Title	ARMixer : live stage monitor mixing through gestural interaction in augmented reality
Sub Title	
Author	黄, 维涵(Huang, Weihan) 南澤, 孝太(Minamizawa, Kōta)
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2021
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2021年度メディアデザイン学 第876号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002021-0876

Master's Thesis
Academic Year 2021

ARMixer: Live Stage Monitor Mixing through Gestural Interaction in Augmented Reality

Keio University
Graduate School of Media Design

Weihan Huang

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Weihan Huang

Master's Thesis Advisory Committee:

Professor Kouta Minamizawa	(Main Research Supervisor)
Professor Akira Kato	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Kouta Minamizawa	(Chair)
Professor Akira Kato	(Co-Reviewer)
Professor Nanako Ishido	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2021

ARMixer: Live Stage Monitor Mixing through Gestural Interaction in Augmented Reality

Category: Design

Summary

Stage monitor mixing plays an important role in live music performances. Existing stage monitoring systems have various problems such as the risk of whistling from wedge monitors and the high cost of in-ear monitors for small venues. Additionally, the communication efficiency of mixing between musicians and sound engineers is also a challenge. finally, the stage metaphor is not widely used as a visual metaphor for audio mixing interfaces.

In this thesis, we introduce ARMixer, a system which uses the stage metaphor for its interface, allows musicians to perform an in-situ self-stage monitor mixing through gestures in augmented reality (AR) and provides a user-friendly, intuitive, and customizable mixing experience. We thoroughly describe the design, implementation, and validation of ARMixer. The results of the two usability tests show that ARMixer has a satisfactory acceptance rate from the participants, and excellent psychoacoustic intuitiveness in terms of mixing parameter controls by gestures and identifying mixing target. Furthermore, the stage metaphor can be applied to stage monitor mixing scenario well.

Keywords:

stage monitor mixing, audio mixing interface, stage metaphor, gesture, augmented reality

Keio University Graduate School of Media Design

Weihan Huang

Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to Professor Kouta Minamizawa for his kindness in supporting and guiding my master's thesis. He always gave me wise advice on the scientific research process.

I would like to thank my sub supervisor Professor Akira Kato for his guidance and encouragement.

I appreciate Project Assistant Professor Yun Suen Pai for his willingness to frequently meet with me and for guiding the technical evaluation.

I appreciate the participants who were involved in the experiment, for their time, patience, and valuable feedback.

Thank you, Stephanie Bourgeois and Zhiqian He for their constant help.

Last but not least, I would like to express gratitude to my parents for the endless support of my study.

Chapter 1
Introduction

In this chapter, we discuss the background on audio mixing, including the audio mixing interface metaphors, various examples of music production using virtual reality, and the problems of the stage monitor mixing systems. We also state the research goals and questions.

1.1. Audio Mixing in Music Production

Audio Mixing

Audio mixing is the process in music production where sounds from multiple signals are integrated into a single stereo or mono channel. The mixed sound signals, that may come from different instruments, vocals or orchestration, are included in a live performance or a recording studio. During the mixing process, the mixer adjusts the frequency, dynamics, sound quality, positioning, reverberation, and sound-field of each original signal individually to optimize each channel, and then overlay it on the final master channel. This process produces a layered effect that the general audience cannot hear in a live recording. For the recorded music, a good mix can sharpen the emotional message of a musical piece, make it more appealing to the listener, and boost commercial success [1]. And for the live performance, the mixes play an enormous role in the live listening experience because there are no second chances for musicians to perform again in each performance.

Audio Mixing Interface Metaphors

As audio mixing plays an important role in the music production process, the audio mixing interface (AMI) metaphor is a topic that cannot be overlooked. Using the visual metaphor of AMI can help users be familiar with the interface layout,

recall mixing knowledge, and complete a meaningful and intuitive mixing task. The research in sonic interaction design has shown that mapping body movements onto parameters of sound can build on the interface with embodied concepts which are much easier for users to learn [2]. Currently there are two cognitive metaphors governing audio mixing production: channel stripe metaphor and stage metaphor.

The channel stripe paradigm (CSP) is the earliest visual metaphor used in AMI and is still widely used today. Tom Dowd is credited with being the first person to create the mixing console with CSP. He used eight faders instead of knobs to adjust audio channels which allowed him to mix more intuitively [2]. Therefore, CSP is presented as repeated vertical strips of controls that feature faders, knobs and buttons to control specific parameters of one-to-one mapping channels (Figure 1.1 Left). It is not only used in physical mixing consoles, but also reflected in common digital audio workstations (DAWs) with the development of digital technology (Figure 1.2)[1][2]. It means that since 1950s, the mainstream AMI has used the CSP and has little changed in its design, both in hardware and software. However, when users use the CSP mixer, the actual behavior conflicts with the psychoacoustics (it is elaborated in Chapter 2). Hence CSP is not the best visual metaphor for AMI design and stage metaphor could be a potential alternative.

The stage metaphor is first proposed by David Gibson [3]. It is based on the concept of "deep mixing" to create a virtual stage that presents each audio channel through colored spheres. The user adjusts the volume, pan and equalizer by changing the position of the spheres in three-dimension space (Figure 1.1 Right). This visualization of the mix is more easily understood and perceived by the user. The stage metaphor is considered to be a viable alternative to CSP [4]. However, in recorded music where a large number of audio channels need to be mixed and processed, using AMI with the stage metaphor would display many spheres so that the interface would become cluttered and difficult to manipulate [5,6]. It is probably why it has not been widely adopted in the professional mixing field. Thus we wonder if there are appropriate mixing scenarios to apply the stage metaphor.

1 Figure 1.2 Left from Steve Harvey. `https://unsplash.com/photos/Q4nTOXmZDFA/`

2 Figure 1.2 Right from Apple. `https://www.apple.com/logic-pro/`

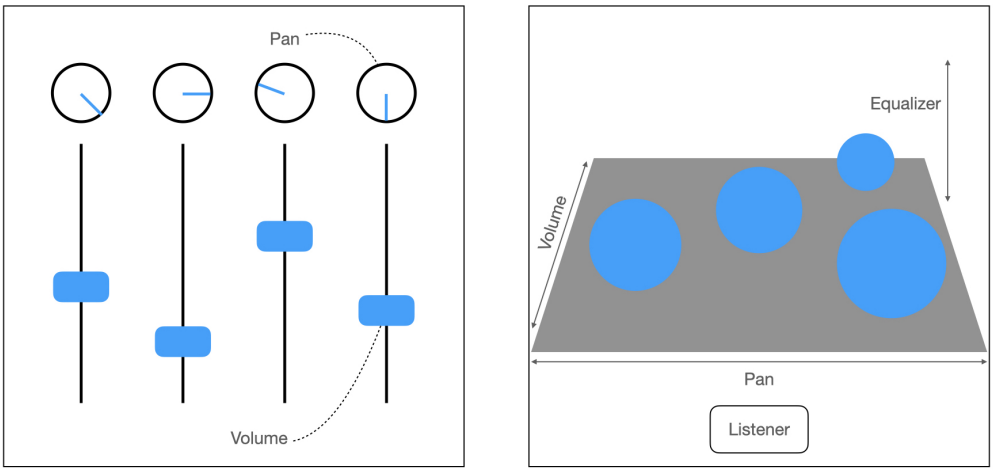


Figure 1.1 Channel Strip Metaphor and Stage Metaphor

Figure 1.2 Physical and Digital Mixing Consoles (Credit: Steve Harvey and Logic Pro from Apple)

1.2. Music Production using Virtual Reality

The possibility of producing music or mixing in virtual space has been practiced along with the development of virtual reality (VR) providing it with immersive, distributed, and collaborative advantages. A study has shown that audio quality and performance are well maintained when audio is extracted and processed in VR, compared to music production using DAW. And the virtual music space consists mainly of timbres, sound parts, and depth space [7].

The use of VR could enhance the auditory and visual experience. Rivas Méndez et al. use the spatial audio recording techniques for capturing live music for reproduction in a six-degree of freedom (6DOF) VR framework, which gives the listeners the ability to move close to sound sources with a high degree of plausibility to match the visuals [8]. DigDrum is a physical drum augmented by VR, the virtual visuals enhance the possibilities of musical expression [9]. LeMo is a two-person collaborative virtual reality sequencer that allows users to communicate about music production through visual representations rather than spoken communication, an approach that expands the interaction of virtual AMI [10].

While consumer market VR headsets mainly provide users with an immersive gaming experience, there are many applications that are similar to DAW or virtual AMI. For example, Virtuoso[3] and Korg Gadget VR[4] simulate a music studio space, provide fully virtual recording, instrumentation, and mixing. In contrast, The Music Room[5] and DearVR[6] allow virtual instruments to be used as MIDI controllers by accessing DAW connections and using the position of space to visually mix individual channels. This type of workflow tends to be more professionally productive. However, the virtual environment is separate from reality. The goal for the musician is to deliver high-quality music, but the results could be diverse

3 Virtuoso: `https://www.vrmusic.site/`

4 Korg Gadget VR: `https://youtu.be/DTUtKwIa3io/`

5 The Music Room: `https://store.steampowered.com/app/431030/The_Music_Room/`

6 DearVR: `https://www.dear-reality.com/`

in different environments. The results produced in VR might not be the most satisfying for the musician because it is not an accurate reflection of reality.

1.3. Stage Monitor Mixing

Stage monitor serves performers by enabling them to hear in real-time the status of themselves and other members so that the performance could be completed successfully. Thus, it is important to prepare the stage monitor mixing before performance.

The most classic stage monitor is the wedge monitors, also known as floor monitors. The wedge monitors (loudspeakers) are fixed around the stage floor and facing the performer, depending on the position of the performers and the microphones (Figure 1.3)[7]. The sound engineer sends the required audio signal into the wedge monitor, allowing the performer to get an accurate signal. However, the wedge monitors have some drawbacks. If the wedge monitors are not set up properly, performers are likely to be disturbed by each other's monitor. Also, the sound from wedge monitors would be "collected" by stage microphone, once it is too loud, not only there is risk of whistling but also affect the sound quality from the main amplifiers. Finally, part of the performers on the stage is not fixed, which means that the monitor sound they hear is always changing. In order to solve this problem, the in-ear monitors become the second option for stage monitors.

In-ear monitors consist of three components: a transmitter, a receiver, and a pair of in-ear headphones. The transmitter sends the mix to a wireless receiver worn around the waist, the receiver has a channel selector and earphone jack, and performers can hear the mix with earphones connected to the receiver (Figure 1.4)[8]. All performers can have their own individual monitor mix through the in-ear monitors. And with the advantage of wireless, performers can move around the stage more. However, with in-ear monitors as a standalone system, the number of performers on stage requires a corresponding number of in-ear monitors,

7 Figure 1.3 from Shure. `https://www.shure.es/musicos/descubre/contenido-didactico/why-use-in-ear-monitor-systems/`

8 Figure 1.4 from Cymatics. `https://cymatics.fm/blogs/production/in-ear-monitors/`

Figure 1.3 Wedge Monitors (Credit: Shure)

which is a high financial cost for some small venues. A system called Giggler aims to solve this problem by allowing performers to use the iOS mixing application to mix monitors for all channels. Moreover, Giggler improved the efficiency of mixing compared to traditional mixers [11]. However, the performers have to use Giggler while playing an instrument, and this switching of the interface between the instrument and the phone screen may interfere with the performers' ongoing task, so the user experience is to be considered.

Finally, regardless of the type of stage monitor used, it is important to communicate clearly with the sound engineer. Not all people know how to communicate with sound engineers [12]. An example of poor communication: Some performers think that the main amplifier is too loud and the wedge monitor's levels are too low so that they cannot hear the monitor mix. It is actually due to the wedge monitors being set up incorrectly. The performers may ask the sound engineer to simply increase the volume, thus destroying the sound-field. This kind of misunderstanding leads us to think about the feasibility of allowing musicians to self monitor their mixing with the Giggler solution, or if it is even realistically possible.

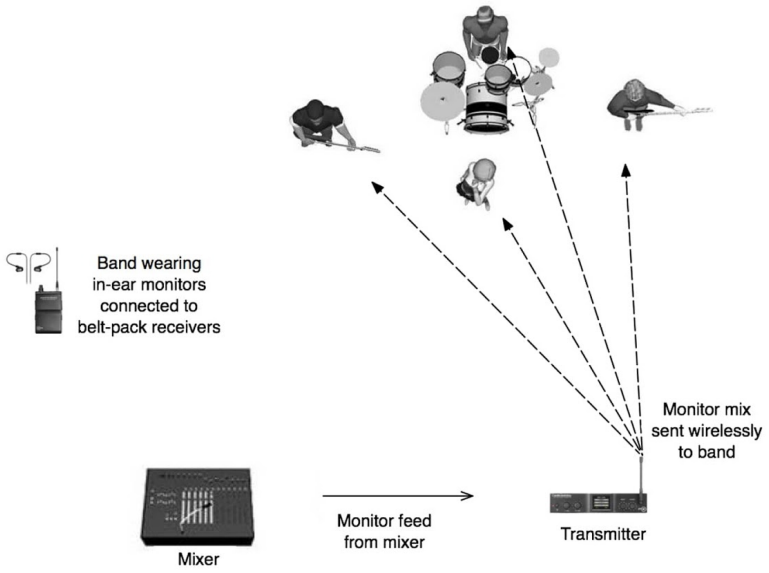


Figure 1.4 In-ear Monitors (Credit: Cymatics)

1.4. Objectives

As the stage metaphor is not widely applied and there are various drawbacks of current stage monitoring system, the main goal of this research is to design a system for musicians to intuitively and efficiently do self monitor mixing on stage, with an audio mixing interface that is mainly designed using the stage metaphor. Based on this concept, we design and develop ARMixer. When musicians with AR HMD face the other members on stage, leveraging augmented reality, the virtual AMI of the corresponding channel is displayed and it can be mixed by gestures. Additionally, the evaluation of this research concentrates on the usability of ARMixer, especially its performance in terms of intuitiveness and efficiency. Therefore, the design research is trying to address the following questions:

1. How intuitive is the ARMixer interactive experience?
2. How efficient is it to use ARMixer to do stage monitoring mixing compared to using a CSP Mixer?

Note that ARMixer is not a replacement for the sound engineers, because they

have responsibilities apart from monitor mixing in a live performance. ARMixer is designed to propose a novel interactive approach to stage monitor mixing for musicians.

1.5. Thesis Structure

This thesis is structured as follows:

- Chapter 1 describes the role of audio mixing in music production, the 2 metaphors for audio mixing interfaces, examples of music production using VR, and 2 existing stage monitor mixing systems. We propose a system for stage monitor mixing leveraging augmented reality based on the lack of appropriate application for stage metaphor interface and the various drawbacks of the current stage monitoring system, and state the research goals and questions.

- Chapter 2 examines the prior works related to this research. We compare the advantages and disadvantages of channel strip metaphor and stage metaphor and find that the intuitiveness of the stage metaphor is compatible with our research goal. We investigate gestural interaction systems and XR systems based on the audio mixing interface with stage metaphor and analyze their potential problems to present the value of this research.

- Chapter 3 describes insights from the user interview into the problems with the stage monitor mixing and proposes a concept of user experience scenarios and a system architecture to address each problem point. We answer the following questions: Why use AR, why is the stage metaphor appropriate, and why use gesture interaction. To validate the concept we present the ARMixer prototype and introduce its interface elements and design process.

- Chapter 4 is based on the design concepts and elements presented in Chapter 3 and describes the prototype goals and implementation process. We focus on the hardware solution and software processing of the ARMixer. Also in accordance with the objectives of this research, we cover the ARMixer usability testing and discussion of the results.

- Chapter 5 summarizes the main discussion of this thesis including research objectives, insights gained from literature review, design concepts, and prototype implementation. Finally, We describe the final conclusions, limitations, and future work.

Chapter 2
Literature Review

In this chapter, we focus on previous research related to the "stage metaphor". We compare stage metaphor and channel strip metaphor. And describe the stage metaphor interfaces applied in gestural interaction systems and XR systems. Finally, we present the innovative value of this research.

2.1. Audio Mixing Interface: Channel Strip Metaphor vs Stage Metaphor

The physical CSP mixer is a conventional metaphor in DAWs. One report on the metaphorical analysis of DAWs compared the interface design of Reason[1] and Live[2] [13]. Both used the CSP mixer metaphor of knobs and faders to help users with experience with electronic music hardware make a seamless transition to using a DAW, but as music software becomes more powerful and common, it is a challenge to make it easy for new users who do not use hardware to use music software. Besides this, the authors suggested that only using the metaphor of physical hardware would retain its drawbacks (e.g. the issue of clutter with multiple audio cables) and lose the flexible nature of the software interface. The music interface needs to reach a balance of realism and abstraction.

Previous studies have explained the problems of channel strip metaphor in terms of user experience and psychoacoustics. The CSP mixers do not represent the position of the sound source in the image. For instance, a channel that would be panned to the right may be at the left on the physical CSP mixer. It forces the

1 Reason: `https://www.reasonstudios.com/en/reason/`

2 Live: `https://www.ableton.com/en/live/`

user to observe the position of each channel knob scale. And when the user uses the faders to control the volume, the position of the faders is actually inversely proportional to apparent depth. When the fader is closest to the user, the sound source is perceived to be farthest away, and this inverse relationship could give the user a sense of dissonance [14]. In addition, CSP could also inhibit the engagement and flow of the mixing process by splitting into separate channels, which overload the user's working memory and reduce aural acuity [4].

A quantitative study comparing the channel strip metaphor and the stage metaphor in terms of speed and accuracy showed that the stage metaphor did not show a significant difference from the channel strip metaphor. The stage metaphor was preferred by subjects for its intuitiveness, innovation, and visual feedback on spatial perception [15]. Another investigation explored the usability of 6 AMIs with the stage metaphor and showed that visually simple interfaces were efficient and reduced the cognitive load. However, it is essential that they prevent the interface from being cluttered when mixing multiple audio channels [6]. Nevertheless, in both studies, the stage metaphor interfaces are in the two-dimensional planes. The concept of the stage metaphor originates from three-dimensional space, and the two-dimensional stage metaphor interface is more limited than the three-dimensional one, so the usability of the stage metaphor interface in three-dimensional space is worth exploring.

2.2. Audio Mixing Interface with Gestural Interaction

The use of gestures to convey information is an important approach to human communication. And the study of gestural interaction is a key element of human-computer interaction (HCI). Examples of using screen-based multi-touch gestures to control AMI on a two-dimensional plane are described in the section above, and the following discussion focuses on the three-dimensional input of fingers and hand position in space.

Thus far, a large quantity of research explored the feasibility of gestural interaction in mid-air on various AMI interfaces. Drossos et al. [16] utilized the gestures of an orchestra conductor for real-time multi-track mixing, and qualitative re-

search showed that the proposed system achieves a excellent user experience and improved artistic expression potential for musicians. Lech and Kostek [17] experimented with a system consisting of a PC, a webcam, a projector and a screen for image projection to demonstrate that using dynamic gesture mixing was more efficient than static gestures.

Additionally, there were examples of using gestures to control the AMI with the stage metaphor. Ratcliffe's research [18] explored the idea of the stage metaphor interface controlled by gestures and developed MotionMix, a mixing system using dynamic gesture interaction. In pilot test, it showed that the system with the stage metaphor was not significantly less accurate than the CSP mixer. Subjects took longer to mix the system without visualization than the MotionMix with visualization, demonstrating that visualization of the stage metaphor could enhance the efficiency of mixing. Wakefield [19] extended the MotionMix interface and developed LAMI which additionally covers equalizer and muting/soloing. While some subjects thought the gesture and interface experience of LAMI was interesting, expert users preferred to use a traditional DAW for mixing. Therefore, according to those two contrasting results, we wonder if more systems using gestures to control the AMI with the stage metaphor need to be evaluated for their performance and usability.

All of the above studies demonstrate the accessibility of using gestures to mix. However, the visual outputs of their proposed systems are two-dimensional such as projectors and monitor screens, so the user's gestural input and visual perceptual output are not in the same environment, similar to the interaction between a keyboard and a screen. Although this interaction is trainable, we believe that integrating the user's input and output in the same space would create a more immersive experience.

2.3. Audio Mixing Interface in XR

XR is an emerging term for all the immersive technologies such as VR and AR. There's a massive productivity gain in letting users spatially organize their multitasking [20]. XR offers the opportunity to have a vast canvas on this purpose.

VESPERS is an audio mixing system that combines stage metaphor and VR. It

utilizes a 24.2 multi-channel speaker array, a VR headset, and VR controller, and represents the sound source as a sphere with conforming stage metaphor. Users can perform audio mixing tasks in a virtual environment based on the position of the sound source. And the real-time mixing output relies on high-density speaker arrays to provide users with immersive audio feedback [21].

Furthermore, the most commonly accepted definition of AR contains three conditions: 1. Combines real and virtual; 2. Interactive in real-time; 3. Registered in 3D [22]. And there were attempts to study the use of AR technology for sound processing. Miyagawa et al. [23] explored the potential of using spatial mapping for music appreciation, placing audio sources in various locations within a physical space and outputting corresponding spatial audio effects. Although it was aimed at music appreciation rather than audio mixing, it still uses the concept of deep mixing. It also uses spheres as a metaphor for the visualization of the audio source. Users commented that it helped them distinguish the instruments in the music and that the dynamics of the spheres were interesting. But sometimes it is tiring to look at because the spheres were too far away. Also recently Bauer and Bouchara [24] proposed an AR composition platform that creates interactive musical experiences. And he emphasized that the visuals and embodied interaction possibilities modified the auditory perception of various mixing elements. Due to the COVID-19 pandemic, this project was temporarily not getting results from user testing. However, AR design needs to be embedded in natural contexts and interact with reality. Both studies use the position in space to implement AR concepts, but these elements are common many places, and therefore do not require a specific space to experience it. Consequently, this application design is not closely connected to data derived from reality.

2.4. Summary

Overview, prior studies show that there is no significant difference in performance between the stage metaphor and channel stripe metaphor. And stage metaphor can provide psychoacoustic intuitiveness when the users perform the audio mixing. But those studies cannot be considered conclusive because regardless the interactive systems are screen-based or gesture-based, the visualizations of the stage

metaphor interface are presented in a two-dimensional plane. The concept of the stage metaphor originates from deep mixing in three-dimensional stage space, and as we all know, three-dimensional space is more potentially extensible than two-dimensional space. Thus, this thesis addresses the further question of what is the user's experience with the stage metaphor in three-dimensional space. Furthermore, the XR systems of audio processing related to the stage metaphor proposed in the past similarly prove its ease of use and enjoyable advantages. However, we believe that mixing through AR requires the connection to contextual awareness, otherwise, the real-world view element of AR is meaningless. Hence at the same time, it remains an open question as to what music scenario is appropriate with the stage metaphor.

It is important to note that previous research suggests the possibility that stage metaphors can be applied to stage monitor mixing [25]. Consequently, this study addresses the usability of the audio mixing interface with the stage metaphor using AR and gestures based in the stage monitor mixing scenario, so far lacking in the scientific literature.

Chapter 3
Concept Design

In this chapter, we describe the problems with stage monitor mixing that are defined from user interviews and propose a concept design and system architecture to solve the problem. Finally, we describe the interface design of the prototype ARMixer used to validate the concept.

3.1. Problem Definition

To explore the workflow and issues of stage monitor mixing for musicians, we invited 3 musicians to conduct a user interview. One of the interviewees had 14 years of experience as a band performer and he is also a music label manager. The other 2 interviewees each had 3 years of performing experience. We found there are several mixing issues in stage monitors.

The interviewees believe that there are various DAWs in the music production industry and musicians choose different DAWs depending on their workflow. The mainstream DAWs are based on CSP design for mixing. Even though musicians sometimes need to work collaboratively across platforms, it is a very low learning cost for them and they can easily adapt to the infrequently used DAW without any negative impact on their productivity. However, musicians in live performance scenarios are limited in the monitor mix.

Before musicians perform live, they need to discuss the setup and mixing effects with a sound engineer (usually from the organization of the venue). There are two main mixing tasks. One is mixing the sound from the main amplifiers, which is responsible for the audience's listening experience, and the other one is stage monitor mixing, which uses in-ear monitoring system to allow the musicians understand what the other members are doing, where they are in the song, and

ensure that the notes and rhythms they are playing fit in with the rest of the band. In general, the musicians discuss with the sound engineer about the mixing from the in-ear monitors based on sound preferences. Sometimes the problem of communication overload happens. For instance, imagine there is a four-piece band (Figure 3.1 Left). For a guitarist, their monitor mix needs may be to reduce the volume of the vocal and increase the volume of the drums in order to focus on the rhythm of the song. The sound engineer works with each musician to perfect their monitor mix. It requires long conversations about specific parameters. In addition, each musician has 4 audio channels in their monitor system, but for the sound engineer, completing all the monitor mixes means processing 16 channels (Figure 3.1 Right). If the drum kit is split into snare, drum, hi-hat, etc., or other synthesizers are added, the number of channels processed increases. The interviewees explained that they mainly play in small venues where the organizer does not provide the mixing console with a sufficient number of channels. So they usually use wedge monitors and their monitor mixes are the same in the general live performance stage. Only large venue systems can support highly customized mixing demands, such as open-air concerts, stadiums, and professional concert halls.

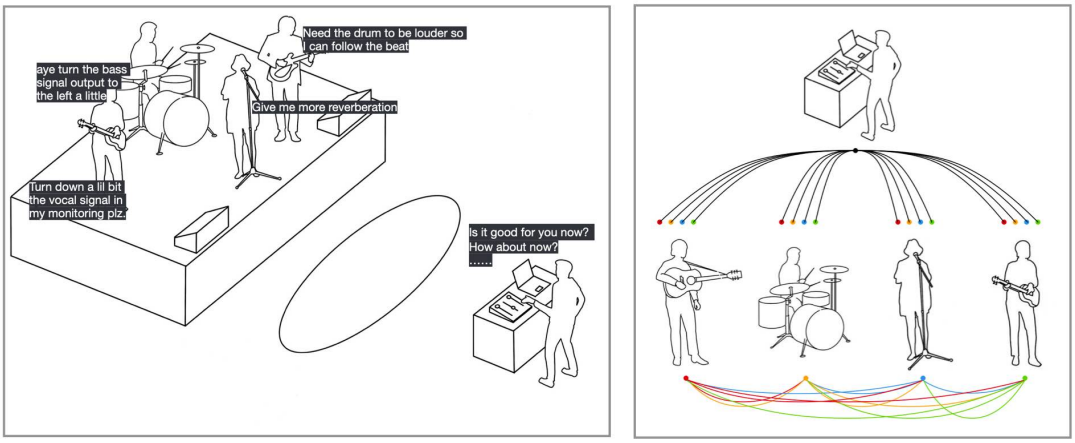


Figure 3.1 Musicians' customized stage monitor mixing demands and sound engineer needs to process multi-track mixes

Also the interviewees complained about the problems in the music practice

room. For amateur bands, it is difficult for the members to get together in the practice room because they all have formal jobs. If an individual member wants to practice as a band, the general solution is to digitalize the other channels and use a PC for the audio output. Therefore they often use a PC with DAW or a physical mixer for mixing. However, they realized that this approach produced a bad user experience because it is easy to get distracted by constantly switching between the instruments and the mixer or PC during the process.

In a nutshell, musicians encounter the excessive communication costs, the limitations of mixing devices that do not support sufficient audio channels, and problems with frequent switching between multiple devices or interfaces. In the next subsections, we describe how augmented reality can be used to solve these problems.

3.2. User Experience Scenario

We propose the stage monitor mixing system leveraging augmented reality for target user musicians to accommodate their demand of efficiency, customized mixing and avoidance of multiple interface switching. We learn from user interviews that musicians actually have basic mixing knowledge, and they are capable of mixing their own monitors. And only the musicians themselves know whether their monitor mixing are adequate for live performances. Therefore, we suggest to transfer the task of monitor mixing directly to the musicians themselves, solving the problem of excessive communication with the sound engineers. However, the area constraints of the stage make it difficult to have a mixing console for each musician, and the audience's visual experience could be compromised. Some designers, such as G. Krishna [26], believe that the best interface is no interface. People tend to try to solve all kinds of problems by adding screens or interfaces, but it actually significantly distract users from their ongoing task. In Figure 3.1, it is also explained that each performer can actually be seen as a single channel. Thus, we propose an augmented reality system that virtualizes the interface of the physical mixing console. For example, when Performer A is facing Performer B, Performer B is mapped with the channel of the corresponding Instrument B. The AMI from Instrument B channel is displayed from Performer A's point of

view and can be mixed intuitively. Likewise, when Performer A is facing another performer or instrument, or even their own instrument, he/she can mix all channels of the monitoring system (Figure 3.2). When a musician needs to process the monitor mixing, the AMI automatically appears in the stage space, instead of remaining fixed in a certain position in the space. AR device features spatial tracking and in-situ visualization. It means that the AR visualization can be accurately placed in the space. For example, placing the visualizer on the guitarist and the controls will be for the guitar channel. Placing the visualizer on the bass player, and the controls will be for the bass channel. On the other hand, with a fixed LCD screen, the user needs to look at the LCD screen and then observe the interface as there is no spatial sense. With AR, users can directly look at the musician to control. Through the natural contextual awareness interaction, users do not actively seek information [27] , therefore this AR system could be one of the paradigms of proactive computing [28].

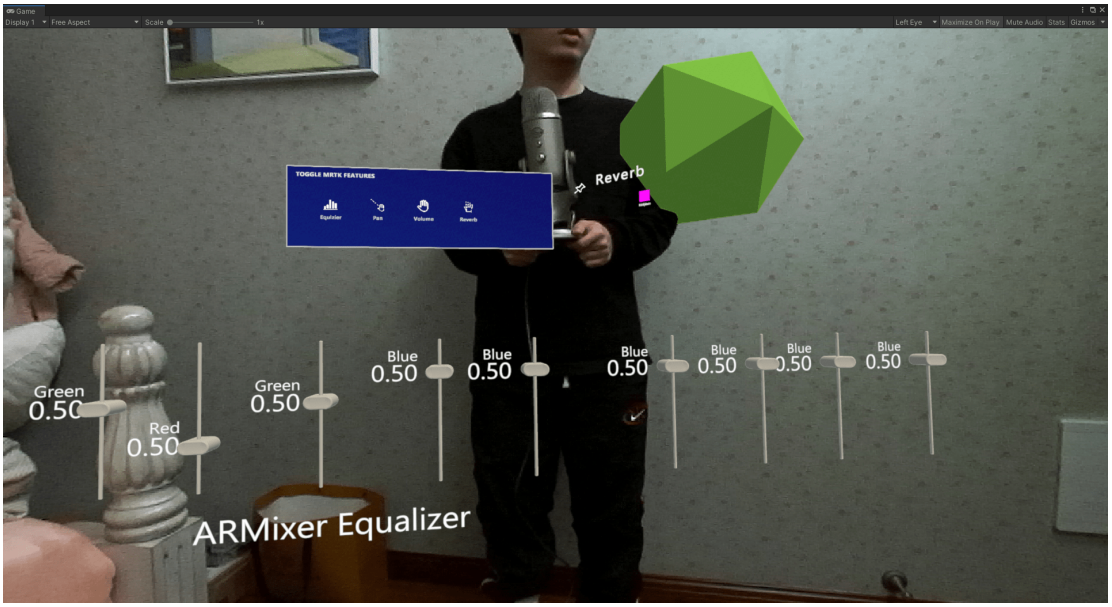


Figure 3.2 Low-fidelity Prototype: User faces the musician with the microphone and the mixing interface of vocal's channel is displayed

To provide simplified mixing interaction for musicians, we adopt the stage metaphor AMI and allow musicians to control it through gestures. The conventional mixing devices can perform the required signal processing, but the in-

teraction is difficult to adapt because different parameters are in different interfaces. Hence the AMI should focus on the natural gestures and movements that musicians might make [29]. Moreover, the stage metaphor is designed to better visualize the absolute and relative spatial distribution of audio channels. But if there are many audio channels to mix in a traditional mixing scenario, the stage metaphor interface becomes cluttered [19]. However, as described in the target scenario, the musician selects the mixing target by orientation, and the displayed AMI through AR represents only a single audio channel. Therefore, there cannot be more than one channel per mix, which means the interface is simple and the stage metaphor is more applicable to this scenario. Lastly, since music performers are on stage with instruments, it is difficult to require them to hold input devices such as hand controller to interact with the AMI at the same time. Using gestures is the natural way of input in this context.

Overall, as shown in Figure 3.3, when the band is preparing the stage monitor mixing before the performance, each musician with an instrument can be treated as a single audio channel (a). The musician with AR HMD faces one of the bandmates (b), the monitor mixing interface for the corresponding audio channel is displayed. The blue sphere represents an audio channel, and the musician can gesture to move the position of the sphere in space to accommodate the effect of monitor mixing (c). Also, he/she can mix the other channels when turning to the other members (d).

3.3. System Architecture

Based on the above user experience scenario, we elaborate its system architecture from two components including display output and audio output.

As shown in Figure 3.4, one of the most promising technologies for AR is optical see-through HMD, which uses a half-transparent mirror that is placed in front of the user's eyes, allowing user see lossless field of view (FOV) [30]. The multiple IR cameras are used for compatibility with people/object recognition, eye-gaze tracking, and hand tracking. Object recognition is used to determine which instrument's channel needs to be processed, with the aid of eye gaze tracking to increase its accuracy [31]. After detection, the corresponding virtual AMI

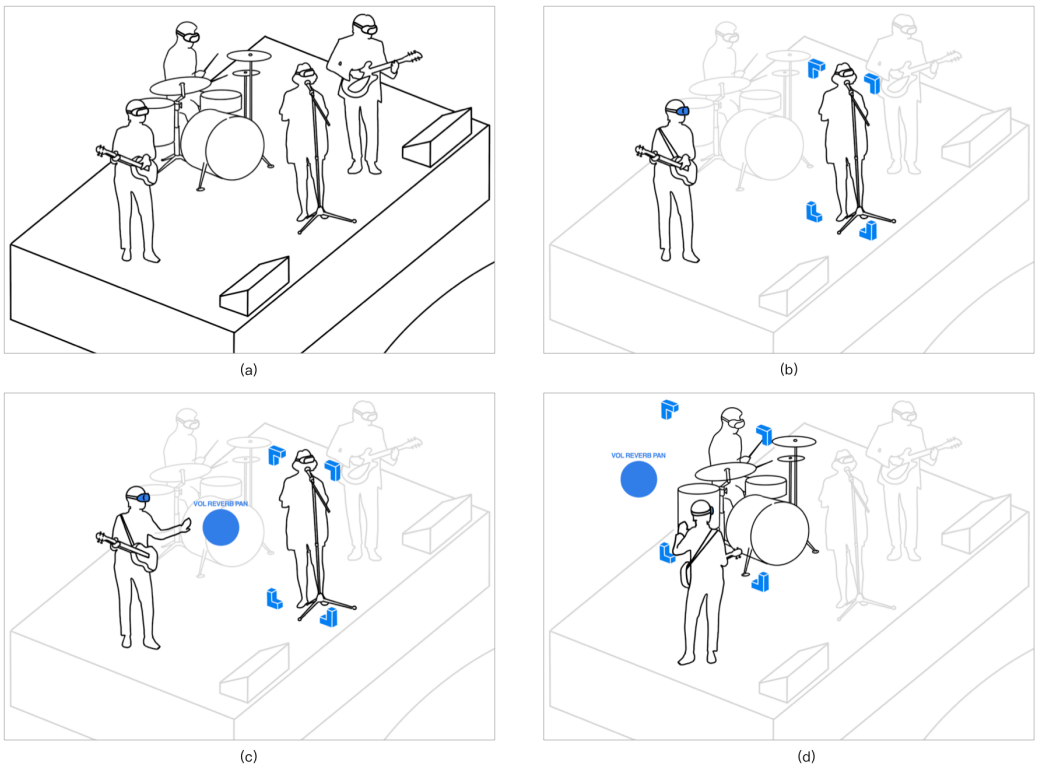


Figure 3.3 Storyboard: Stage Monitor Mix leveraging Augmented Reality

is placed in the real world.

At the same time, the system has a receiver to collect all audio signals for mixing. Hand tracking is used to implement various gestures in mid-air to monitor mix and eventually output to the user's stage monitor headphones.

We suggest that the hardware of this AR system is the lightweight AR glasses. The user would not be distracted by it, and it would not interfere with the performance when using it on stage.

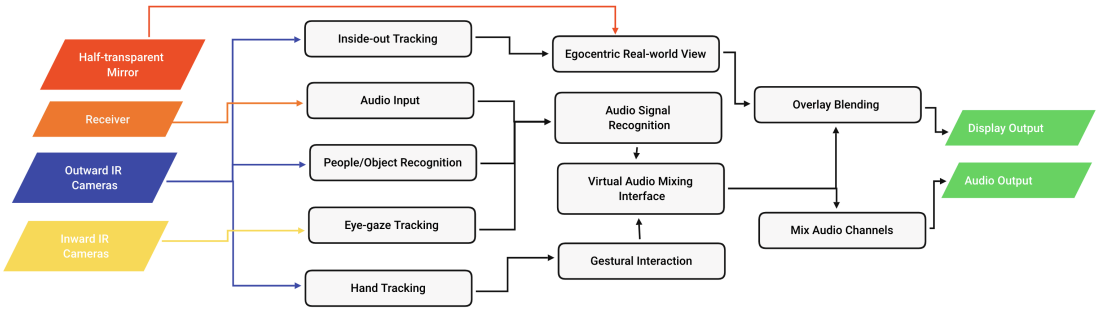


Figure 3.4 The System Workflow

3.4. ARMixer

To validate whether our proposed concept addresses the research objectives, we design the prototype named ARMixer[1]. In this section, we describe ARMixer's interface elements with stage metaphor and its gestural interaction.

In user interviews, we learned that volume, pan, reverb and equalizer are the most common mixing parameters used by musicians in stage monitor mixing. It is also confirmed that the visualization of the equalizer is one of the most needed parameters for future AMI design [32]. To avoid conflicts in gesture interaction, volume, pan and reverb are placed in the same interface layer. And as discussed in Section 3.2, the virtual AMI is displayed when the mix target is recognized. Following the stage metaphor paradigm, the single audio channel is represented by a blue sphere. In addition, the equalizer control is arranged in another interface

1 Introduction video is available in Vimeo `https://vimeo.com/574414567`

layer.

Volume Control

When the user holds their right hand out with the palm facing upwards and moves it up and down along the y-axis in space ↕, the audio volume increases and decreases accordingly. When the volume increases, the opacity of the blue sphere increases at the same rate. Similarly, it becomes more transparent when the volume decreases (Figure 3.5). In the previous study, the AMI with stage metaphor for volume control used the z-axis motion (Figure 3.6)[2] because the volume should be lower when the sphere is farther away from the user, and higher when the sphere is closer to the user [3]. Although this design leverages the relationship between the position of spatial audio and the user's psychological perception, it is difficult to understand for users unfamiliar with spatial audio concepts. ARMixer's volume gesture control is inspired by the conductor in a symphony. As shown in figure 3.7[3], generally when the conductor raises their hand, the orchestra and other sound effects will be very strong and loud, and this everyday gesture metaphor of raising one's hand could be more in line with the user's mental model.

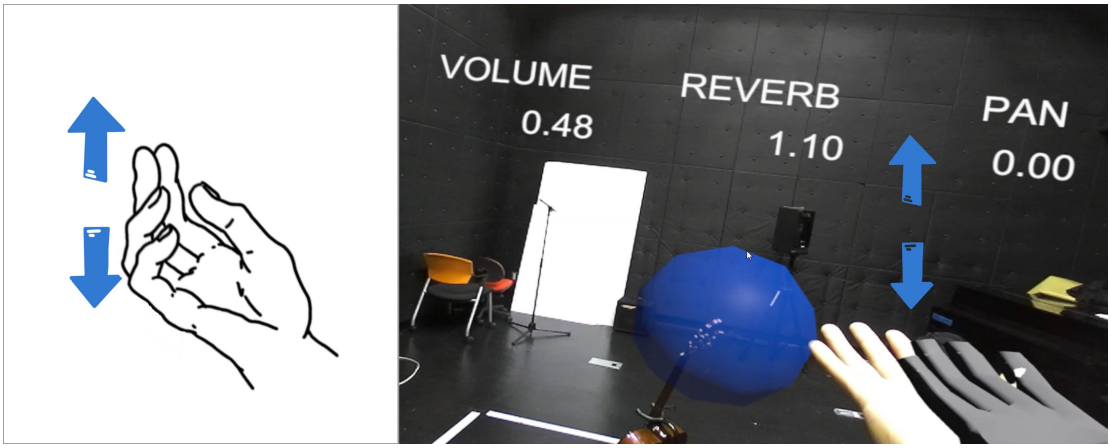


Figure 3.5 Volume Control

2 Figure 3.6 from "The Art of Mixing: A Visual Guide to Recording" by David Gibson

3 Figure 3.7 from Kazu Ota. `https://unsplash.com/photos/ohXrVLI1MLw/`

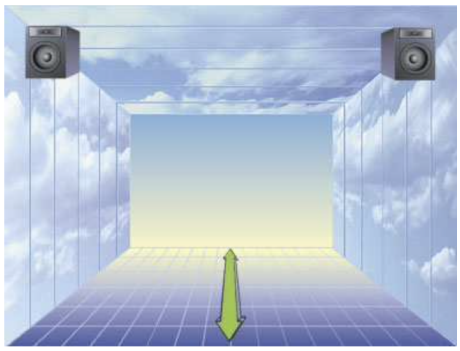


Figure 3.6 Volume: Front-to-Back Placement (Credit: David Gibson)

Figure 3.7 Conductor's Gesture (Credit: Kazu Ota)

Pan Control

Pan control is used to "move" or pan the apparent position of a single sound channel between two outputs, usually "left" and "right" for stereo output. As the user keeps their left palm facing downwards and moves it left or right along the x-axis in space in front of them ↔, the audio source pans left or right accordingly (Figure 3.8). At the same time, the blue sphere moves along the x-axis to provide a visual cue for controlling the panning.

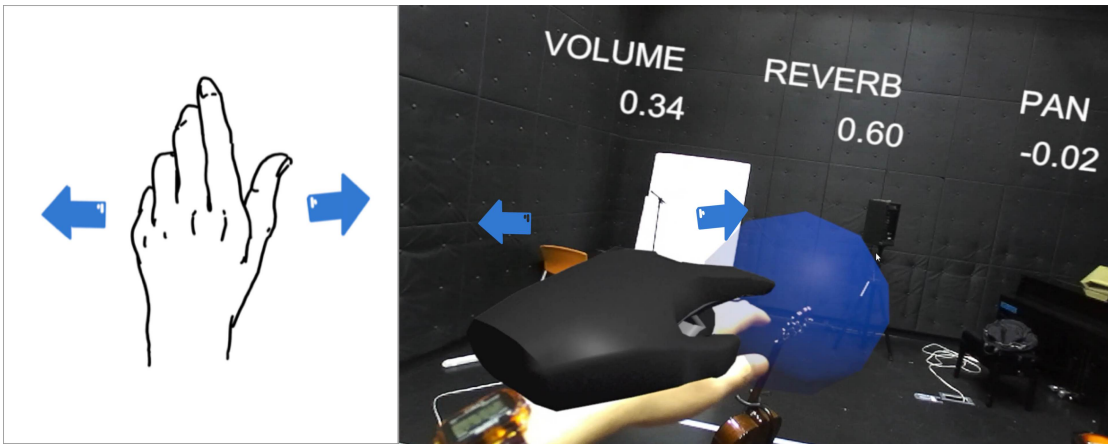


Figure 3.8 Pan Control

Reverb Control

Reverb or reverberation is created when a sound is reflected causing numerous

reflections to build up and then decay as the sound is absorbed by the surfaces of objects in the space. It is related to spatial sound effect and the size of the sphere represents the strength of the audio reverberation. Therefore, when the user zooms into the floating sphere, the reverb becomes stronger as the sphere gets bigger. The user can expand the sphere by pinching their two hands simultaneously (Figure 3.9). This design approach is used because the pinch gesture is stable and precise [33,34] and it provides the "tactile" sensation between between the thumb and index finger which enhances user perception of the accuracy of gesture interactions in mid-air.

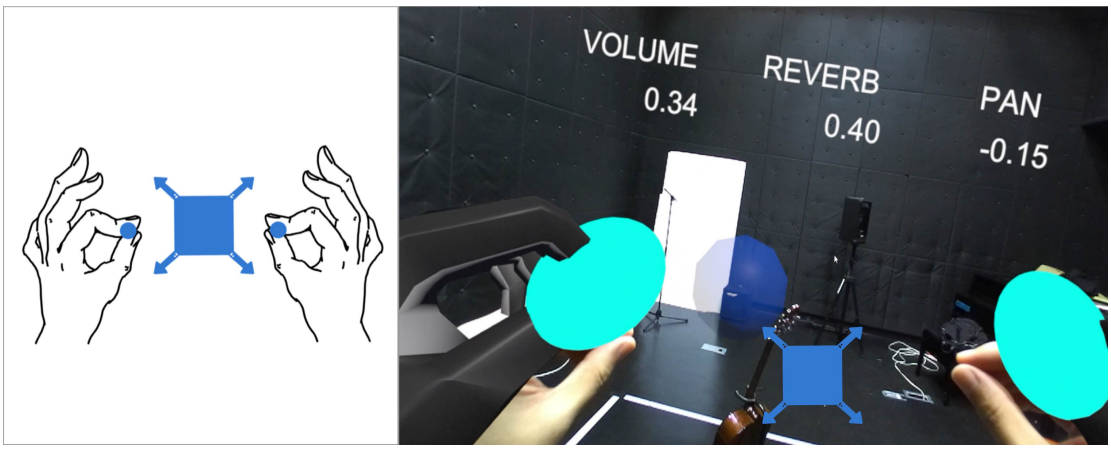


Figure 3.9 Reverb Control

Equalizer Control

Equalizer or EQ is the process of adjusting the balance between frequency components within an electronic signal. We provide high (12kHz), mid (2.5kHz) and low (80Hz) frequencies for the user to adjust with the equalizer. To improve the accuracy and to avoid conflicts in gesture recognition with one hand, we use two-handed collaboration to control the EQ balance. The user can select the fader by pinching with their right hand and moving up and down along the y-axis in space ↕, while opening and closing their left palm to switch the frequency (Figure 3.10).

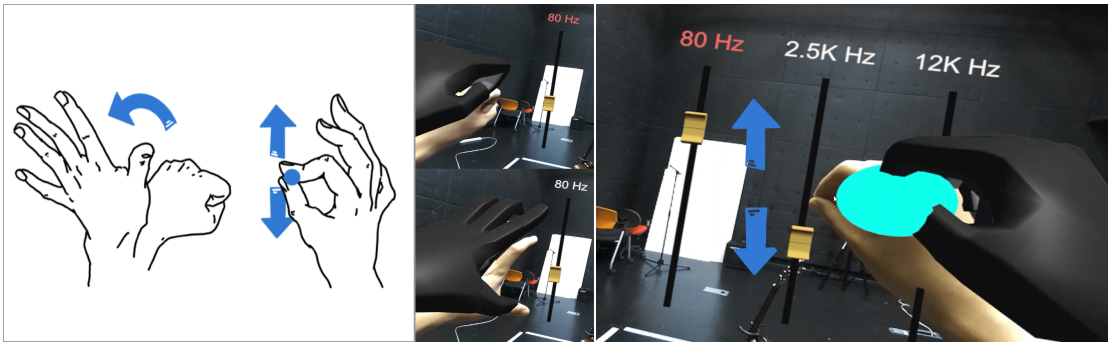


Figure 3.10 Equalizer Control

Chapter 4
Proof of Concept

In this chapter, we describe the prototype goals and implementation process. We focus on the hardware solution and software processing of the ARMixer. Also in accordance with the objectives of this research, we cover the ARMixer usability test and discussion of the results.

4.1. Prototype

4.1.1 Prototyping Goals

In summary, following the concept of leveraging AR to stage monitor mixes and the interface elements of ARMixer in Chapter 3, the ARMixer prototype needs to accomplish the following three goals:

- Have a virtual AMI with single audio channel placed in the real world, enabling intuitive AR interaction.
- The user is able to control the virtual AMI with the stage metaphor via gestures to adjust volume, pan, reverb, and equalizer mixing parameters.
- The system is compatible with real-time monitor mixing of audio input and output.

4.1.2 Implementation

The implementation of the system consists of hardware and software parts.

- Hardware: Video see-through AR head mounted display (HMD)

- Software: ARMixer's audio mixing interface and its gesture interaction

Hardware

The hardware system solution used in ARMixer is the video see-through head-mounted display (HMD). There are three fundamental techniques for displaying visuals in AR. The most prevalent method for overlaying virtual objects in real environments via smartphones and tablets is called handheld AR. However, considering the user experience of the ARMixer, it is difficult for the user to hold the instrument while holding another electronic device. Therefore, handheld AR is not used in our prototyping. The other two techniques are based on HMDs and are classified as optical see-through (OST) and video see-through (VST). For OST, the real world is seen through the slanted semi-transparent mirrors placed in front of the user's eyes (Figure 4.1 Left)[1]. These mirrors are also used to display the computer-generated images and thus combine virtual objects and real-world views [35]. The OST headsets currently in the market, such as Hololens 2 and Magic Leap (Figure 4.2 Right)[2] have a small field of view (FOV). Consequently, it is difficult for users to interact with virtual objects in a small rectangular area (Human FOV is around 150° × 120°; the FOV of Hololens 2 is 43° × 29°; the FOV of Magic Leap One is 40° × 40°.). Moreover, the use of mirrors and lenses reduces the brightness and contrast between virtual and real-world perceptions. The VST presents video feeds from cameras mounted on the head gear (Figure 4.1 Right)(e.g. Gear VR see Figure 4.2 Left), and computer-generated images are electronically combined with real-world video representations. The VST provides a larger see-through FOV than the OST. Although there is latency in matching video and computer-generated images in VST, it is possible to synchronize the latency of both for the natural AR interaction. We consequently chose the VST for prototyping. In the following paragraphs we describe the details of the combination of VST HMD.

As shown in Figure 4.3, we simulated the VST AR HMD by assembling the

1 Figure 4.1 from Niteesh Yadav. `https://blog.prototypr.io/understanding-display-techniques-in-augmented-reality-c258b911b5c9/`

2 Figure 4.2 from Niteesh Yadav. `https://blog.prototypr.io/understanding-display-techniques-in-augmented-reality-c258b911b5c9/`

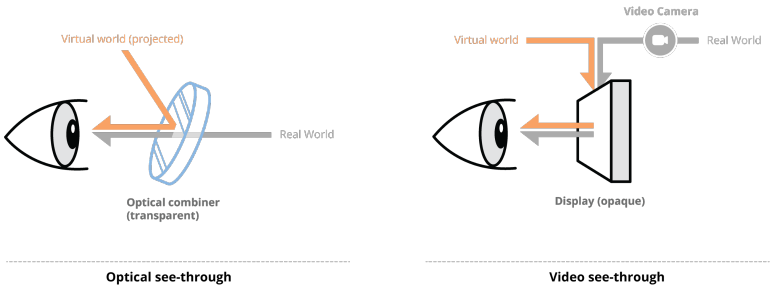


Figure 4.1 Tow Head-mouted Dsiplay Techniques in Augmented Reality (Credit: Niteesh Yadav)

Figure 4.2 Augmented Reality Headsets (Credit: Niteesh Yadav)

Oculus Quest 2[3] VR headset with the Zed Mini[4] stereo camera. The Leap Motion[5] hand controller was also attached in front of the VR headset for compatibility with natural gesture interaction. In addition, we used an audio interface (Focusrite Scarlett 2i2[6]) with a microphone or other instruments connected as the real-time audio input and output devices in the program. Finally, we used a PC with an Intel i7 processor, 16G of RAM, and an Nvidia GeForce GTX 1660Ti for computing. This setup (Figure 4.4) was able to accommodate the demands of ARMixer.

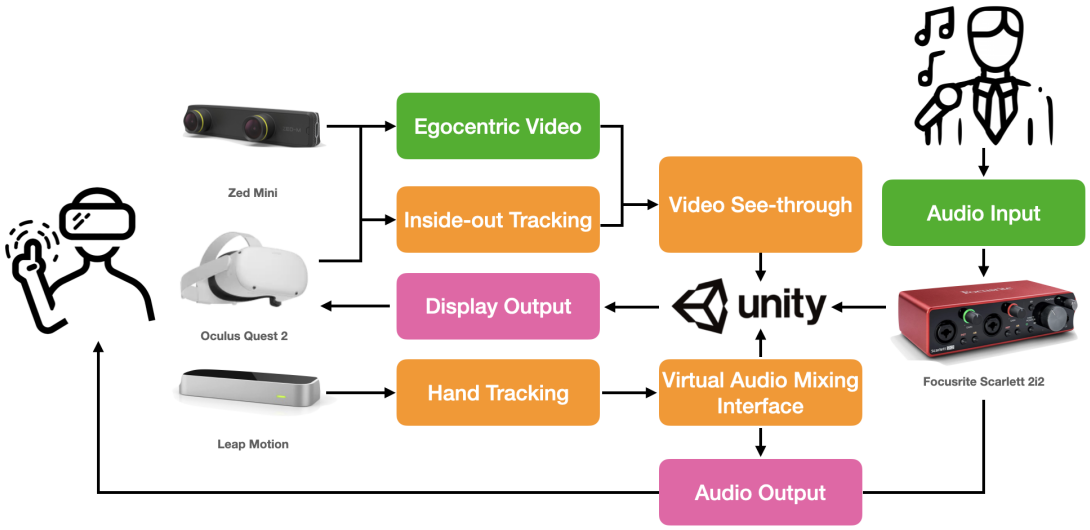


Figure 4.3 ARMixer System Overview

An Oculus Quest 2 was used to render and display the ARMixer's content. It is a standalone headset with six degrees of freedom (6DOF), tracking the movement of both head and body, then translating them into VR/AR with realistic precision. We chose the Quest 2 for two main reasons. First, Quest 2 uses four cameras for inside-out tracking and no external sensors are required. It allowed the virtual

3 Oculus Quest 2: `https://www.oculus.com/quest-2/`

4 Zed Mini: `https://www.stereolabs.com/zed-mini/`

5 Leap Motion: `https://www.ultraleap.com/product/leap-motion-controller/`

6 Focusrite Scarlett 2i2: `https://focusrite.com/en/usb-audio-interface/scarlett/scarlett-2i2/`

Figure 4.4 ARMixer Setup

objects in the program not to have position calibration issues due to tracking. And it saved time in prototyping and experimentation for equipment setup. Second, Quest 2 is compatible with PC VR streaming when connected over USB or Wi-Fi. It allowed us to preview ARMixer's content in real time.

The ZED Mini is a stereo camera that provides see-through video and real-time depth and environment mapping. When we attached a Zed Mini in front of the Quest 2, the Zed Mini became an egocentric video device that displayed the real-world view from the user's point of view. And with a FOV of up to 90° × 60°, the Zed Mini provided more space for ARMixer's interactive content.

Leap Motion is an optical hand tracking module that captures the hand movements with high accuracy. ARMixer interactions were all gesture-based and allowed users to manipulate the virtual AMI directly.

Software

The hardware provides the basic capabilities, and the core content of ARMixer is in the software. Unity is a cross-platform game engine that supports building VR/AR platforms. The ARMixer prototype was completely developed in Unity[7] (version 2019.4.26f1 LTS). The following are the resources and components used in this project.

- Virtual Reality SDKs[8] allows the virtual contents from Unity to be visible in the Quest 2 without any external plug-ins.
- Unity Audio Mixer[9] allows the user to mix multiple audio sources, apply effects, and perform mastering. We used the Audio Mixer to implement the editing of the ARMixer's 4 mixing parameters.
- Stereolabs ZED-Unity plugin[10] comes with a stereo camera prefab to imple-

7 Unity: `https://unity.com/`

8 Virtual Reality SDKs: `https://docs.unity3d.com/2019.2/Documentation/Manual/VROverview.html/`

9 Unity Audio Mixer: `https://docs.unity3d.com/2019.4/Documentation/Manual/AudioMixer.html/`

10 Stereolabs ZED-Unity plugin: `https://github.com/stereolabs/zed-unity/`

ment video see-through augmented reality with the Quest 2.

- Leap Motion Unity Modules[11] allows physical hand motion tracking and gesture interaction with the user interface.

Regarding the ARMixer software process, as shown in Figure 4.5, first we developed a VR version to implement the gesture interaction and interface elements described in Chapter 3. And we wrote custom scripts for the behavior of each gesture concerning the 4 mixing parameters volume, pan, reverb, and equalizer. After ensuring that the gesture interaction and mixing effects worked as demanded in the VR environment, the Zed Rig prefab was applied as the main camera in the Unity project to implement video see-through, and the main camera of Leap Rig prefab was disabled because Zed camera was used for the final image instead [36]. In addition, because the Leap Motion was attached below the Zed Mini, the virtual hand generated by Leap Motion did not match the user's physical hand perfectly. It may confuse the user and could negatively impact the user experience. To solve the problem, we created a child object named Leap Offset under the Zed Rig prefab, and the transformation position of the Leap Rig prefab needed to be followed with Leap Offset. Lastly, we manually adjusted the position of the Leap Offset to ensure that the virtual hand and the physical can overlap in the AR environment. Through this method, we transformed a completely virtual world into an AR world where virtual objects and reality were mixed.

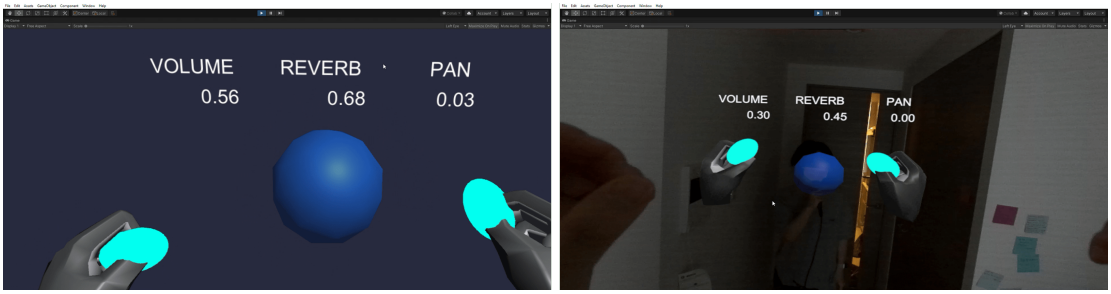


Figure 4.5 Software Processing: ARmixer in VR and Video See-through AR

11 Leap Motion Unity Module: `https://github.com/leapmotion/UnityModules/`

4.2. Usability Test

4.2.1 Purpose of Testing

To evaluate whether ARMixer achieves the main goal of this research described in Chapter 1, the usability test needs to verify 3 hypotheses:

1. Users with ARMixer could intuitively recognize the mixing target.
2. ARMixer's gestural interactions and interface animations allow users to monitor mixes intuitively and easily.
3. Users could save time by using ARMixer to monitor mix a single audio channel compared to using a CSP Mixer.

We set up 2 experiments based on these hypotheses. In addition, user feedback provided guiding direction for ARMixer iterations.

4.2.2 Experiment Setup

We conducted 2 experiments and there were 11 participants in total. Experiment 1 was a quantitative study to verify hypothesis 2 and hypothesis 3. And Experiment 2 was a simulated live performance to verify hypothesis 1 and hypothesis 2.

Experiment 1
We recruited 7 musicians (3 males and 4 females) for this testing. Their ages ranged from 22 to 26 (mean 24.29, SD 1.38) years old, music experience ranged from 3 to 13 years, stage performance experience ranged from 3 to 6 years. 2 of them know the basics of mixing, 3 have mixing experience but don't use it often, and 2 have been mixing for the last 2 years. They have all experienced VR HMDs but all were through VR gaming.

Participants were given an introduction and spent 10 minutes experiencing and familiarizing themselves with the gestures of the ARMixer before the experiment. As shown in Figure 4.6, each participant was required to mix a vocal channel in a pre-recorded song using the ARMixer and the CSP Mixer separately, with no restrictions on the effect of the mix, but the second mix was required to sound close

to the first. The order in which the ARMixer and CSP Mixer were used alternated from test to test. The completion time results were recorded when participants indicated finishing the mix. Finally participants filled out a questionnaire and conducted user interviews. The control variables were the mixing task, mixing parameters (volume, pan, reverb and equalizer), audio output device, and sample rate. The vocal channel of the selected song[12] was emphasized, allowing the participants to hear the mixing effects easily when they were mixing.

Experiment 2

There were 4 participants in this experiment (3 males and 1 female). Their ages ranged from 24 to 29 (mean 25.75, SD 2.36) years old, music experience ranged from 7 to 10 years, stage performance experience ranged from 2 to 7 years. Their main roles in the band are guitarist, vocalist, keyboardist, and drummer. 2 of them know the basics of mixing, 1 has mixing experience but don't use it often, and 1 has been mixing for the last 2 years. They have all experienced VR/AR HMDs.

The experiment venue was in a music practice room. Participants were divided into two groups, making a two-piece band consisting of a vocalist and a guitarist. As in Experiment 1, each participant was given a 10-minute warm-up session. Each band was required to play a different song. As shown in Figure 4.7, the participants with the ARMixer were asked to perform a real-time stage monitor mix of another band member's channel, and the experiment ended when the participants were satisfied with the mix. Finally, participants completed a questionnaire and user interviews. There were 4 real-time stage monitor mixes including 2 vocal channel mixes and 2 guitar channel mixes.

4.3. Discussions

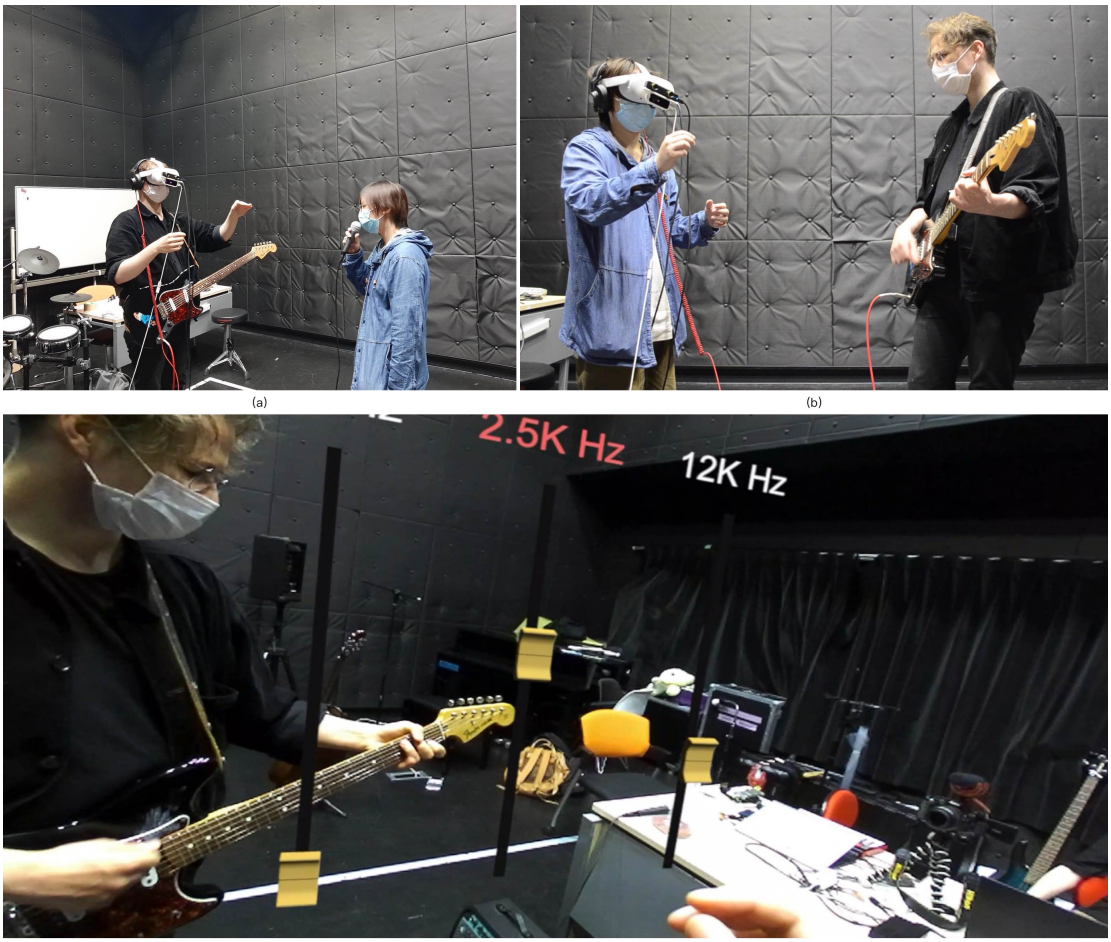
Usability

Usability is defined in the ISO 9241- 11:1998 as "the extent to which a product can

12 lamajet: `https://open.spotify.com/track/2f16dixJtp6DiUolB42cr1?si=eJJUD9APS4-9CuWnI1ZZ4g&dl_branch=1/`

a: Mixing by ARMixer; b: Mixing by CSP Mixer; c: ARMixer's egocentric view

Figure 4.6 Experiment Setup 1



a: Real-time vocal channel mixing; b: Real-time guitar channel mixing; c: ARMixer's egocentric view

Figure 4.7 Experiment Setup 2

be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [37]. We used the System Usability Scale (SUS) for evaluating the overall usability of the ARMixer experience [38]. SUS is composed of 10 statements that are scored on a 5-point scale of strength of agreement (Table 4.1). And the SUS scores ranges from 0 to 100, where higher scores indicate better usability.

I think that I would like to use ARMixer frequently.
I found the ARMixer unnecessarily complex.
I thought the ARMixer was easy to use.
I think that I would need the support of a technical person to be able to use ARMixer.
I found the various functions in ARMixer were well integrated.
I thought there was too much inconsistency in ARMixer.
I imagine that most people would learn to use ARMixer very quickly.
I found the ARMixer very cumbersome to use.
I felt very confident using the ARMixer.
I needed to learn a lot of things before I could get going with ARMixer.

Table 4.1 The 10 System Usability Scale Statements

As shown in 4.8[13], the average SUS score from Experiment 1 was 74.29 (SD 8.50). This meant that ARMixer is an acceptable product that is at least passable and close to "GOOD". We noted that the ARMixer's SUS scores decreased as a result of the strength of agreement with the following statement:

I think that I would need the support of a technical person to be able to use ARMixer.

4 participants agreed and 1 participant fully agreed with this statement. They thought that they would need to learn a new set of gestures to perform mixing tasks with ARMixer, compared to the CSP Mixer interface they were used to.

13 Figure 4.8 from 10up. `https://10up.com/blog/2018/testing-the-gutenberg-publishing-userflow/`

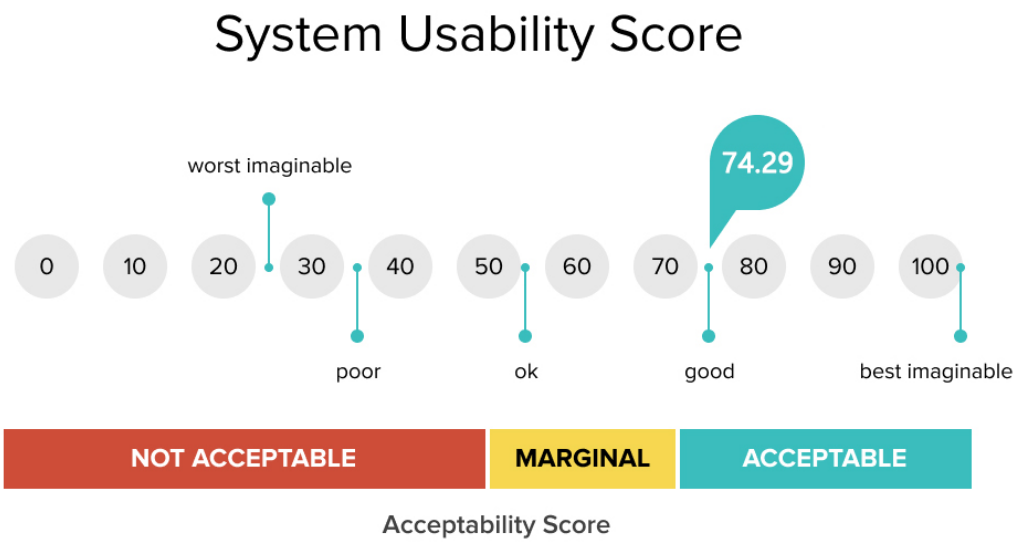


Figure 4.8 The Acceptability Result by SUS scores from Experiment 1 (Credit: 10up)

And because of the imperfections of ARMixer's technology, participants who had not tried AR HMD stated that they experienced motion sickness at the beginning and needed guidance from a technical person on how to reduce motion sickness manually. Thus the extra learning cost of ARMixer most negatively affects its SUS scores.

Interestingly, the average SUS score form Experiment 2 was 64.36 (SD 5.54)(Figure 4.9). 3 of the participants in Experiment 2 gave low SUS scores compared to the scores from Experiment 1. All scores were 62.5, 67.5, 57.5, and 70. Unlike the participants in Experiment 1, these 4 participants had to live perform with their instruments while mixing. They reported that although the mixing interface was intuitive, wearing a big and heavy HMD to play instruments distracted them from the live performance (Figure 4.10).

On the other hand, after understanding the design goals of ARMixer and experiencing it, most participants expressed a preference for using ARMixer. Moreover, they thought its gesture design had potential to be learned quickly by most people.

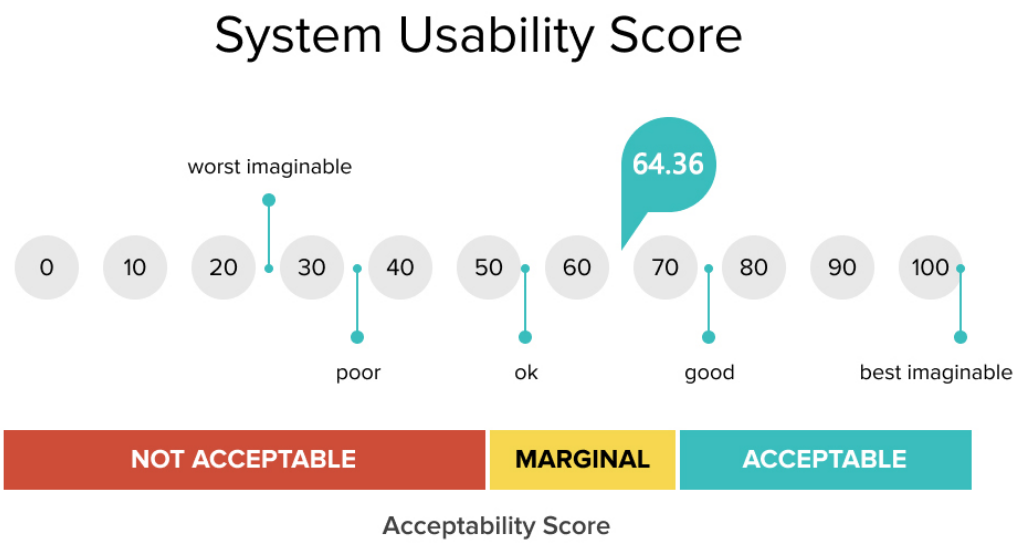


Figure 4.9 The Acceptability Result by SUS scores from Experiment 2

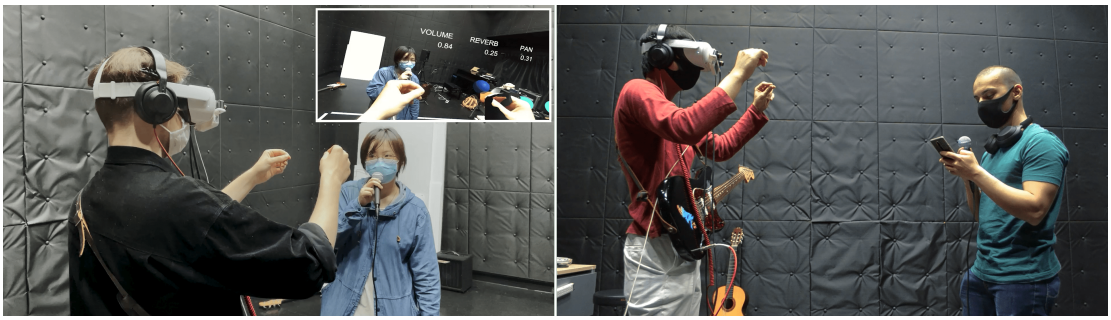


Figure 4.10 Participants with ARMixer performed real-time stage monitor mixing

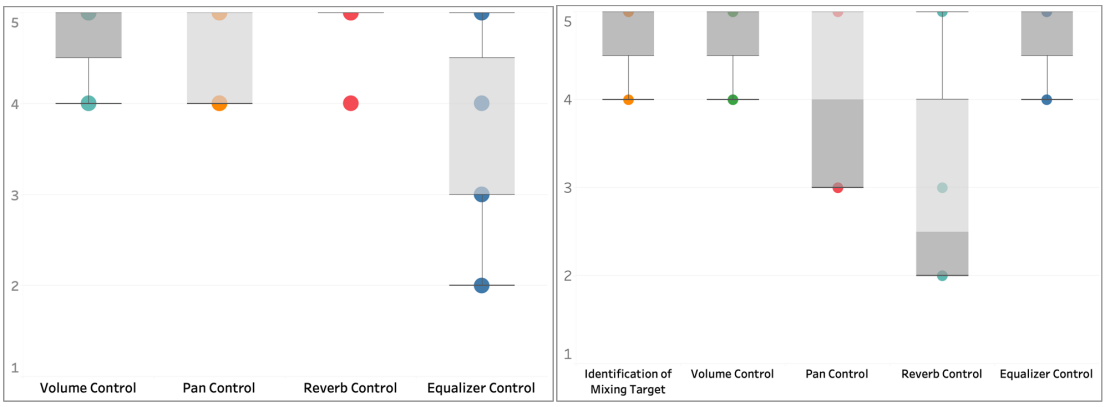
Intuitiveness

We evaluated the intuitiveness of the ARMixer experience in terms of psychology, which includes 5 criteria. There is 4 intuitiveness of the metaphorical relationship between gestural interaction and parameter control. The last one regarding the band member is mapped to the corresponding mixing channel when the user with ARMixer faces the member. In this discussion we refer to this criterion as the "identification of mixing target" and it is only available in Experiment 2. We used the 5-point Likert scale to quantitatively measure the 5 criteria, which scales from 1 to 5 points, with fully disagree (1), disagree (2), neutral (3), agree (4), and fully agree (5) [39].

As shown in Figure 4.11, the lower whiskers of the volume, pan, reverb controls were above 4 from Experiment 1, indicating strong agreements on all terms. Participants have expressed different opinions in the equalizer control, 1 of 7 commented that the visualization of the reverb control was not as intuitive and simple as the former three. Also, since the equalizer interface is close to the traditional CSP Mixer, some participants would be satisfied with the familiar interaction approach.

In Experiment 2, the lower whiskers of the volume, equalizer controls, and the identification of mixing target were above 4, suggesting that these three criteria had excellent intuitiveness. Pan control also performed well as it had a median of 4. The participants from Experiment 2 showed diverse opinions. We speculate that the level of user understanding of reverb may have a correlation with the perceived intuitiveness of the reverb control, as all participants who regularly use mixing production gave a score of 5, while those who only know the basics of mixing gave a score of 2 or 3.

In addition, from answers to the question of "The most satisfactory and unsatisfactory parameter controls", the reverb control got 6 votes and volume control got 4 votes in terms of most satisfaction, and equalizer control was the most unsatisfactory by 10 votes. Some participants claimed that the visualization of volume, pan, and reverb was simple and easy to understand, while the equalizer control required multi-step manipulation. In particular, reverb control provided visual representation of spatial sound effects. It also demonstrates the strength of the stage metaphor. Others thought that the pinch gesture for reverb was the



Left: Intuitiveness Performance in Experiment 1; Right: Intuitiveness Performance in Experiment 2

Figure 4.11 Intuitiveness Evaluation with the 5-Point Likert Scale

most accurate in ARMixer.

Therefore, although participants have expressed different preferences in the reverb control, it is indeed considered the most satisfying interaction. All medians are ranged from 3 to 4 in other intuitiveness evaluations. The results are promising that ARMixer provides an intuitive and simple stage monitor mixing experience and it is consistent with Hypothesis 1 and Hypothesis 2. Participants are satisfied with this "WhatYou See Is What You Get" AR experience.

Efficiency

This quantitative study (only available in Experiment 1) was designed to measure the time performance on ARMixer and CSP Mixer. 7 participants were given the assignments of mixing a single vocal channel on the same song using both mixing tools. Table 4.2 shows the mixing completion time for each participant. Almost every completion time using CSP Mixer was ranged from 60s to 130s. In contrast, the variability of the results using ARMixer was high. It is noteworthy that Participant 2 forgot the correct gesture control during the test resulting in an exceptionally high completion time, despite having been instructed to "warm-up" prior to the formal test. It indicates that the learning cost of gesture interaction affects the efficiency performance of ARMixer.

We performed a *t*-test on this study. As shown in Table 4.3, the average com-

	P1	P2	P3	P4	P5	P6	P7
CSP Mixer	64s	91s	119s	82s	131s	89s	68s
ARMixer	56s	250s	108s	166s	149s	159s	38s

Table 4.2 Raw Data of Completion Time in Seconds

pletion time on ARMixer is higher than on the CSP Mixer. And the p-value is higher than 0.05, indicating that the time participants spent to complete the mixing task on ARMixer had no significant difference with the CSP Mixer statistically. Therefore, according to Hypothesis 3, we cannot claim that the users can save time by using ARMixer to monitor mix a single audio channel compared to CSP Mixer. However, from answers to the question of "Whether ARMixer can help reduce communication with sound engineers", participants expressed strong agreement. The experimental setup for this scenario requires a professional sound engineer to monitor mix with the musicians. We will explore the quantitative analysis of this mode in comparison to ARMixer.

	ARMixer	CSP Mixer
Mean	132s	92s
SD	72.09s	24.90s

p-value = 0.0834

Table 4.3 Completion Time Results in Seconds

User Feedback

We conducted the user interview on ARMixer experience. Some comments are as below:

"It wasn't as disorienting as I expected before I used it. Seeing the mixing target through the HMD felt natural."
"I focused on the mixing target when I started to use it"
"I couldn't access as many parameters at the same time as I could with a traditional mixer"
"Cool technology"

"It is more intuitive than a traditional mixer"
"The AR HMD limited my view of my instrument"
"I noticed the weight of the headset"
"I focused on the blue ball first"
"It would be better to use a few fingers to mix"
"Moving my hand up and down to control volume was pretty cool and easy to understand"
"Volume control is the most useful to help my performance"

We compiled all the comments and elaborated the comprehensive user feedback from the following three angles.

AR Experience: Taking advantage of the immersion from AR, ARMixer allowed participants to concentrate more on the audio mixing interface and mixing target. Instead, participants had to examine the parameter positions of each knob or tried to randomize the knobs to know the corresponding effect by the sound feedback when they used the CSP Mixer. It was inconvenient for them to check repeatedly. Another participant who is a drummer thought the ARMixer would enhance spatial perception. This is because the guitarist and bassist are in their left and right positions on the stage when the drummer performs, but there is no equivalent mixing effect from wedge monitors. ARMixer would be able to intuitively mix based on spatial position. However, all participants agreed that the big and heavy AR HMD combining the three devices was the biggest obstacle to use the ARMixer. The first was the physical pressure from HMD, and the second was that the big HMD distracted them from their live performance. We are expecting lightweight AR optical glasses with wide FOV to be applied in ARMixer in the future.

Gestural Interaction: Despite the intuitive metaphor of the gestures designed, most participants thought that the accuracy of the ARMixer's gesture recognition should be improved. The most prominent problem was that when the user stopped mixing, the rest of the gestures would continue to be recognized by the system, resulting in the system continuing to perform the mixing task incorrectly and the mixing effect is immediately altered. It is one of the ongoing issues with gesture interaction and the problem ultimately arises from the unreliability

of gesture classification [40]. We believe that it would be undesirable to let the user correct the error, undo commands and try again through interface design since such errors occur frequently. In ARMixer's case, as described in Chapter 3, we believe that future gesture errors can be reduced with the aid of eye-gaze tracking, which allows the system to identify the user's behavioral state, and machine learning to classify various gestures. Moreover, users holding their arms in mid-air for long periods of time causes arm fatigue, a phenomenon known as the "gorilla-arm effect" [41]. All participants reported no arm fatigue during the experiment, but the results could be different if the participants were assigned the task of mixing multiple channels.

Potential Extension: Participants provided three directions for future extensions to ARMixer. The first is that the virtual audio mixing interface could be expanded with more mixing parameters and features such as more EQ frequencies, echoes, and compressor. The second is that ARMixer can combine traditional instruments with digital audio technologies such as synthesizers. The third is that ARMixer can serve not only as the stage monitor mixing, but perhaps as a collaborative tool for band members on stage, or even for controlling stage lighting.

Chapter 5
Conclusion

Stage monitors help musicians hear their own and other members' performance status clearly on stage to guarantee a stable show. Hence stage monitor mixing is particularly important. Nowadays, stage monitor systems have drawbacks, such as the risk of whistling from wedge monitors and the high cost of in-ear monitors for small venues. More importantly, communication between musicians and sound engineers is also a challenge. Furthermore, The intuitive stage metaphor interface is not widely utilized. Therefore, combining these factors, this research aims to design an intuitive and efficient self-stage monitor mixing system for musicians, with the interface designed with the stage metaphor. We propose the concept of the monitor mixing system using an augmented reality head-mounted display and gesture interaction, and implement the prototype ARMixer to validate its usability, intuitiveness and efficiency.

The target area of the literature review was "stage metaphor". Previous studies have shown that compared to the channel stripe metaphor, the stage metaphor has a psychoacoustic intuitive sense and excellent usability in the audio mixing interface field. However, a serious problem with the stage metaphor is that the interface becomes cluttered and difficult to manipulate when multiple mixing channels need to be processed. Furthermore, in all the proposed systems of the stage metaphor interface with gestural interaction, the visualizations are two-dimensional which is contrary to the original concept of stage metaphor —— three-dimensional deep mixing. Finally, we examined the relevant metaphorical XR systems, however there were no usability studies. Moreover, the proposed AR systems did not have a close connection with natural contexts. In this thesis, the important questions associated with stage metaphor are how does the usability of the audio mixing interface with the stage metaphor perform in a three-dimensional environment and what music scenario can be appropriate for the application of the stage metaphor.

Thus we explored the scenario based on stage monitor mixing and solved the problems with design.

We proposed the stage monitor mixing system leveraging augmented reality and gestural interaction for target user musicians to accommodate their demand of efficiency, customized mixing, and avoidance of multiple interface switching. And it is also a self-mixing system because musicians are capable of processing their own monitor mixes. We incorporated the advantage of the intuitiveness of AR and the stage metaphor, assuming that a musician on stage with their instrument can be treated as one of the audio channels. When the user with the AR HMD faces that musician, the corresponding virtual audio mixing interface with stage metaphor is displayed on the real stage. At the same time, only one audio channel is processed per mix which means the interface is simple and the stage metaphor is appropriate to be used in this monitor mixing scenario. Furthermore, the user can control the mixing interface through metaphorical gestures. Therefore, the user's gestural input and visual feedback are combined in an immersive 3D reality/virtual environment and result in an intuitive "What You See Is What You Get" experience. Finally, We designed the ideal system architecture of a lightweight AR glasses for this concept and proposed a minimum viable prototype ARMixer and its interface elements.

To validate our proposed concept, we used the video see-through AR solution to implement the prototype ARMixer. We created an AR HMD compatible with see-through video and hand tracking by combining a Quest 2, Zed Mini, and Leap Motion. The gesture interaction and interface were edited in Unity. Consequently, the AR view and mixing effects could be output in real-time. We conducted 2 usability experiments with 11 musicians in total: one was a quantitative study in which individual participants mixed the same digital music using ARMixer and a traditional CSP Mixer, the other was a qualitative study in which 2 two-piece bands simulated a stage monitor mixing scenario, using ARMixer to perform real-time instrument channel mixes. As a result, ARMixer is acceptable and has potential to be a good product. Also in the intuitive evaluation, its interactions including AR and parameter controls provided a psychologically intuitive experience to the user. Finally, the quantitative analysis illustrated that there was no significant difference in efficiency performance between ARMixer and CSP Mixer,

implying that it cannot be stated that using ARMixer is more efficient than using CSP Mixer. However, the participants strongly agreed that ARMixer was able to reduce communication time with sound engineers and expressed enthusiasm for using ARMixer.

In summary, the users deemed it acceptable to mix via gestures in augmented reality. And the stage metaphor can be applied to this scenario well. ARMixer provides excellent psychoacoustic intuitiveness and usability so that users can mix immersively and easily. The current ARMixer does not perform efficiently compared to the CSP Mixer because there are some limitations. The first is that the overly big and heavy HMD distracts users from their performance, and the second is that the low accuracy of hand tracking causes the mixing results to be mis-processed. The third is that due to the COVID-19 pandemic, we are temporarily unable to organize a user study with a band of more than 4 people and sound engineers to simulate real user scenarios. In future ARMixer research, we will explore the use of AR glasses compatible with eye-gaze tracking, support multi-channel audio recognition and use machine learning to improve the accuracy of hand tracking. We will also iterate the design in terms of collaboration between musicians on stage. Finally, we hope to conduct more tests with professional musicians and an updated prototype to receive more accurate insight.

References

[1] Roey Izhaki. *Mixing audio: concepts, practices, and tools*. Routledge, 2017.

[2] Mads Walther-Hansen. New and old user interface metaphors in music production. *Journal on the Art of Record Production*, (11), 2017.

[3] David Gibson. The art of mixing: A visual guide to recording. *Engineering, and Production*, 236, 1997.

[4] Josh Mycroft, Tony Stockman, and Joshua D Reiss. Audio mixing displays: The influence of overviews on information search and critical listening. In *International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2015.

[5] Steven Gelineck and Anders Kirk Uhrenholt. Exploring visualisation of channel activity, levels and eq for user interfaces implementing the stage metaphor for music mixing. In *2nd AES Workshop on Intelligent Music Production*. Audio Engineering Society, 2016.

[6] Christopher Dewey and Jonathan P Wakefield. Formal usability evaluation of audio track widget graphical representation for two-dimensional stage audio mixing interface. *Journal of the Audio Engineering Society*, 2017.

[7] Sijie Wang and Weiwei Yu. Space elements of computer music production based on vr technology. *IEEE Access*, 2020.

[8] David Rivas Méndez, Calum Armstrong, Jessica Stubbs, Mirek Stiles, and Gavin Kearney. Practical recording techniques for music production with six-degrees of freedom virtual reality. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[9] Silvin Willemsen, Anca-Simona Horvath, and Mauro Nascimben. Digidrum: A haptic-based virtual reality musical instrument and a case study. In *Proceedings of the 17th Sound and Music Computing Conference*, pages 292–299, 2020.

[10] Liang Men and Nick Bryan-Kinns. Lemo: supporting collaborative music making in virtual reality. In *2018 IEEE 4th VR workshop on sonic interactions for virtual environments (SIVE)*, pages 1–6. IEEE, 2018.

[11] Andries Valstar, Min-Chieh Hsiu, Te-Yen Wu, and Mike Y Chen. Giggler: An intuitive, real-time integrated wireless in-ear monitoring and personal mixing system using mobile devices. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 971–974, 2015.

[12] Brad Herring. *Sound, lighting and video: A resource for worship*. Routledge, 2012.

[13] Matthew Duignan, James Noble, Pippin Barr, and Robert Biddle. Metaphors for electronic music production in reason and live. In *Asia-Pacific Conference on Computer Human Interaction*, pages 111–120. Springer, 2004.

[14] Jarrod Ratcliffe. Hand motion-controlled audio mixing interface. In *Proceedings of NIME*, 2014.

[15] Steven Gelineck, Dannie Michael Korsgaard, and Morten Büchert. Stage-vs. channel-strip metaphor: Comparing performance when adjusting volume and panning of a single channel in a stereo mix. In *New Interfaces for Musical Expression*, pages 343–346. Louisiana State University, 2015.

[16] Konstantinos Drossos, Andreas Floros, and Konstantinos Koukoudis. Gestural user interface for audio multitrack real-time stereo mixing. In *Proceedings of the 8th Audio Mostly Conference*, pages 1–6, 2013.

[17] Michal Lech and Bozena Kostek. Testing a novel gesture-based mixing interface. *Journal of the Audio Engineering Society*, 61(5):301–313, 2013.

[18] Jarrod Ratcliffe. Motionmix: A gestural audio mixing controller. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[19] Jonathan Wakefield, Christopher Dewey, and William Gale. Lami: A gesturally controlled three-dimensional stage leap (motion-based) audio mixing interface. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[20] Alex Volgan. Designing vr tools: The good, the bad, and the ugly, 2015. URL: `https://medium.com/@LeapMotion/designing-vr-tools-f9d484dd244b/`.

[21] Richard Graham and Seth Cluett. The soundfield as sound object: Virtual reality environments as a three-dimensional canvas for music composition. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[22] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[23] Shoki Miyagawa, Yuki Koyama, Jun Kato, Masataka Goto, and Shigeo Morishima. Placing music in space: A study on music appreciation with spatial mapping. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, pages 39–43, 2018.

[24] Valentin Bauer and Tifanie Bouchara. First steps towards augmented reality interactive electronic music production. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 90–93. IEEE, 2021.

[25] Steven Gelineck and Dannie Michael Korsgaard. Stage metaphor mixing on a multi-touch tablet device. In *Audio Engineering Society 2014*. Audio Engineering Society, 2014.

[26] Golden Krishna. *The best interface is no interface: The simple path to brilliant technology*. Pearson Education, 2015.

[27] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented real-

ity. *IEEE transactions on visualization and computer graphics*, 23(6):1706–1724, 2016.

[28] David Tennenhouse. Proactive computing. *Communications of the ACM*, 43(5):43–50, 2000.

[29] Mike Wozniewski, Zack Settel, and Jeremy R Cooperstock. A spatial interface for audio and music production. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx06), Montreal, Canada*, page 271. Citeseer, 2006.

[30] Jaron Lanier, Victor Mateevitsi, Kishore Rathinavel, Lior Shapira, Joseph Menke, Patrick Therien, Joshua Hudman, Gheric Speiginer, Andrea Stevenson Won, Andrzej Banburski, et al. The realitymashers: Augmented reality wide field-of-view optical see-through head mounted displays. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 141–146. IEEE, 2016.

[31] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. Gaze+ pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, pages 99–108, 2017.

[32] Christopher Dewey and Jonathan Wakefield. Elicitation and quantitative analysis of user requirements for audio mixing interface. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[33] Andrew D Wilson. Robust computer vision-based detection of pinching for one and two-handed gesture input. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 255–258, 2006.

[34] Zhanpeng Huang, Weikai Li, and Pan Hui. Ubii: Towards seamless interaction between digital and physical worlds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 341–350, 2015.

[35] Jannick P Rolland, Richard L Holloway, and Henry Fuchs. Comparison of optical and video see-through, head-mounted displays. In *Telemanipulator and Telepresence Technologies*, volume 2351, pages 293–307. International Society for Optics and Photonics, 1995.

[36] Stereolabs. Using leap motion and zed mini for ar hand tracking, 2018. URL: https://www.stereolabs.com/blog/leap-motion-zed-mini-hand-tracking-ar/.

[37] ISO 9241-11. Iso 9241-11: 1998, ergonomic requirements for office work with visual display terminals (vdts)—part 11: Guidance on usability. 1998.

[38] John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189, 1996.

[39] William MK Trochim. Research methods knowledge base: Likert scaling. *Retrieved on January*, 15:2009, 2006.

[40] Amy J. Ko. User interface software and technology, 2017. URL: https://faculty.washington.edu/ajko/books/user-interface-software-and-technology/.

[41] Sebastian Boring, Marko Jurmu, and Andreas Butz. Scroll, tilt or move it: using mobile phones to continuously control pointers on large public displays. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, pages 161–168, 2009.