many types of sounds, and the first idea was whether we could swap birdsong. Birdsong usually has an initial position that is not particularly clear but at the same time its position is relatively fixed (think of a bird perched on a tree and chirping). This seemed like a good choice, but when we tested it, problems arose. It is difficult for us to capture the calls of birds in nature, so we can only rely on mobile phones to play the chirping of birds. Another point worth noting is that when we actually use the YAMNet model to classify the sound while turning on the bird calls on the mobile phone, the model cannot quickly identify the bird calls. Even after a long period of recognition, it occasionally flashes when the classification type is bird calls. Its confidence is only about 0.2, which undoubtedly means that for our scenario, it is not a wise choice to use bird calls as the target sound.

Another option is the sound of cars or car engines. When we consider it from the perspective of pedestrians on the road, car sounds as target sounds seem to be a good choice. Because choosing car sounds is equivalent to providing pedestrians with a safety alert when they recognize car sounds, which is perfect for the future application of the model. However, if you want to do this from the perspective of a bystander, you may have to use a camera for object detection instead of simple sound replacement. When we look at it from the perspective of a car passenger or driver, we can more easily predict the position of the engine relative to the user, and it is a more controlled working environment (for example, less background noise). However, this means that the sound source is always fixed, because the engine is always in a relatively fixed position.

Or a bicycle bell. From the perspective of a cyclist, the bell always comes from in front of you, which is similar to the engine and the passenger; from the perspective of a pedestrian, the sound source is too random, because the bicycle can appear from anywhere and move in a short time (the bicycle may have passed the pedestrians during a bell), which will cause the microphone to collect the sound in a short time and cannot determine the correct position of the sound source.

Therefore, we will use the sound type of Speech as our target sound. First, we found that in the Unity scene, Speech is easy to identify as a sound type, that is, the YAMNet model can better judge the sound type of Speech compared to other sounds, and the confidence value of the sound can also be at a higher value (0.5-0.9). Secondly, the sound of Speech is also

relatively fixed and has obvious position perception (when two people are talking, we can usually put the microphone in the middle to feel the left and right direction of the virtual sound). In addition, it appears as a Speech type rather than a Human Voice type, which means that the Speech type of sound is usually a long-term source (that is, the Long-Term part in the scene can also become a Speech type). This long-term and relatively fixed sound source greatly facilitates the microphone to collect and determine the direction of the sound source and reduce errors.

3.1.2 Defination and Design of YAMNet Model

Sound event classification is very important in building up a full Unity scene. In the constructed Unity scene, we should make sure that the sound which triggers the dog's barking has to be of human speech; hence, we need to identify the speech we need in a real outdoor environment. However, many challenges are seen in the classification of sound events. First, different sounds have different acoustic attributes. Some sounds are of short durations, for example, a gunshot; some are very long, for example, the sound of speaking; and some have different frequencies, like music. Moreover, in the sound event detection process, if the target sound to be detected is far away from the microphone, the priority of the target sound received by the microphone is low compared to other sounds in the environment, making it difficult to detect. Finally, in real life, sound events are often mixed with several types of sounds, making several sound events occur simultaneously. This also makes detection very difficult.

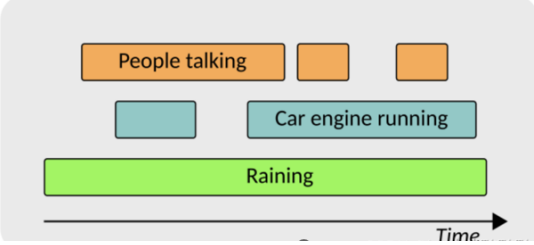

Figure 3.1: Sound Event Detection

Owing to the aforementioned issues, we will classify sound events using the YAMNet model. YAMNet is an audio event classification model based on convolutional neural network (CNN), used for the identification and classification of audio events. Lightweight based on MobileNetV1 architecture, the model is fit for resource-constrained environments: mobile devices and embedded devices. The specific operations for audio classification are:

1.Audio pre-processing The input audio data will be pre-processed, wherein input is taken in fixed-sized audio chunks of a pre-determined length. It is then converted into a spectrogram, giving a visual representation of the frequency of an audio signal against time. Spectrograms are the input of the YAMNet model.

2.Feature extraction: The preprocessed spectrogram is passed to the YAMNet convolutional neural network, which extracts high-level features from the spectrogram that go on to represent time and frequency patterns of the audio signal. Subsequent layers of convolution and pooling operations in the YAMNet convolutional neural network will extract high-level features from the spectrogram.

3.Audio event classification The features are then extracted and fed into a fully connected layer, which finally outputs a probability distribution using the Softmax function, based on the possibility that the audio segment belongs to the different categories. YAMNet is a

pretrained model using Google's AudioSet Ontology and is able to recognize 521 different audio events.

4.Classification result output: The model assigns a label to an audio clip from the output probability distribution, and it gives a confidence level in between 0 to 1 to represent how much the audio is similar to that label. The closer to 1, the more similar it is, and vice versa; thus, it indicates what is the most probable kind of audio event that corresponds to an audio clip.

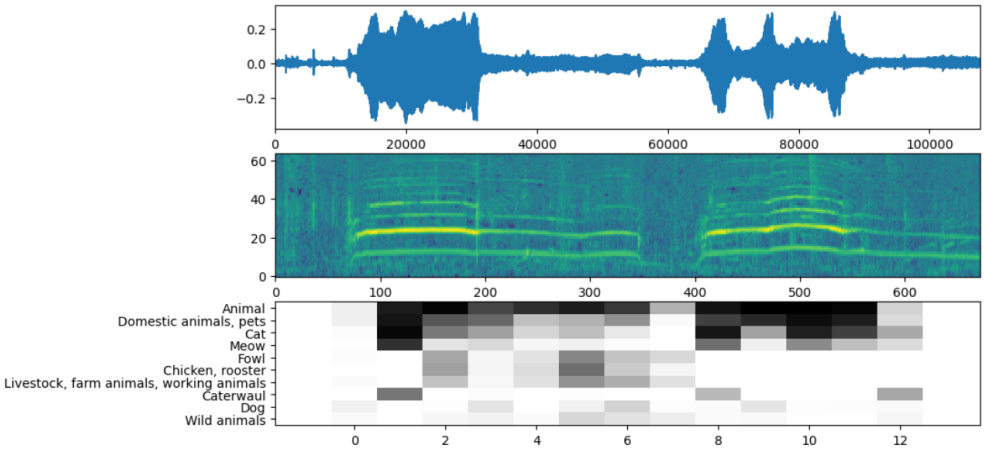

Figure 3.2: Information for the Visualization by YAMNet (waveforms, spectrograms, and inference)

3.2 Change Real Sound into Virtual

3.2.1 Locate Sound Source

Regarding the location and detection of sound events, we studied the paper by Archontis Politis, who brought forth the fact that convolutional neural networks can be used to locate and detect overlapping sound sources. However, based on our constructed scenario, we found that a simpler way can also be used to locate the sound.

We found that the microphones were stereo and what we needed to do was perform a simple positioning instead of an accurate coordinate in 3D space. For such a relatively simple positioning task, one idea is to run the YAMNet model on each channel of the microphone and make some kind of calculation out of it. For example, if the speech confidence in the left channel is 0.75 and in the right channel it is 0.25, then probably the speech is closer to the microphone on its left side than to the right side.

However, when we actually operated, we found that the volume difference between the left and right channels can also be used to estimate the direction of the sound source. This project meets our needs to position without stress. The basic idea is to make a judgment on the direction from which the sound is coming by the relative volume value in the left and right channels of the stereo audio data captured from the microphone or other audio acquisition device.

1.In our specific steps: First, collect the sound data of the left and right channel of the microphone. The data point of each channel represents the volume (amplitude) of the channel

at a certain moment.

```
void StartMicrophone()
{
    audioSource.clip = Microphone.Start(null, true, 10, 44100);
    audioSource.loop = true;
    while (!(Microphone.GetPosition(null) > 0)) { }
    audioSource.Play();
}
```

Figure 3.3: Microphone Recording

```
StartMicrophone();

//Screen.sleepTimeout = SleepTimeout.NeverSleep; // Never sleep so that logging device stays awake while running.
model = ModelLoader.Load(modelFile);
worker = WorkerFactory.CreateWorker(model, WorkerFactory.Device.GPU);
ClassMap();
if (inputClip != null)
{
    PredictAudioFile(inputClip);
}
StartCoroutine(LogTimer()); // Start log timer
```

Figure 3.4: Start Microphone Recording

2.Volume calculation: In the code, the sample data in the current audio clip is obtained through audioSource.clip.GetData(samples, position). This data is an alternating array containing left and right channels. We iterate through the array, and for absolute volume values of the left and right channels, respectively, we can maintain two variables: 'leftSum' and 'rightSum', which store the total volume for the left and right channels, respectively.

```
//Create a method to analyze audio data and estimate sound source direction
//compare the energy of the left and right channels to determine which direction the sound is more likely to come from.
void Update()
{
    if (audioSource.clip == null || Microphone.GetPosition(null) <= 0)
    {
        //Debug.LogWarning("Microphone is not ready or no data available.");
        return;
    }

    float direction = calculateDirection(); // 获取计算的方向

    // 检查方向变化是否显著
    if (Mathf.Abs(direction - lastDirection) > directionThreshold)
    {
        UpdateSoundDirection(direction);
        lastDirection = direction;
    }
    // 检查模型的预测结果并播放狗叫声
    //CheckAndPlayDogBarkingSound(direction);
}
```

Figure 3.5: volume Calculation

3.Direction estimation: The left and right channel volumes will then be subtracted to arrive at the difference, that is, 'rightSum - leftSum'. The difference obtained will signify the direction of sound. If 'rightSum' is greater than 'leftSum', it means that the right channel volume is greater; then the sound is inferred to come from the right. On the other hand, if 'leftSum' is larger than 'rightSum', it is taken as an inference that the sound is from the left.

This way we can easily say where the sound source is: left or right. That makes it intuitive. Anti-harvesting is great for the sound source estimation in this plane to simulate the real experimental environment.

```
private float calculateDirection()
{

    Debug.Log("calculateDirection method called.");

    if (audioSource.clip == null)
    {
        Debug.LogWarning("audioSource.clip is null.");
        return 0.0f; // 如果没有音频片段，无法计算方向
    }
    int sampleSize = 1024;  // 样本窗口的大小 sample size
    float[] samples = new float[sampleSize * 2];  // 立体声通道 Dual Channel

    int position = Microphone.GetPosition(null) - sampleSize;
    Debug.Log($"Microphone position: {position}");
    if (position < 0)
    {
        Debug.LogWarning("Microphone position is less than sample size.");
        return 0.0f;  // 确保不会使用负索引 Make sure don't use negative indexes
    }
    audioSource.clip.GetData(samples, position);

    float leftSum = 0, rightSum = 0;
    for (int i = 0; i < sampleSize; i++)
    {
        leftSum += Mathf.Abs(samples[2 * i]);
        rightSum += Mathf.Abs(samples[2 * i + 1]);
    }
    Debug.Log($"LeftSum: {leftSum}, RightSum: {rightSum}, Difference: {rightSum - leftSum}");

    return rightSum - leftSum;  // 右边为正，左边为负    Right side is positive, left side is negative
}
```

Figure 3.6: Direction Estimation

3.2.2 Add 3D Sound Effects to virtual Audio

In Unity, in order to add 3D sound effects to the virtual sound of a dog barking, we need to use some kind of spatial audio engine to spatialize the sound and have the user wear headphones. Usually, you can use the 3DTuneIn toolkit, or you can use Google Resonance, Steam Audio or Oculus/Meta's solution.

```
private void UpdateSoundDirection(float direction)
{
    Debug.Log($"Calculated Direction: {direction}");
    // 确定音源的x坐标，10代表向右，-10代表向左
    //Determine the x-coordinate of the sound source, 10 represents right, -10 represents left
    float xPosition = direction > 0 ? 10 : -10;

    // 根据方向设置音源的具体位置
    //Set the specific location of the sound source according to the direction
    Vector3 position = new Vector3(xPosition, soundSourceTransform.position.y, soundSourceTransform.position.z);

    // 应用位置变更
    //Apply location changes
    soundSourceTransform.position = position;
    Debug.Log($"Direction: {direction}, New Position: {position}");
}
```

Figure 3.7: Define Virtual Location

The way we use it is to use the 3D sound effect settings of the AudioSource component. The implementation of 3D sound effects depends on the relative position between the sound source and the listener, as well as the spatial sound effect settings of the AudioSource component. To implement its specific steps, we first need to make sure that the AudioSource component is set to 3D sound effects, that is, set the "spatialBlend" value to 1.0. Then adjust the effect of the 3D sound effect by configuring the parameters of "Volume Rolloff" (determines how the sound decays with distance) and "Min Distance and Max Distance" (the

distance range where the sound starts and completely decays). Then, the ”UpdateSoundDirection” function is used to update the dog barking sound at a specific position in 3D space. Finally, the ”PlayDogBarkingSound” method uses ”AudioSource.PlayClipAtPoint” to play the dog barking sound at a specific position. In this way, the user can feel the dog barking sound coming from the direction of the sound source ”Speech” in the headphones.

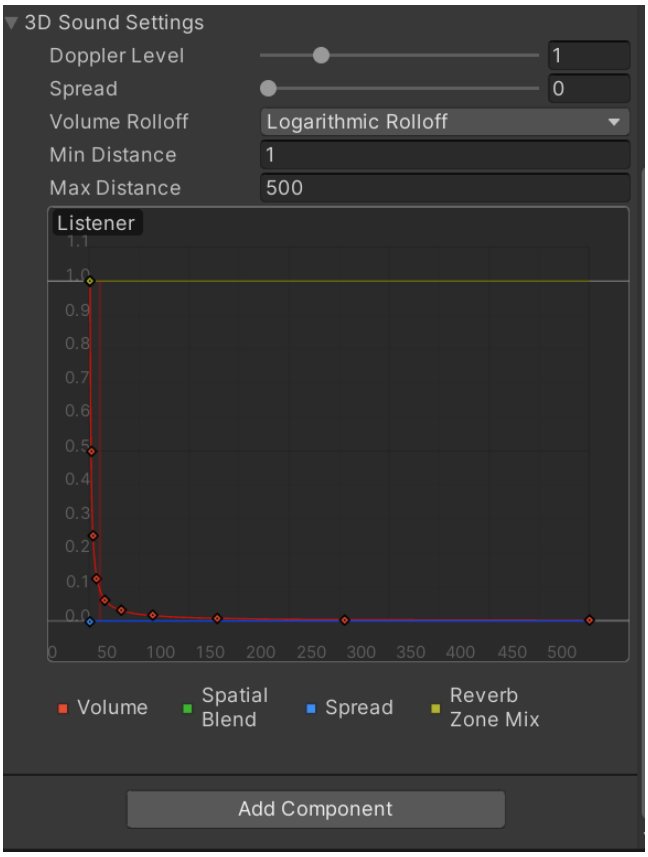

Figure 3.8: 3D Effects Settings

3.3 Start and Stop virtual Audio

Generally speaking, starting and ending the virtual sound is a relatively simple part. We only need to add a condition when playing and ending. As for the selection of this condition, we have selected the confidence of the sound type ”Speech” through constant comparison. When the confidence of ”Speech” is greater than or equal to 0.5, we will start to quickly calculate the direction of the sound source at this time and pass it to the ”UpdateSoundDirection” function, and then use this function to play the dog barking sound in a specific 3D space; and when the confidence of ”Speech” is less than 0.5, the system will stop it through the function ”StopDogBarkingSound”.

3.4 Issues in Building Scene

3.4.1 Error Location in Sound Source Detection

When we are doing sound localization, errors are always inevitable, and some of them are typical cases and worth analyzing.

Unlike the final code, the initial sound source localization code takes into account the running speed of the system. The positioning idea is: when the confidence of ”Speech” is greater than or equal to 0.75, the microphone starts recording the sound, and the 3D sound effect of the dog barking is played by identifying the direction of the sound source at this moment. However, the result of this method is that the dog barking in the headset is always on the left side of the user, regardless of whether the sound source is on the left or right. By continuously outputting and viewing the debug log, we found that the microphone position is always ”-1024” and has never changed, which makes the calculated direction always 0 and the new direction defaults to (-10.00, 0.00, 0.00). Through searching, the possible reason for this problem is that when the system obtains audio data from the microphone, the sample position attempted to be obtained may be before the start position of the audio buffer, which then leads to a negative index.

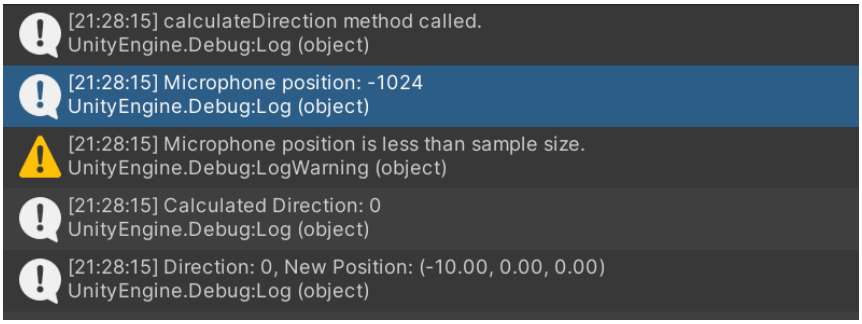

Figure 3.9: Caption

Simply put, this is probably because the moment when the confidence of ”Speech” was recognized to be greater than or equal to 0.5 was too short, causing the microphone to not record enough audio data and thus unable to determine the left and right of the sound source. Therefore, we can change our thinking. Instead of judging the direction of the sound only when the confidence of ”Speech” is greater than or equal to 0.5, the system will collect audio data through the microphone at all times from the moment the program is opened, and when the confidence of ”Speech” is greater than or equal to 0.5, the direction of the sound source at that time is calculated and passed to the dog barking to make it play in the correct position. Obviously, this method is successful, and the dog barking finally successfully appears in the correct direction of the sound source. Although the dog barking occasionally appears in the opposite direction due to some interference caused by noise or echo, this is an acceptable error and most of the time, the dog barking appears in the correct position.

3.4.2 Error in Clip Stopping

The problem of this barking dog in need of silencing, when the speech sound is not recognized, actually spread throughout the project. In version 1.0 of the project, the project was only capable of recognizing the type of sound and converting speech into dog barking with no 3D sound effects and directions from the sound source. Our project is based on version 1.5 and is already implemented to stop barking with a dog. Although we developed the function ”StopDogBarkingSound()”, in fact, it did not pause the dog barking properly; however, this did not affect program operation, and we put it on the shelf just for the moment, waiting till the end to fix this function.

By constantly analyzing the code and looking for materials, we found out that we were using the PlayClipAtPoint() method in the PlayDogBarkingSound() and StopDogBarkingSound() methods; according to the documentation (https://docs.unity3d.com/ScriptReference/Audio

-Source.PlayClipAtPoint.html), this process creates a new audio source. Although the StopDogBarkingSound() function will stop only the source of audioSource, not the one that was created with PlayClipAtPoint, so the dog will just continue barking.

```
private void PlayDogBarkingSound()
{
    if (dogBarkingClip != null)
    {
        //AudioSource.PlayClipAtPoint(dogBarkingClip, Camera.main.transform.position); // 在主摄像机位置播放狗叫声
        // 使用3D音频播放
        //AudioSource.PlayClipAtPoint(dogBarkingClip, soundSourceTransform.position);
        // 如果当前没有播放的狗叫声，则创建一个新的 AudioSource 来播放
        //If there is no dog barking sound currently playing, create a new AudioSource to play it.
        if (barkingAudioSource == null)
        {
            barkingAudioSource = gameObject.AddComponent<AudioSource>();
            barkingAudioSource.spatialBlend = 1.0f;  // 3D音效
            barkingAudioSource.rolloffMode = AudioRolloffMode.Logarithmic;  // 对数衰减
            barkingAudioSource.minDistance = 1;
            barkingAudioSource.maxDistance = 50;
        }

        barkingAudioSource.clip = dogBarkingClip;
        barkingAudioSource.transform.position = soundSourceTransform.position;
        barkingAudioSource.Play();
    }
}

private void StopDogBarkingSound()
{
    //if (audioSource.isPlaying)
    //{
    //    Debug.Log("Stopping dog barking sound.");
    //    audioSource.Stop();
    //}
    if (barkingAudioSource != null && barkingAudioSource.isPlaying)
    {
        barkingAudioSource.Stop();
    }
}
```

Figure 3.10: How to Stop Dog Barking

Chapter 4: Evaluation

4.1 Evaluation Overview

4.1.1 Test targets and metrics

The present test is aimed at verifying that the Audio Augmented Reality project software can correctly identify and convert real sound into virtual sound and play it back toward the type of sound direction in order to create a fake effect. Whether the system crashes and user experience are also important test goals, besides these three required test goals.

The number of system crashes, response time of the system, and user-experience satisfaction are all needed to be able to measure the different precisions of the quantitative indicators of success. Furthermore, to assure the benefits of the testers enrolled in the experiment, including privacy and safety, it is important to have them co-sign a consent form as participants with the experimenter, make sure they are all over 18 years of age, and are voluntary participants in this test, co-creating and filling out the questionnaire on related questions, and participating in possible interviews after the test.Moreover, the correlation between auditory performance and user satisfaction has been recorded, and it has been recorded that good auditory performance is related to effective sound processing abilities[23].

4.1.2 Test Method

First of all, the test environment requirements. We require that the test needs to be carried out in an open real-life environment with a computer equipped with a corresponding 3D sound Bluetooth headset. The second is the participants. The participants consist of 4 women and 2 men, and their ages range from 20 to 50 years old.Previous studies identified age and gender as critical factors in the interaction with and processing of auditory stimuli[32]. Thus, balancing gender and using a wide variation in age may lead to a more complete understanding of how well the system performs the functions of AAR.

The test process is that the participants first sign the participant consent form, agree to the matters required by the experiment, and then the researchers inform the participants of the general content of the next experiment to prepare the participants[24]. After understanding(providing participants with comprehensive information enhances their understanding of and engagement with the research process[3]), the participants put on headphones with left and right channels. After confirming that the participants' headphones are connected to the computer, the researchers click to open the Unity scene and start the test. The participants closed their eyes throughout the process. Because the study is to convert speech to dog barking, the researchers will have a normal conversation on the left or right side of the test computer. The participants need to remember whether there is a dog barking sound in the headset along with the speech, the direction of the dog barking, and whether the dog barking can stop at the end of the speech. After a series of tests, such as 10 conversations in different positions, to make the participants say left or right, record these directions and calculate the direction accuracy at the end to determine whether the system can correctly complete the AAR function.

4.2 Functional evaluation

4.2.1 System Function Overview

The purpose of this project is to provide three functions: sound classification, sound source localization, and environmental sound simulation through advanced audio augmented reality (AAR) technology. For sound classification, by integrating the YAMNet audio event classification model, the system can identify and classify various sounds in the surrounding environment[7], such as speech, music, bird, dog, silence, etc.; for sound source localization, the left and right channels of the dual-channel microphone receive the sound strength to locate the sound source; for environmental sound simulation, similar to binaural positioning, the time difference and level difference between the left and right ears can locate the sound source[14], the audio output is adjusted according to the location of the sound source to adjust the 3D sound effect to achieve 3D simulation of spatial sound. In addition, the project also supports the conversion of real-time environmental sound into virtual sound effects, such as converting actual human voices into animal calls, to enhance the user's auditory experience and immersive feeling.

The main goal of the functional evaluation is to verify the actual performance and stability of the various functions of the project to ensure that it meets the predetermined technical requirements and user needs. Our evaluation focuses include: sound source localization accuracy: testing the accuracy and response time of the system in identifying and locating sound sources; audio event classification effect: verifying the recognition rate of the YAMNet model for different audio events in practical applications; authenticity of 3D sound simulation: evaluating the system's sound performance when simulating different sound source positions and distances; system usability: testing the system's compatibility and stability on different devices and in real environments[2].

4.2.2 Functional testing methods

Many methods are used in the process of functional testing for examining the system performance comprehensively. The first is the evaluation of the system in different environments through field testing, such as in indoor, outdoor, noisy, and quiet scenes. Testing under actual usage conditions can show a comprehensive understanding of the system's performance. It helps to reveal adaptability and stability under various conditions, in particular its stability in real environments, so as to give developers suggestions for improving it.

At the same time, operation by the user is one indispensable part of testing. We invited target users to participate in order to test the usability and functionality of the system from a user perspective. This kind of testing method helps in collecting direct feedback from users and identifying possible deficiencies in the design of the user interface and whether the functional implementation meets users' expectations. This will ensure that the system design is more user-intuitive and friendly, hence improving the experience[9].

Stress testing also helps identify exactly how stable a system is, and what its processing power might become with an increased amount of usage. In a general perspective, this testing approach can uncover the problems that might occur in the system under high traffic and big data and ensure that the system can still operate stably under high load. The long-term attack from our project would check for the system's stability, which will be very important in determining the reliability of the system when it copes with long-term use[11].

In brief, the system is analyzed with respect to levels from comprehensive testing using the method of field tests, user operating testing, and stress testing. This multi-level test-

ing method gives strong guarantees in terms of ensuring the stability, reliability, and user satisfaction of the system under diverse use scenarios.

4.2.3 Functional testing results

In this functional test, some key performance indicators of the system were evaluated, while the respective results in different environments showed the performance of the system. Sound source localization accuracy is important to evaluate how well a system works in a different environment. The sound source localization accuracy of the system reaches 80

A test of the model YAMNet was done, with the obtained results being that the classifier based on that model perfectly coped with the recognition of common environmental sounds, especially different types of traffic sounds, human voices, and natural sounds, with an average rate of more than 90

What's more, testers were highly recognized in the 3D sound simulation system. The feedback given by the user presented that the orientation of the sound source was simulated effectively, and the hearing space sense was restored. Especially on head-mounted devices, the performance is outstanding, and users can accurately perceive the direction of the sound. This suggests that the system is pretty authentic in the simulation of 3D sound, and in the future, it can continue to be optimized so that it produces a full enhanced spatial sense on the spatial distance on 3D sound effects, with changes in strong and weak, thereby enhancing the real experience of the user.

In short, the precision of sound source localization in a noisy environment still needs further improvement; the audio event classification model should be optimized to cover a broader class of sounds and the accuracy and response speed of the classification model increased while also improving cross-platform compatibility of this system. Moreover, learning more advanced algorithms for sound source localization and dynamic sound source processing technologies to enhance the ability of the system in tracing moving sound sources will also be key features that will increase user satisfaction. These directions of improvement will boost not only the performance and stability of the system but also bring users a richer and more accurate audio experience.

4.3 Non-functional evaluation

Non-functional evaluation is equally important during system development and testing. If compared to simple functional implementation, it is about performance concerning performance, reliability, security, maintainability, and scalability of a system. The subsequent paper, therefore, discusses the non-functional evaluation framework for the project in detail to make sure that during practical application, the system can be stable, secure, and feasible for future expansion.

First and foremost, performance evaluation is one of the essentials in non-functional evaluation, which pays attention to the response speed and processing capacity of the system in actual operation. Among these, the response time of the system is the key indicator. Meanwhile, the immediacy of the system and the smoothness of the users could be achieved by averaging the time taken for the system in response to what the user hears, such as the time for sound source localization and audio classification. Besides, the parallel processing capability is another critical concern when conducting performance evaluation on the system. This tests the capability of the system to process several audio inputs, channelled through it from different environments simultaneously, hence assessing its performance when under high load[20].

Reliability evaluation: This is a concern that the system will continue to operate stably under predetermined conditions without crashing or losing any data due to errors. Whether the error occurred in a system provides error-handling capability is the key point of this evaluation. Ideally, it should handle various operating mistakes properly and avoid program crash or functional failure. Meanwhile, the recovery capability is an important aspect in the assessment of reliability. When a failure occurs, the system should recover to normal as soon as possible for continuity. Otherwise, allowing the system to run continuously over a period of time, such as 72 hours, the stability of the system could be checked concerning observed issues like memory leaks or performances that deteriorate[16].

Finally, the objective of the maintainability assessment is to ensure that the system is easy to be upgraded and maintained, enables quick error repair, and adapts to changes in the environment.

Code quality is the foundation of maintainability. Code static analysis tools can be used to evaluate the complexity and standardization of the code in order to ensure that the code is readable and maintainable[22]. The integrity of technical documents is guaranteed. Detailed, easy-to-understand documents will allow developers and maintainers to quickly understand and use the system. Systems designed by modules have higher maintainability. It will also be ensured that the degree of modularity is such that each component in the system can be independently upgraded or replaced, reducing the level of complexity in doing maintenance or upgrades. In a word, the performance, reliability, maintainability, etc. of the system can be comprehensively nonlinear-functionally evaluated to gain an in-depth understanding of various performances of the system in actual applications. Such evaluation results provide not only a scientific basis for the optimization and upgrade of the system but also a solid foundation for the future development direction and strategy formulation.

4.4 Data Collection and Analysis

Collection and analysis of data are essential ingredients of any system development and evaluation process. As a matter of fact, the precision and depth of evaluation are directly related to data collection and analyses. Grounded on comprehensive methods of data collection and detailed techniques for data analysis, this research work was to have an in-depth understanding of functional performance and user experience in depth for drawing valid inferences that support subsequent decision-making and improvement efforts.

User surveys and face-to-face interviews are some of the important methods of data collection that elicit direct user experience data. This survey contains a number of questions ranging from functionality to the general user experience of the user base. The questions on the survey come in open-ended and closed-ended formats to ensure, from different perspectives, an account of users' thoughts and experiences with the system. Accuracy Scoring System: The scoring system we use for accuracy enables us to quantify user satisfaction with, and feedback on, features; creates foundational data for subsequent statistical analysis.

Meanwhile, face-to-face interviews as a method of qualitative data collection provide an opportunity for in-depth interviews with some selected core users[25]. In this treatment, we should be in a position to go in-depth with users about their experiences with the system, including perceived strengths and areas where improvement is possible.

In the interviews, particular attention is paid to real-operation troubles that users are facing and specific improvement proposals that can be put forward by them. The interviews provide great insight into user behavior and allow the finding of very small changes in the needs and expectations of users.

We will thus use both quantitative and qualitative approaches to process and interpret the data at the data analysis stage. Quantitative analysis refers to processing survey data through use of statistical methods in calculating major indicators, means, and variance; performing a trend analysis so as to observe changes of such indicators with time. In that respect, we shall objectively be able to establish the performance of the system and the trends of users' satisfaction[35].

In this respect, we will use the method of content analysis for qualitative data, such as interview content: coding and categorization of information. This involves the identification of main themes and patterns in user feedback; this needs to be done in order to find detailed opinions and feelings expressed by users. This deep analysis will help us catch users' specific views about system functionality and design more precisely.

Besides, data visualization is also one privileged means of presenting these multivariate data sets. Indeed, charts and graphs—like performance indicator line charts and user satisfaction bar charts—present the results of such an analysis intuitively, which gives decision-makers and development teams a fast understanding of key messages and trends. The results from data collection and analysis provide a basis not only for understanding system performance and user satisfaction but also for the scientific optimization and expansion of functions in the future. These will also have a direct bearing on the optimization of system performance, enhancing the experience of users, and Markup of development strategies in the future to ensure continuous progress at success.

Chapter 5: Conclusion

The project targeted the development of an application able to realize audio event recognition in real time and execute advanced spatial sound processing. Large progress was achieved by exploiting state-of-the-art audio analysis technology and 3D-sound simulation techniques. The conclusion outlines the state of the principal achievements and exposes a comprehensive agenda of future research and development.

This application has remarkably identified various audio events and dynamically changed the audio outputs to enhance the effect of 3D auditory by embedding the YAMNet model in Unity. This is especially good in an environment with multi-channel audio processing when the spatial and directional nature of sound is amplified several-fold. This not only enriches the auditory experience but makes it much more immersive, another feature which has been quite well-received in applications such as gaming and in virtual reality scenarios.

The performance of the application was optimized throughout its development. It received systematic cures for the first noticeable delays and occasional recognition errors. The iterative refinement of algorithms and fine-tuning of model parameters brought this into a robust application able to respond with high speed and accuracy in real time. These enhancements have so far meant that the application meets rigid demands of real-time audio processing, hence securing reliability and efficiency across various operational contexts.

The feedback received from the end users has been very satisfactory concerning the sound effects the application processes in 3D. Especially for immersive environments, users have appreciated improved spatial sound effects where the depth and realism provided by the application enhance the user experience greatly. Such feedback is priceless since it confirms that what has currently been implemented effectively points in the direction of further enhancements.

The project has pointed out several avenues of further development and refinement in the future:

Model Training and Optimization: Future work will consider increasing the training data to further enhance the robustness and accuracy of the model in diverse and complicated environments. Enhancement of the model's computational efficiency is also emphasized for the purpose of efficient real-time processing.

Adaptability across Scenarios: The current system, while highly effective in indoor controlled environments, will find its extension to outdoor or noisier environments of prime importance. How the algorithm can adapt to such variable acoustic conditions effectively forms the thrust area of further research.

Development of Interactive Features: The incorporation of more user-interactive features, such as setting the direction of sound and adjusting certain sound effects to individual likings, will make the system even more responsive to users' specific needs and preferences. This way, users will be able to have an audio tailored to their preference, which will increase user interaction and overall satisfaction.

Advanced Audio Processing Technologies: This could also be achieved by integrating more advanced audio processing technologies in the future, such as echo cancellation and volume balancing, to provide all-rounded audio[31]. These are technologies that would widen the coverage area for problems in audio processing, thus enhancing the usefulness of the application.

Commercialization and Market Expansion: The commercial exploitation of such technology, especially in collaboration with game developers or film and music industries, may well translate into huge economic benefits. This project has managed to attain a solid foundation in audio recognition and spatial sound-processing technologies. It promises to take the auditory experience to newer dimensions in each and every form of multimedia and industries with its latest innovations and strategic market expansions so as to bring forward the frontiers of the audio technology industry. With this, we could visualize users getting easy access to a richer real auditory experience, improving user interactions with audio within daily-use contexts[15].

Appendix A: First appendix

A.1 Section of first appendix

Appendix B: Second appendix

Bibliography

[1] R. Baumgärtel, H. Hu, M. Krawczyk-Becker, D. Marquardt, T. Herzke, G. Coleman, K. Adiloğlu, K. Bomke, K. Plotz, T. Gerkmann, S. Doclo, B. Kollmeier, V. Hohmann, and M. Dietz. Comparing binaural pre-processing strategies ii. *Trends in Hearing*, 19:233121651561791, 2015.

[2] O. Bălan, A. Moldoveanu, and F. Moldoveanu. Multimodal perceptual training for improving spatial auditory performance in blind and sighted listeners. *Archive of Mechanical Engineering*, 40:491–502, 2015.

[3] N. Chapman, R. McWhirter, M. Armstrong, R. Fonseca, J. Campbell, M. Nelson, M. Schultz, and J. Sharman. Self-directed multimedia process for delivering participant informed consent. *BMJ Open*, 10:e036977, 2020.

[4] L. Cliffe, J. Mansell, J. Cormac, C. Greenhalgh, and A. Hazzard. The audible artefact. 2019.

[5] L. Cliffe, J. Mansell, C. Greenhalgh, and A. Hazzard. Materialising contexts: virtual soundscapes for real-world exploration. *Personal and Ubiquitous Computing*, 25:623–636, 2020.

[6] Alan B Craig. Understanding augmented reality: Concepts and applications. 2013.

[7] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona. 3d tune-in toolkit: an open-source library for real-time binaural spatialisation. *Plos One*, 14:e0211899, 2019.

[8] A. Dam, A. Siddiqui, C. Leclerq, and M. Jeon. Extracting a definition and taxonomy for audio augmented reality (aar) using grounded theory. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66:1220–1224, 2022.

[9] F. Ehrler, T. Weinhold, J. Joe, C. Lovis, and K. Blondon. A mobile app (bedside mobility) to support nurses' tasks at the patient's bedside: usability study. *Jmir Mhealth and Uhealth*, 6:e57, 2018.

[10] M. Farmani, M. Pedersen, and Z. Tan. Informed sound source localization using relative transfer functions for hearing aid applications. *Ieee/Acm Transactions on Audio Speech and Language Processing*, 25:611–623, 2017.

[11] R. Gonçalves, T. Rocha, J. Martins, F. Branco, and M. Au-Yong-Oliveira. Evaluation of e-commerce websites accessibility and usability: an e-commerce platform analysis with the inclusion of blind users. *Universal Access in the Information Society*, 17:567–583, 2017.

[12] J. Hollebon, F. Fazi, and M. Gálvez. A multiple listener crosstalk cancellation system using loudspeaker-dependent regularization. *Journal of the Audio Engineering Society*, 69:191–203, 2021.

[13] H. Kim, L. Hernaggi, P. Jackson, and A. Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images. 2019.

[14] L. Kraljević, M. Russo, M. Stella, and M. Sikora. Free-field tdoa-aoa sound source localization using three soundfield microphones. *Ieee Access*, 8:87749–87761, 2020.

[15] K. Li. Catr: combinatorial-dependence audio-queried transformer for audio-visual video segmentation. 2023.

[16] K. Li, M. Yu, L. Lu, J. Zhai, and W. Liu. A novel reliability analysis approach for component-based software based on the complex network theory. *Software Testing Verification and Reliability*, 28, 2018.

[17] M. McGill, S. Brewster, D. McGookin, and G. Wilson. Acoustic transparency and the changing soundscape of auditory mixed reality. 2020.

[18] O. Mohareri and A. Rad. A vision-based location positioning system via augmented reality: an application in humanoid robot navigation. *International Journal of Humanoid Robotics*, 10:1350019, 2013.

[19] P. Mokhtari, H. Kato, H. Takemoto, R. Nishimura, S. Enomoto, S. Adachi, and T. Kitamura. Further observations on a principal components analysis of head-related transfer functions. *Scientific Reports*, 9, 2019.

[20] V. Nagaraju, L. Fiondella, P. Zeephongsekul, C. Jayasinghe, and T. Wandji. Performance optimized expectation conditional maximization algorithms for nonhomogeneous poisson process software reliability models. *Ieee Transactions on Reliability*, 66:722–734, 2017.

[21] A. Neidhardt, C. Schneiderwind, and F. Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms - literature review and theoretical framework. *Trends in Hearing*, 26:233121652210929, 2022.

[22] S. Paradkar. A framework for modeling non-functional requirements for business-critical systems. *SSRN Electronic Journal*, 2021.

[23] M. Park, J. Song, S. Oh, M. Shin, J. Lee, and S. Oh. The relation between nonverbal iq and postoperative ci outcomes in cochlear implant users: preliminary result. *Biomed Research International*, 2015:1–7, 2015.

[24] M. Prictor, M. Lewis, A. Newson, M. Haas, S. Baba, H. Kim, M. Kokado, J. Minari, F. Molnár-Gábor, B. Yamamoto, J. Kaye, and H. Teare. Dynamic consent: an evaluation and reporting framework. *Journal of Empirical Research on Human Research Ethics*, 15:175–186, 2019.

[25] E. Raita. User interviews revisited. 2012.

[26] R. Ranjan and W. Gan. Natural listening over headphones in augmented reality using adaptive filtering techniques. *Ieee/Acm Transactions on Audio Speech and Language Processing*, 23:1988–2002, 2015.

[27] S. Russell, G. Dublon, and J. Paradiso. Hearthere. 2016.

[28] C. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki. Design theory for binaural synthesis: combining microphone array recordings and head-related transfer function datasets. *Acoustical Science and Technology*, 38:51–62, 2017.

[29] D. Satongar, C. Pike, Y. Lam, and A. Tew. The influence of headphones on the localization of external loudspeaker sources. *Journal of the Audio Engineering Society*, 63:799–810, 2015.

[30] J. Segura-García, J. Navarro, J. Pérez-Solano, J. Montoya-Belmonte, S. Felici-Castell, M. Cobos, and A. Aranda. Spatio-temporal analysis of urban acoustic environments with binaural psycho-acoustical considerations for iot-based applications. *Sensors*, 18:690, 2018.

[31] Z. Sun, P. Sarma, W. Sethares, and Y. Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. 2019.

[32] B. Tobey, M. Gallego-Abenza, D. Reby, and N. Mathevon. Dog-directed speech: why do we use it and do dogs pay attention to it? *Proceedings of the Royal Society B Biological Sciences*, 284:20162429, 2017.

[33] L. Tronchin, F. Merli, M. Manfren, and B. Nastasi. Validation and application of three-dimensional auralisation during concert hall renovation. *Building Acoustics*, 27:311–331, 2020.

[34] Y. Vazquez-Alvarez, M. Aylett, S. Brewster, R. Jungenfeld, and A. Virolainen. Designing interactions with multilevel auditory displays in mobile audio-augmented reality. *Acm Transactions on Computer-Human Interaction*, 23:1–30, 2015.

[35] M. Walji, E. Kalenderian, M. Piotrowski, D. Tran, K. Kookal, O. Tokede, J. White, R. Vaderhobli, R. Ramoni, P. Stark, N. Kimmes, M. Lagerweij, and V. Patel. Are three methods better than one? a comparative assessment of usability evaluation methods in an ehr. *International Journal of Medical Informatics*, 83:361–367, 2014.

[36] Y. Wu, W. Che, and B. Huang. An improved 3d registration method of mobile augmented reality for urban built environment. *International Journal of Computer Games Technology*, 2021:1–8, 2021.

[37] J. Yang, A. Barde, and M. Billinghurst. Audio augmented reality: a systematic review of technologies, applications, and future research directions. *Journal of the Audio Engineering Society*, 70:788–809, 2022.

[38] T. Yang and J. Kang. Perception difference for approaching and receding sound sources of a listener in motion in architectural sequential spaces. *The Journal of the Acoustical Society of America*, 151:685–698, 2022.

[39] Z. Zhou, A. Cheok, X. Yang, and Y. Qiu. An experimental study on the role of software synthesized 3d sound in augmented reality environments. *Interacting With Computers*, 16:989–1016, 2004.