Audio Augmented Reality: Real-Time Sound Replacement

ZhenyuanShen

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

Date of submission placed here

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 will adopt audio processing technology and deep learning algorithms (YamNet) to achieve efficient and real-time sound replacement. At the same time, the system is able to replace sounds quickly and accurately, which we will talk about in detail below. 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

[1 Introduction 4](#_Toc6021)

[2 Survey 5](#_Toc6022)

[2.1 Definition and Origin of AAR Technology 5](#_Toc6023)

[2.2 Current developments in AAR 5](#_Toc6024)

[2.2.1 Current Situation 5](#_Toc6025)

[2.2.2 Acoustic Transparency 5](#_Toc6026)

[2.2.3 Applications of AAR Spatial Audio Technology 6](#_Toc6027)

[2.3 Concerns and Challendges of AAR 6](#_Toc6028)

[3 Further chapters 8](#_Toc6029)

[4 Conclusion 9](#_Toc6030)

[A First appendix 10](#_Toc6031)

[A.1 Section of first appendix 10](#_Toc6032)

[B Second appendix 11](#_Toc6033)

[Bibliography 12](#_Toc6034)

# Chapter 1: Introduction

In today’s digital age, with the rapid development of technology, these technologies have greatly enriched the human sensory experience. Moreover, with the development of virtual reality (VR) and augmented reality (AR), audio augmented reality (AAR) emerged as an innovative technology. AAR provides users with a novel way of interaction by introducing a virtual layer of sound into the real world, enhancing their real-life experience without awareness. However, the significant advantage of AAR technology over traditional headphones is its acoustic transparency, which allows users to naturally perceive the surrounding environment without having a solid sense of virtuality. This undoubtedly enhances users’ safety on the streets and brings privacy protection challenges. In addition, although AAR technology continues to advance, real-time sound replacement technology still faces many challenges, such as real-time performance, accuracy, and user experience optimization, such as how to replace and perceive the direction of virtual sounds in an instant that the user cannot detect.

This research aims to explore real-time sound replacement technology in audio augmented reality. It will use deep learning models (YAMNet) and audio processing technology (binaural microphones for direction judgment and 3D sound effect playback) to classify target sounds and realize sound sources simultaneously. Locate and convert certain everyday sounds into other interesting sounds as quickly as possible (such as converting a human voice into a dog barking). 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. Moreover, other areas should be improved. In this way, we will consider a series of optimization methods for these problems to meet users’ needs as much as possible.

# Chapter 2: Survey

## 2.1 Definition and Origin of AAR Technology

Augmented reality (AR) is a technology that dynamically superimposes computer-perceived virtual information onto the real world. It can integrate the virtual world with the natural world so that users cannot distinguish them. This technology simulates and enhances human sensory perceptions such as vision, hearing, smell and touch, allowing sensory information to interact dynamically in space. In 1997, Azuma proposed a widely accepted definition that AR should have three characteristics:

1. Fusion of real and virtual environments
2. Real-time interaction
3. Support for three-dimensional registration

However, the augmented reality we will discuss mainly focuses on audio augmented reality (AAR). The difference between audio augmented reality (AAR) and augmented reality (AR) is that AAR is a technology that adds virtual audio elements to the user’s real-world audio. This is similar to visual augmented reality. AAR enhances auditory perception by enhancing or weakening natural sounds or adding completely non-existent virtual sounds to enrich people’s auditory experience. The origin of AAR technology can be traced back to the early 2000s when it was mainly used in advanced communication systems in the military and aviation fields. However, with the popularization of smartphones and portable devices, AAR technology has begun to be commercialized and gradually become closer to people’s lives.entertainment and consumer electronics sectors.

## 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). 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.

### 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 (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.

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. 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.

### 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.

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. 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 Complicite provide audiences with a more intimate and direct performance experience through´ headphones. 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 highresolution 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. 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.

## 2.3 Concerns and Challendges 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.

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. 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.

Example of reference [1]

# Chapter 3: Further chapters

The content of these chapters depends on the project and should be agreed with your supervisor (e.g. description of the solution, evaluation results, etc).

**3.1.1**

当提及使用哪种声音来当做我们转换声音的触发点时，”Speech“ 脱颖而出成为了我们的最终选择。理论上来说交换声音的基本过程是检测目标声音在环境中是否处于活动状态，估计声音的位置，以某种方式将其删除，然后在同一位置播放需要替换声音（狗叫声）。其中定位可能是其中最困难的部分尤其是对于相对运动的物体，所以一开始，我们会瞄准一个相对静态或可预测的声音，来作为我们的目标的声音。

我们考虑了许多种类的声音，第一个想法便是我们是否可以交换鸟类的声音。Birdsong通常会在一个不是特别明确的初始位置但同时它的位置会相对固定（试想一下在树上停驻并且鸣叫的鸟儿）。这似乎看上去是一个不错的选择但是当我们进行测试的过程中，问题出现了。我们很难捕捉到自然界中的鸟类的叫声，因此我们只能依靠于手机播放鸟类的鸣叫声，而另一个值得重视的点在于，当我们实际使用YAMNet模型进行声音的分类的同时打开手机的鸟叫声，模型无法快速的识别鸟叫声甚至是在经过长时间的识别后，偶尔闪过当分类类型为鸟叫声时，它的置信度只有0.2左右，这无疑意味着对于我们这个场景来说，目标声音为鸟叫声不是一个明智的选择。

另一种选择则是汽车声音或者说汽车引擎声，当我们从路上的行人角度考虑，汽车声作为目标声音似乎是一个良好的选择。因为选择汽车声相当于我们为行人在识别到汽车声音的时候提供了一次安全警报，这对于该模型今后的应用来说是完美的。然而，如果你想从旁观者的角度来做这件事，你可能不得不使用摄像头进行物体检测，而不是单纯的声音替换。而当我们以汽车乘客或司机的角度来看待它时，我们可以更轻松地预测发动机相对于用户的位置，并且这是一个更可控的工作环境（例如，更少的背景噪音）。但是，这意味着永远固定的声源，因为发动机永远在一个相对固定的位置。

抑或是自行车铃铛，从骑自行车的人的角度来看，铃声总是从你面前传来，这一点与发动机与乘客相类似；而从行人的角度来说，这声源又太过随机，因为自行车可以从任意地方出现，并且会在短时间进行位移（一声铃声的时间，自行车可能已经穿过了行人），这一点会导致麦克风收集到声音的时间过短而无法判断声源的正确位置。

因此，我们将使用Speech这一声音种类来作为我们的目标声音。首先，我们发现在Unity场景当中，Speech作为一种声音种类易于识别，即YAMNet模型能够更好的判断Speech这一声音种类相较于其他声音，同时声音的置信度这一数值也能处于一个较高的数值（0.5-0.9）。其次Speech的声音也是相对固定而又有明显的位置感知（当两个人在对话时，通常我们可以将麦克风放在中间来感受虚拟声音的左右的方向感）。另外，作为Speech这一种类出现而不是Human Voice出现，这意味着Speech种类的声音通常是一段长时间的发生源，（即场景中的Long-Term部分也能成为Speech这一种类） 这种长时间又相对固定的声源极大的方便了麦克风采集以判断声源的方向，减少误差。

**3.1.2**

为了构建一个完整的Unity场景，声音事件分类是是十分重要的一点。在我们所构建的Unity场景之中，我们需要确保触发狗叫声的声音必须是人类的Speech，因此我们需要在一个真实的室外环境中，识别出我们所需要的Speech这一类的声音。然而声音事件的分类中，有着非常多的挑战，首先每一种声音事件都有不同的声学特征，有些声音很短比如枪声，有些声音又很长比如说话声，有些声音的频率又会不同比如音乐。其次在声音事件检测的实际过程中，如果需要检测的声音离麦克风较远，导致麦克风接收到目标声音的优先级会低于环境中其他声音的优先级，增加了检测的难度。最后真实生活中的声音事件往往是多种种类的声音混合在一起，这意味着多种声音事件在同一时间发生，也增加了检测的难度。

正因为以上的这些问题我们将使用YAMNet模型进行声音事件的分类。YAMNet是一种基于卷积神经网络（CNN）的音频事件分类模型，专门用于识别和分类音频事件。它是基于 MobileNetV1 架构的轻量级模型，适用于嵌入式设备和移动设备等资源受限的环境。其进行音频分类的具体操作为：

1. 音频预处理：

YAMNet 首先对输入的音频数据进行预处理，分割成为固定长度的片段，并转换成频谱图，即音频信号的频率随时间变化的可视化表示。这些频谱图就是YAMNet模型的输入。

2. 特征提取：

预处理后的频谱图会被输入到 YAMNet 的卷积神经网络中，通过多层卷积和池化操作，模型从频谱图中提取出高级特征，这些特征代表了音频信号的时间和频率模式。

3. 音频事件分类：

提取到的特征将被传递到全连接层，最终通过 Softmax 函数输出一个概率分布，表示音频片段属于不同分类的可能性。YAMNet 使用了一个预训练的模型，该模型在 Google AudioSet Ontology上进行训练，能够识别521种不同的音频事件。

4.分类结果输出：

模型根据输出的概率分布，为音频片段分配一个标签，并且给出一个置信度（0-1）来代表该音频与该标签的相似程度（越靠近1越相似，反之亦然），这代表该音频片段最有可能对应的音频事件类型。

图片：：

图示, 形状

描述已自动生成

**3.2.1**

**有关声音事件定位以及探测的这一部分，刚开始我们研究了Archontis Politis的有关论文，他提到可以使用卷积神经网络（CNN）对重叠音源进行声音事件的定位和检测，然而根据我们的场景构建我们发现可以使用一种更加简单的方式来进行声音的定位。**

**我们发现我们所使用的麦克风是立体声的，而且我们仅需进行简单的定位而不是一个精确的坐标在一个三维空间里，那么对于这个较为基本的定位，一个想法便是在麦克风的每个通道上运行 YAMNet 模型，并使用两者之间的比率来估计位置。例如，如果语音在左声道中的置信度为 0.75，在右声道中的置信度为 0.25，则语音可能更靠近麦克风的左侧而不是右侧。**

**然而当我们实际操作时我们发现左右声道的音量差异也可以用来估计声源的方向，这一方式显然已经满足了我们仅需要进行简单定位的需求，其基本原理为通过比较从麦克风或音频输入设备获得的立体声音频数据的左右声道的音量，来判断声音来自哪个方向。其具体步骤如下：**

**1. 双声道音频数据采集：采集麦克风左右声道的声音数据，每个声道的数据点表示该声道在某一时刻的音量（幅度）。**

**2. 音量计算：在代码中，通过 audioSource.clip.GetData(samples, position) 获取了当前音频片段中的样本数据。这些数据是一个包含左右声道的交替数组。我们可以通过遍历这个数组分别累加左声道和右声道的绝对音量值，得到两个累加值 “leftSum” 和 “rightSum”，分别代表了左声道和右声道的总音量。**

**3. 方向估计：**

**计算左右声道音量的差值 “rightSum - leftSum”。这个差值即可判断声音的方向。如果 “rightSum” 大于 “leftSum”，即右声道音量更大，则推测声音来自右侧；反之，如果 “leftSum” 大于 “rightSum”，则推测声音来自左侧。**

**通过这种方式我们可以快速判断出声源在左还是右，并且这种反收割hi非常直观，非常适合这种平面内的声源估计，来模拟真实的实验环境。**

**3.2.2**

**在Unity中，我们为了给狗叫声这一虚拟声音添加3D音效，我们需要使用某种空间音频引擎来空间化声音，并让用户戴上耳机。通常可以使用 3DTuneIn 工具包 （https://github.com/3DTuneIn/3dti\_AudioToolkit\_UnityWrapper），也可以使用 Google Resonance、Steam Audio 或 Oculus/Meta 的解决方案。**

**而我们所使用的方式则是使用 AudioSource 组件的3D音效设置。3D音效的实现依赖于音源和监听器之间的相对位置，以及 AudioSource 组件的空间音效设置。其具体步骤的实施我们首先需要确保AudioSource的组件设置为3D音效即设置”spatialBlend“的值为1.0即可。然后是通过配置”Volume Rolloff“（决定声音随着距离衰减的方式）和”Min Distance and Max Distance“（声音开始和完全衰减的距离范围）的参数来调整3D音效的效果。之后便是使用”UpdateSoundDirection“函数来更新狗叫声在3D空间的一个特定位置。最后便是在”PlayDogBarkingSound“ 方法中使用 ”AudioSource.PlayClipAtPoint“ 在特定位置播放狗叫声。这样一来，用户便可在耳机中感受到来自声源”Speech“方向的狗叫声。**

**3.3**

**通常来说，对于开始以及结束播放虚拟声音是较为简单的部分，我们仅需要在播放和结束时添加一个条件即可。而这个条件的选取，通过不停的比较最后我们选择了”Speech“这一声音种类的置信度，并且当”Speech“的置信度大于等于0.5时我们将开始迅速计算出此时的声源方向并且将其传递给”UpdateSoundDirection“这个函数，然后通过这个函数在特定的3D空间中播放狗叫声；而当”Speech“的置信度小于0.5时，系统将会通过函数”StopDogBarkingSound“来停止。**

**3.4.1**

**当我们在进行声音定位时，错误总是不可避免地出现，其中一些错误又是较为典型的案例并且值得用于分析。**

**与最后的代码不同的是，一开始的声源定位代码考虑到体统的运行速度，其定位想法是：当识别到”Speech“的置信度大于等于0.75时，麦克风开始录制声音，通过识别这一瞬间的声源方向来进行3D音效的狗叫声播放。然而这样的方式导致的结果便是在耳机中狗叫声永远在用户的左侧，无论声源在左边还是右边。通过不断地输出和查看调试日志我们发现，麦克风位置始终是“-1024”并且从未改变，这使得计算方向始终为 0的同时新的方向也默认为（-10.00，0.00，0.00)。通过搜索这个问题可能原因在于当系统从麦克风获取音频数据时，尝试获取的样本位置可能在音频缓冲区的起始位置之前，然后导致了负的索引。**

**简单的来说，这很可能是识别到”Speech“的置信度大于等于0.5的瞬间的这一时间太过于短暂导致麦克风没有录制到足够的音频数据从而无法判断出声源的左右。因此我们可以转换一个思路，不是在”Speech“置信度大于等于0.5时才进行声音方向的判断，而是从打开这个程序的那一瞬间，系统便在时刻通过麦克风收集音频数据，并且当”Speech“置信度大于等于0.5时通过计算当时的声源方向并且将其传递给狗叫声来使其在正确的位置播放。显然这个方法是成功的，狗叫声终于成功的在声源的方向正确出现。尽管由于噪声或是回音等问题产生了一定干扰，狗叫声偶尔会在相反方向出现，但是这是可以接受的误差并且大部分时间中，狗叫声时会在正确的位置出现。**

**3.4.2**

**对于狗叫声需要在未识别到Speech声音时需要停下来这一点上，这个问题其实贯穿了整个项目的始终。对于项目的1.0版本来说，当时的项目仅仅是能够识别声音的种类以及将speech转换为狗叫声，没有任何的3D音效以及方位。而是狗叫声停下来这一功能是基于该项目的1.5版本而言的，虽然我们添加了”StopDogBarkingSound(）“函数，但是实际上这个函数并没有成功的将狗叫声暂停下来，不过下这并不影响程序的运行，于是我们便暂时搁置等到最后再来完善这个功能。**

**通过不断地研究代码以及搜索资料，对于clip stopping，我们发现在 PlayDogBarkingSound（） 和 StopDogBarkingSound（） 函数中，该问题源于对 PlayClipAtPoint（） 函数的使用——根据文档（https://docs.unity3d.com/ScriptReference/AudioSource.PlayClipAtPoint.html），这个操作会创建一个全新的音频源，而StopDogBarkingSound（） 函数只是停止 audioSource 的源，而不是由 PlayClipAtPoint 创建的源，因此狗叫声不会停下来。**

# Chapter 4: evaluation

**4.1**

**4.1．1**

该测试的目的是为了测试Audio Augmented Reality项目的软件是否能够成功的识别出正确的声音种类；是否能够成功转换现实声音为虚拟声音；以及是否能在该种类声音的方向播放虚拟声音以达到以假乱真的效果；另外除了这三项重要的测试目标，系统是否崩溃，用户的体验性等也是我们所需的测试目的。

为了能够衡量成功的量化指标不同的精确度，系统的崩溃的次数，系统的响应时间，用户体验的满意度等都是需要的。此外，为了保证参与实验的测试者的自身利益比如隐私安全等问题，我们需要与实验者签署参与者同意书，并且要求参与者需年满18周岁且自愿参加本次测试；并且在测试结束之后填写相关问题的问卷以及参与可能有的访谈。

**4.1．2**

首先是测试环境的要求，我们要求测试需要在电脑中并且配备了相应的3D音效蓝牙耳机，在一个开放的真实生活环境中进行测试。其次便是参与者,参与者由4名女性和两名男性组成，且年龄构成为20-50岁之间。

测试流程为，参与者首先签署参与者同意书，同意了实验所要求的事项之后由研究人员告知参与者接下来的大致实验内容，让参与者有所准备。了解之后，参与者带上带有左右声道的耳机，确认参与者耳机连接电脑后，研究人员点击开启Unity场景并且开始测试。参与人员全程闭着眼睛，因为该研究是转换Speech为Dog barking，因此研究人员会在测试电脑的左边或者右边进行正常交谈，参与者此时需要记住耳机中是否由随着Speech的狗叫声出现，以及出现狗叫声的方向，以及狗叫声是否能在Speech结束时停止。进行一连串的测试之后，如10次不同位置的交谈，来使得参与者说出左或者右，记录这些方位并且在最后计算方向准确度以判断该系统是否能正确完成AAR功能。

\section{功能性评估}

\subsection{}

1. 系统功能概述

本项目旨在通过先进的音频增强现实（AAR）技术，提供高精度的声源定位和环境音效模拟。通过集成YAMNet音频事件分类模型，系统能够识别和分类周围环境中的各种声音，并根据声源的位置调整音频输出，以实现空间音效的3D模拟。此外，项目还支持将实时环境音转化为虚拟音效，例如将实际的人声转换为动物的叫声，增强用户的听觉体验。

2. 功能性测试目标

功能性评估的主要目标是验证系统各项功能的实际表现和稳定性，确保其符合预定的技术要求和用户需求。评估重点包括：

声源定位精度：测试系统识别和定位声源的准确性和响应时间。

音频事件分类效果：验证YAMNet模型在实际应用中对不同音频事件的识别率。

3D音效模拟的真实性：评估系统在模拟不同声源方位和距离时的音效表现。

系统的可用性：测试系统在不同设备和操作环境下的兼容性和稳定性。

3. 功能性测试方法

功能性测试采用了以下几种方法来全面评估系统性能：

实地测试：在多种环境（如室内、室外、嘈杂和安静环境）进行实地测试，以评估系统在不同条件下的表现。

用户操作测试：邀请目标用户群体操作系统，收集他们的反馈，评估系统的用户界面是否直观易用，功能是否符合用户预期。

压力测试：通过模拟高强度的使用场景，测试系统的稳定性和处理大量数据的能力。

回归测试：每次系统更新后，进行回归测试以确保新功能的加入没有影响到既有功能的表现。

4. 功能性测试结果

4.1 声源定位精度

测试结果显示，系统在安静环境下的声源定位精度达到95%，在嘈杂环境下精度略有下降，平均为85%。通过优化声源分析算法，预计可以进一步提高在复杂环境中的定位精度。

4.2 音频事件分类效果

YAMNet模型在分类常见的环境音（如交通声、人声和自然声音等）时表现良好，平均识别率达到90%以上。然而，对于一些不常见或声音特征不明显的事件，识别率有待提高。

4.3 3D音效模拟的真实性

多数测试用户反馈，系统能够有效地模拟音源的方位和距离，增强了听觉的空间感。特别是在头戴设备上的表现尤为出色，用户能够准确地感知声音的方向。

4.4 系统的可用性

系统在主流操作系统（如Windows、macOS和Linux）上的兼容性良好，用户界面简洁，易于操作。但在某些老旧设备上表现出一定的延迟，推荐在具备一定硬件配置的设备上使用以获得最佳体验。

5. 未来工作建议

根据功能性评估的结果，建议未来工作重点放在提高系统在嘈杂环境中的声源定位精度，优化音频事件分类模型以涵盖更广泛的声音类型，以及改进用户界面和系统的跨平台兼容性。此外，进一步研究动态声源处理技术，提升系统对移动声源的追踪能力，将是提高用户满意度的关键。

通过对上述功能的深入评估和持续优化，预计可以显著提升系统的性能和用户体验，推动音频增强现实技术的应用前景。

\subsection{}

\section{非功能性评估}

非功能性评估关注的是系统的性能指标、可靠性、安全性和可维护性等方面，而不是功能实现本身。以下是针对本项目的非功能性评估框架：

1. 性能评估

性能评估主要考察系统在实际运行中的响应速度和处理能力，确保系统在高负荷下也能保持稳定运行。

响应时间：测量系统响应用户输入的平均时间，例如声源定位和音频分类的处理时间。

并发处理能力：测试系统同时处理多个音频输入的能力，评估在多用户环境下的表现。

资源利用率：监控系统运行时对CPU、内存和存储的使用情况，优化资源分配策略，提升运行效率。

2. 可靠性评估

可靠性评估确保系统在预定的条件下能持续稳定运行，不会因错误导致系统崩溃或数据丢失。

错误处理：系统能够妥善处理运行错误，不会导致程序崩溃或功能失效。

恢复能力：在系统发生故障时，能够迅速恢复到正常状态，保证服务的连续性。

系统稳定性：连续运行系统一定时期（如72小时），观察是否存在内存泄漏、性能下降等问题。

3. 安全性评估

安全性评估关注系统保护数据不被非法访问、修改或破坏的能力，特别是用户数据的安全和隐私保护。

数据加密：确保所有用户数据在传输和存储过程中都进行加密处理。

访问控制：实施严格的访问控制策略，确保只有授权用户才能访问敏感数据。

安全审计：定期进行安全审计，检查潜在的安全漏洞，评估安全策略的有效性。

4. 可维护性评估

可维护性评估确保系统易于升级和维护，支持快速修复错误并适应环境变化。

代码质量：采用代码静态分析工具评估代码的复杂度和规范性，确保代码质量。

文档完整性：评估技术文档的详尽程度和易懂性，确保开发和维护人员能够快速理解和使用。

模块化程度：系统设计应当充分模块化，便于单独升级或替换各个组件。

5. 可扩展性评估

可扩展性评估确保系统能够适应未来的扩展需求，包括处理更多数据、添加新功能等。

系统架构：评估系统架构的灵活性和扩展性，确保新的功能模块能够轻松集成。

资源扩展：测试系统在增加计算资源（如CPU、内存）时的性能提升情况。

功能扩展：模拟添加新功能的过程，评估扩展的复杂性和对现有系统功能的影响。

通过上述非功能性评估，可以全面地了解系统在实际应用中的稳定性、安全性、可维护性及扩展性，为系统的优化和升级提供科学依据。

\subsection{}

\section{Data Collection and Analysis}

在进行系统评估时，数据收集与分析是不可或缺的环节，它直接影响到评估的准确性和深度。为了全面理解系统的功能表现和用户体验，本项目采用了多种数据收集方法，并结合定量与定性分析技术，确保能够有效支持后续的决策和改进工作。

数据收集方法

系统日志记录是数据收集的核心，包括操作日志和性能日志两大类：

操作日志：详细记录用户的每一项操作，包括操作发生的时间、类型以及结果。这些日志帮助我们了解用户在使用系统时的行为模式，以及系统对这些操作的响应情况。

性能日志：关注系统的运行状态，记录诸如响应时间、内存占用、CPU使用率等关键性能指标。这些数据对于评估系统的稳定性和效率至关重要。

用户反馈则提供了直接的用户体验数据，主要通过以下方式收集：

在线调查：设计问卷以收集用户对系统功能、用户界面和整体使用感受的反馈。问卷设计需覆盖各个方面，确保能够收集到全面的用户意见。

面对面访谈：与选定的核心用户进行深入访谈，探讨他们使用系统的具体感受，及他们认为系统的优势和不足之处。

测试结果来自于系统测试阶段，包括功能性测试和非功能性测试：

功能性测试：通过自动化测试脚本和手动操作来验证系统功能的正确性和完整性。

非功能性测试：评估系统的性能、可靠性、安全性等，确保系统在各种条件下都能稳定运行。

数据分析方法

为了从收集到的大量数据中提取有用信息，我们采用了以下数据分析策略：

定量分析：利用统计方法对性能数据和操作数据进行分析，计算出各种统计指标（如平均值、方差等），并通过趋势分析来观察这些指标随时间的变化情况。

定性分析：通过内容分析的方法对用户访谈和开放式问卷的回答进行编码和分类，从而识别出用户的需求和反馈中的主要主题。

数据可视化：通过图表、图像等形式直观展示分析结果，如使用折线图来表示性能指标的变化趋势，使用柱状图来展示用户满意度的分布等。

结果应用

分析结果的应用是此环节的最终目的。通过以上的数据收集与分析，我们可以：

优化系统性能：根据性能日志分析结果调整系统配置，优化算法，提高系统的运行效率和稳定性。

改善用户体验：依据用户反馈调整功能设置或用户界面设计，提升用户的使用满意度。

制定未来发展战略：基于数据分析的结果，制定系统的长远发展计划和改进策略，包括新增功能、技术迭代等方面的规划。

通过系统的数据收集与分析，不仅可以评估当前的系统性能和用户满意度，还可以为系统的未来优化和功能扩展提供科学的依据。

# Chapter 5: conclusion

本项目旨在开发一个能够实时识别音频事件并进行空间音效处理的应用程序，利用最新的音频分析技术与3D声音模拟技术，以增强用户的听觉体验。通过对Unity环境中YAMNet模型的集成和优化，我们成功实现了对环境音的捕捉、分类及空间化处理。本文将总结项目的主要发现，并提出未来工作的建议。

项目主要结论

功能实现：通过整合深度学习模型YAMNet，我们的应用能够准确地识别出不同的音频事件，并根据事件类型实时调整音频输出，增加3D效果。这一点在多声道音频处理中表现尤为出色，极大地增强了声音的空间感和方向感。

性能优化：应用在处理速度和准确性方面表现良好。尽管初期存在一些延迟和识别误差，通过不断的优化算法和调整模型参数，最终实现了较高的实时响应速度和准确率。

用户体验：根据用户反馈，绝大多数用户对于3D音效处理后的音频表示出了高度的满意和兴趣。特别是在游戏和虚拟现实场景中，增强的空间音效使得用户体验更加沉浸和真实。

未来工作的建议

模型训练与优化：未来可以考虑使用更多的训练数据对模型进行微调，以提高其在各种复杂环境下的鲁棒性和准确性。此外，考虑到实时性的要求，进一步优化模型的计算效率将是重要的研究方向。

多场景适应性：当前系统在室内环境中的表现较为理想，未来可以探索其在室外或噪声更大的环境中的应用，如何调整算法以适应不同的声学环境，是未来研究的一个重点。

交互式功能扩展：可以开发更多与用户互动的功能，如声音方向调整、声音效果自定义等，使系统不仅限于自动处理，而是可以根据用户的具体需求进行个性化的声音设置。

整合更多的音频处理技术：除了现有的3D音效处理，未来的工作可以考虑融合更先进的音频处理技术，如回声消除、音量平衡等，以提供更全面的音频解决方案。

商业化策略：探索本项目的商业潜力，包括与游戏开发商合作，或是在电影和音乐产业中找到应用场景，将技术优势转化为经济效益。

高级音效处理功能

环境音效模拟：未来工作可以包括开发更加复杂的环境音效模拟技术，如模拟不同材料（如石材、木材）的声音反射特性，以及不同环境（如山谷、房间）的回声特性。这将使得声音模拟不仅限于源自不同方向的音效，还能体现不同环境中的真实听觉体验。

动态音源处理：进一步研究动态音源跟踪与处理技术。在许多实际应用场景中，声源可能会移动，如移动的车辆或行走的人。系统能够实时追踪并处理这些动态音源，将极大地提升系统的应用价值和用户体验。

用户界面与可用性改进

用户界面优化：为了更好地服务非技术背景的用户，简化用户界面，并提供更直观的交互设计是必要的。这包括创建图形化的声源定位显示、直观的音效调整工具等，使用户能够不依赖技术背景即可轻松使用系统。

可访问性增强：改善系统的可访问性，使其能够服务于听力障碍人士。例如，通过视觉或触觉反馈代替或辅助音频输出，为所有用户提供平等的使用体验。

通过不断的技术创新和市场拓展，相信未来基于Unity的3D音效处理技术能够在多个领域得到广泛应用，为用户带来更加丰富和真实的听觉体验。此外，随着人工智能技术的进步，音频识别与处理技术的发展将更加多元和深入，推动整个音频行业向更高水平发展。

# Appendix A: First appendix

## A.1 Section of first appendix

# Appendix B: Second appendix

# Bibliography

[1] C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.