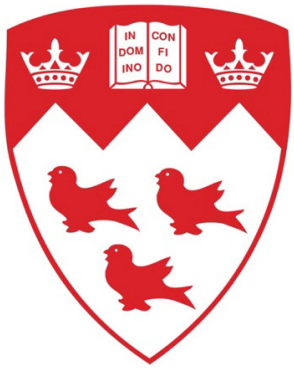
Capturing Orchestral Music for Three-Dimensional Audio Playback



Will Howie

Department of Music Research

Schulich School of Music

McGill University, Montreal

Submitted March 2018

A thesis submitted to McGill University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

This thesis details the design, implementation, and evaluation of a novel technique for orchestral music capture for three-dimensional audio reproduction. The technique is optimized for Japan Broadcasting Corp. (NHK)'s "22.2 Multichannel Sound" three-dimensional audio playback format. The design of the technique draws upon previous research in spatial hearing, music recording for stereo and multichannel playback environments, concert hall acoustics, spatial impression in multichannel audio, and subjective evaluation and analysis of reproduced sound. Preliminary experiments investigate immersion and envelopment in three-dimensional music recording, as well as the relationship between microphone polar patterns and vertical height channel signal capture. A novel technique for three-dimensional orchestral music recording is then introduced. The technique is designed to capture a fully immersive sound scene featuring a cohesive orchestral image with realistic horizontal and vertical extent, stable sound source imaging, natural ensemble and scene depth, and a highly enveloping ambient sound field. A series of formal and informal subjective evaluations show that the proposed technique achieves these sonic imaging goals, and is suitable for 3D commercial music production and immersive content creation for broadcast. This new microphone technique is also applicable to other genres of music recording, as well as productions optimized for smaller-scale 3D audio formats. Further investigation finds 22.2 Multichannel Sound to be perceptually unique among common 3D audio formats with respect to the reproduction of acoustic music. A library of high-quality 3D audio test material was created for this research, which will be made available to other researchers for future studies.

RÉSUMÉ

Cette thèse détaille la conception, la mise en œuvre et l'évaluation d'une nouvelle technique en prise de son orchestrale en trois dimensions. La technique est optimisée pour le format de lecture audio tridimensionnelle "22.2 Multichannel Sound" de Japan Broadcasting Corp. (NHK). La conception de la technique s'appuie sur des recherches antérieures dans l'audition spatiale, l'enregistrement de musique pour les environnements de lecture stéréo et multicanaux, l'acoustique des salles de concert, l'impression spatiale dans l'audio multicanal et l'évaluation subjective du son reproduit.

Les expériences préliminaires font focus sur l'immersion et la sensation d'enveloppement d'enregistrement musical 3D ainsi que la relation entre la courbe de directivité des microphones et la prise de son de canaux verticaux surélevés. Une technique novatrice en prise de son 3D pour musique orchestrale en est découlée. Cette technique est conçue pour une prise de son pleinement immersive caractérisée par une image orchestrale cohérente avec étendue verticale et horizontale réaliste, une image de la source stable, une perception de profondeur ainsi qu'une ambiance hautement décorrélée. Une série d'évaluations subjectives formelles et informelles montre que la technique proposée atteint ces objectifs d'imagerie sonique, et convient à la production de musique commerciale 3D et à la création de contenu immersif pour la diffusion. Cette nouvelle technique de microphone est également applicable à d'autres genres d'enregistrement de musique, et aux productions optimisées pour formats audio 3D à plus petite échelle. Les recherches approfondies révèle que le son multicanal 22.2 est perceptuellement unique parmi les formats audio 3D courants en ce qui concerne la reproduction de la musique acoustique. Une librairie d'enregistrements 3D haute-qualité servant de matériel test a ainsi été créé et sera mis à la disposition d’autres chercheurs pour de futures études.

ACKNOWLEDGEMENTS

I would like to express my appreciation to my co-supervisors, Prof. Richard King and Prof. Wieslaw Woszczyk for their guidance throughout this process. A huge thanks to Ieronim Catanescu, Yves Méthot and Julien Boissinot for their outstanding technical support and know-how.

Thanks to Bryan Martin, Matt Boerum, Denis Martin, Dave Benson, Dr. Sungyoung Kim, and Dr. Doyuen Ko for their profoundly valuable advice over the years, and for helping to keep me sane.

In Japan: Thanks to Toru Kamekawa, Dr. Atsushi Marui, Misaki Hasuo, and Hidetaka Imamura for their help in facilitating the listening tests and presentations that took place at Tokyo University of the Arts. And thank you to Kensuke Irie for his continual support and help at the NHK.

Special thanks to my family for being a fantastic bunch of super encouraging weirdos.

This research would not have been possible without extraordinary financial support from the Social Sciences and Humanities Research Council (SSHRC), as well as generous support from McGill University, the Schulich School of Music, and the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT).

PREFACE

This dissertation is presented as a manuscript thesis, wherein the bulk of the document is taken from previously published papers. For Chapters 3 through 6, the text appears mostly as originally published, apart from changes to formatting, figures, tables, and reference numbers. Some text and references have been updated or eliminated to reflect the grouping of these manuscripts into a single document. Also, some discussion sections have been expanded to include information omitted from the original published versions owing to document length limitations. Section 5.7 of Chapter 5 contains new information from a follow up study that took place after the original paper was completed. Written consent to reproduce previously published material in this manuscript has been granted by my various co-authors, as well as the Managing Editor of the Journal of the Audio Engineering Society. All figures within this manuscript fall under the sole property and ownership of the author, Will Howie, unless otherwise noted.

ORIGINAL SCHOLARSHIP AND DISTINCT CONTRIBUTIONS TO KNOWLEDGE

The study undertaken in **Chapter 3.2** is the first (and currently only) formal investigation of the use of different microphone polar patterns for height channels for music recording, in terms of listener preference. Results of the study, as well as general impressions reported by the investigators, give current 3D audio practitioners valuable insight into the design and execution of 3D microphone arrays. Observations from **Chapter 3.1** help confirm that height channels increase immersion in music recordings, as well as the importance of lateral reflected sound energy for achieving strong levels of listener envelopment.

Chapter 4 introduces a novel approach to orchestral music recording for 22.2 Multichannel Sound: a combination of omnidirectional microphones for orchestral sound capture, directional bottom channel microphones to capture floor reflections and vertical orchestral imaging, and an ambience array designed to capture many points of decorrelated reflected sound energy. Chapter 4's investigation into listener perception of the bottom channels in a 22.2 system is the first of its kind, and contributes to a greater understanding of the role of bottom channels in 3D music reproduction. The recording methodology described in Chapter 4 is easily scalable to other current immersive audio formats, and functions as a valuable guide for content creators working in commercial 3D music production, broadcast, film, video game audio, and virtual reality. The resultant test recordings have already been used in several further studies by researchers at McGill, BBC, and Rochester Institute of Technology, and constitute part of a larger body of high-quality three-dimensional music recordings created for this thesis.

Chapter 5's perceptual comparison of three different orchestral capture methods optimized for 22.2 Multichannel Sound is the first of its kind. This study is also one of a very small number of formal investigations into perceptual differences between several microphone arrays optimized for 3D music capture. Additionally, no known previous research or publication has detailed the simultaneous execution of multiple 22.2-optimized orchestral recording techniques. Methodology and results from this study provide valuable practical, perceptual, and aesthetic insights for current 3D audio practitioners, particularly those using high channel-count formats, such as 22.2 or 11.1. The consistently poor performance of ambisonics-based music recording techniques observed in previous studies comparing two-dimensional microphone techniques is seen again here, for a three-dimensional audio environment.

Chapter 6 details the first known perceptual comparison of different 3D audio playback formats that focuses specifically on listener discrimination. It is also one of the few studies related to 3D audio where the stimuli are sourced from 3D music recordings that have been deemed "critical testing material". The creation of additional high-quality 3D audio stimuli for listening tests is something several previous studies had concluded was necessary. Chapter 6 is also the first published study to compare 22.2 with three other standardized 3D audio formats. Section 6.2 introduces several new concepts in three-dimensional music recording, particularly the construction of microphone arrays that prioritize capturing a wide range of spectral content from each instrument, as well as a realistic presentation of horizontal and vertical extent within the sound scene. The results of this study suggest that within the context of music reproduction, far greater perceptual differences exist between 22.2 and other 3D audio formats than have been reported in previous research.

CONTRIBUTIONS OF AUTHORS

For all previous published work presented in this thesis (Chapters 3-6) I was the principal author, and was responsible for all background research, development of the research questions, design and implementation of new recording techniques, design and administration of listening tests, and interpretation of the results. Listed below are the contributions of my various co-authors.

Chapter 3.1

W. Howie and R. King, "Exploratory microphone techniques for three-dimensional classical music recording," in *AES Convention 138*, Warsaw, 2015.

Richard King aided in designing the experimental recording technique, and framing the discussion of the results.

Chapter 3.2

W. Howie, R. King, M. Boerum, D. Benson, A. Han, "Listener preference for height channel microphone polar patterns in three-dimensional recording," in *AES Convention 139*, New York, 2015.

Richard King aided in the mixing and level matching of stimuli, and framing the research question. Matt Boerum built the Max/MSP patch used to administer the listening test. Dave Benson conducted the statistical analysis of the listener data. Alan Han assisted with the stimulus recordings.

Chapter 4

W. Howie, R. King, D. Martin, "A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound," in *AES Convention 141*, Los Angeles, 2016.

Richard King aided in framing the research question and interpreting the results. Denis Martin created the Max/MSP patch used to administer the listening test, conducted the statistical analysis of the listener data, and aided in interpreting the results.

Chapter 5

W. Howie, R. King, D. Martin, F. Grond, "Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio," in *AES Convention 142*, Berlin, 2017.

Richard King aided in the mixing and level matching of test stimuli. Denis Martin created the Max/MSP patch used to administer the listening test, conducted the statistical analysis of the listener data, and aided in interpreting the results. Florian Grond built a spatial decoder for the Eigenmike audio files, and contributed to the discussion and explanations of recording using higher order ambisonics. For the content found in the addendum, Toru Kamekawa provided a venue for a 2nd listening test, and assisted in preparation of the experiment. Misaki Hasuo assisted in administering the listening tests.

Chapter 6

W. Howie, R. King, D. Martin, "Listener Discrimination Between Common Speaker-Based 3D Audio Reproduction Formats," J. Audio Eng. Soc., Vol. 65, No.10, 2017, pp. 796-805.

Richard King aided in framing the research question. Denis Martin created the Max/MSP patch used to administer the listening test, conducted the statistical analysis of the listener data, and aided in interpreting the results.

CONTENTS

List of Tables

List of Figures

1 INTRODUCTION

1.1 Motivation

Listening to live orchestral music is an experience that is unique in terms of the size and breadth of the ensemble's sonic image, intensity and extremes of musical dynamic range, and integration of the concert venue's acoustic signature with the music. As countless genres and styles of music have come and gone, orchestral music has maintained popularity with audiences, not only within the context of live concert attendance, but also for commercial music recordings, film scores, video game soundtracks, and live broadcasts over television and internet streaming. With the exception of binaural recording techniques [1], music recording and reproduction has, until very recently, been primarily one or two-dimensional in nature. Stereo and traditional 5.1 surround sound reproduction systems reproduce sound only in the horizontal plane, at ear level: a compromised listening experience that does not deliver a fully immersive or realistic auditory scene. Three-dimensional sound reproduction with vertically oriented "height" channels has been shown to improve the depth, presence, envelopment, naturalness, and intensity of music recordings [2], [3], [4], bringing the listener closer to an ideal, more realistic listening experience.

In recent years, numerous 3D audio formats for cinema, home theatre, and broadcast have been introduced [5], [6], [7]. The most advanced and robust of these new audio reproduction formats is Japan Broadcasting Corp. (NHK)'s "22.2 Multichannel Sound" (22.2), the immersive audio component of their ultra-high resolution 8K video broadcast format: Super Hi-Vision [8]. Utilizing ten playback channels (loudspeakers) at ear level, nine above the listener, and three at floor level (Figure 37), 22.2 has the potential to create highly realistic, richly enveloping presentations of recorded music. 22.2 is also an ideal format for recreating the rich sonic experience of listening to a live orchestra. Five frontal loudspeakers at ear level, with a reproduction angle of 120° (Figure 51) allow for the presentation of an orchestral image that closely matches the true horizontal extent of the ensemble. The three bottom channels, which are unique to 22.2 among currently standardized 3D audio formats, vertically extend the orchestral image to the floor, as it would appear from the conductor's perspective. An even spatial distribution of numerous height and surrounding channels ensures an accurate recreation of the performance venue's early and late reflected sound energy. Owing to a downward compatibility of number and position of loudspeakers, any music capture technique designed for 22.2 could easily be adapted to other common 3D audio formats. Surprisingly then, very few authors have discussed techniques for orchestral music recording for 22.2 or any other 3D playback format, and none have addressed the importance of lower vertical channels for the reproduction of a realistic, vertically anchored orchestral image [9], [10].

Within the current literature on acoustic music recording for 3D audio, there is also a decided lack of empirical studies aimed at comparing or evaluating proposed recording techniques. Ideally, any newly developed technique for orchestral music capture would be validated in terms of its potential as a viable system for broadcast and commercial recording through subjective or objective means. Another important consideration is the practicality of

the recording playback environment. NHK has developed a range of simplified playback systems to deliver 22.2 to consumers [11]. However, the technical demands for producing and reproducing content for 22.2 are still high. This begs the question: for the reproduction of orchestral music, could a lower-channel count, simpler 3D audio format deliver a listening experience that is perceptually indistinguishable from that of 22.2?

1.2 Research Goals

This thesis has four primary research goals:

1. To develop a technique for orchestral music capture, optimized to exploit the full potential of the 22.2 format to deliver an ideal listening experience.
2. To confirm the validity of the technique developed in (1) for broadcast and commercial recording through formal subjective evaluations by trained listeners.
3. To determine whether the same experience delivered by 22.2 for the reproduction of acoustic music can be achieved through the use of a 3D audio format with a reduced channel count.
4. The creation of high-quality three-dimensional audio recordings that can be used as stimuli for current and future studies by researchers at McGill University and elsewhere.

1.3 Structure of Thesis

This thesis is comprised of the following chapters:

Chapter 1 | Introduction summarizes the motivation and goals of the research.

Chapter 2 | Background provides an overview of previous literature in several areas pertinent to the development and evaluation of three-dimensional music capture techniques: spatial hearing, stereo, multichannel, and three-dimensional acoustic music recording techniques, three-dimensional audio formats, and considerations for subjective and objective analysis of reproduced sound.

Chapter 3 | Preliminary Experiments details the development of several small-scale three-dimensional microphone arrays, all optimized for acoustic music capture. This research focuses on gaining a better understanding of what is necessary to achieve natural instrument or ensemble images within a 3D audio playback environment, and what kind of sonic information is required for achieving strong levels of listener envelopment. This chapter also explores the relationship between microphone polar patterns and height channel sound capture, and whether a strong preference exists among listeners.

Chapter 4 | A Three-Dimensional Orchestral Music Recording Technique, optimized for 22.2 Multichannel Sound is introduced. The technique builds on experiments from Chapter 3, as well as previous research in music recording and spatial impression for multichannel audio. A test recording is shown to perform well in informal listening sessions at five different 22.2 playback environments in Canada and Japan. The recording's sonic image is observed to remain consistent across multiple playback environments. A subjective listening test shows that within the context of dynamic orchestral music, subjects can successfully differentiate between playback conditions with and without bottom channels.

Chapter 5 | Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio follows directly from the informal evaluations from Chapter 4. The proposed technique is compared with a current production standard for orchestral music recording for 22.2, as well as a spherical higher order ambisonics capture system. Results of a formal subjective listening test show the proposed technique performs as well or better than the other techniques under investigation for all subjective attributes examined: "clarity", "scene depth", "naturalness", "environmental envelopment", "sound source envelopment", and "quality of orchestral image". These results are confirmed in a follow-up study using a different demographic and cultural group as listeners. Preference between the recording techniques under investigation is also examined.

Chapter 6 | Listener Discrimination Between Common Speaker-Based 3D Audio Reproduction Formats addresses the 3rd research question of the thesis: *having designed, implemented, and tested a technique for orchestral music capture optimized for 22.2, can the same listening experience be achieved with a simpler, reduced-channel 3D audio format?* 22.2 is shown to deliver an acoustic music listening experience that is significantly perceptually different from other currently standardized 3D audio formats. This chapter also details the creation of a small library of high-quality music recordings for multiple 3D audio formats, which can be used by other researchers in future studies.

Chapter 7 | Conclusions: A summary of the general conclusions from each chapter is given, as well as additional discussion. Possibilities for future work are also addressed.

2 BACKGROUND

Within the published articles that comprise the bulk of this thesis, previous research is often cited or referred to, but not deeply explained, for the sake of brevity. What follows is a more in-depth review of several areas of pertinent audio and auditory scholarship.

2.1 Spatial Hearing

The mechanisms that allow humans to locate sound in three-dimensional space are complex and varied; an understanding of these mechanisms is necessary for any deep discussion of three-dimensional sound capture. Blauert's *Spatial Hearing* [12] is a classic foundational text on the psychoacoustics of human sound localization, detailing much of the author's own research, as well as work by others. For further scholarship in this field, *Principles and Application of Spatial Hearing* [13] is an excellent reference.

2.1.1 Localization and Localization Blur

"Localization" refers to the relationship between the actual position of a sound source in physical space, and the perceptual location of the corresponding auditory event. Accuracy of localization can be expressed in terms of "localization blur": a measurement of the "smallest possible change in position of the sound source that produces a just-noticeable change in position of the sound event. [12]" Humans are able to locate sound events with a remarkable

degree of accuracy, particularly in the forward, horizontal plane. Figure 1 shows the results of a large-scale investigation into localization blur in the horizontal plane [12]. As can be seen, human acuity in sound source localization is best in the frontal area, with an increase in localization blur as the sound sources approach a fully lateral position, relative to the direction of the head. Moving behind the head, localization blur decreases as the sound source position approaches 180º. It should be noted that for narrow band signals (e.g., sine tones) a phenomenon commonly known as "front/back confusion" may occur, wherein the auditory event is perceived as being in a location that is the mirror opposite of the sound source. For example, a narrow-band sound event at 30º could be incorrectly perceived as being located at 150º.

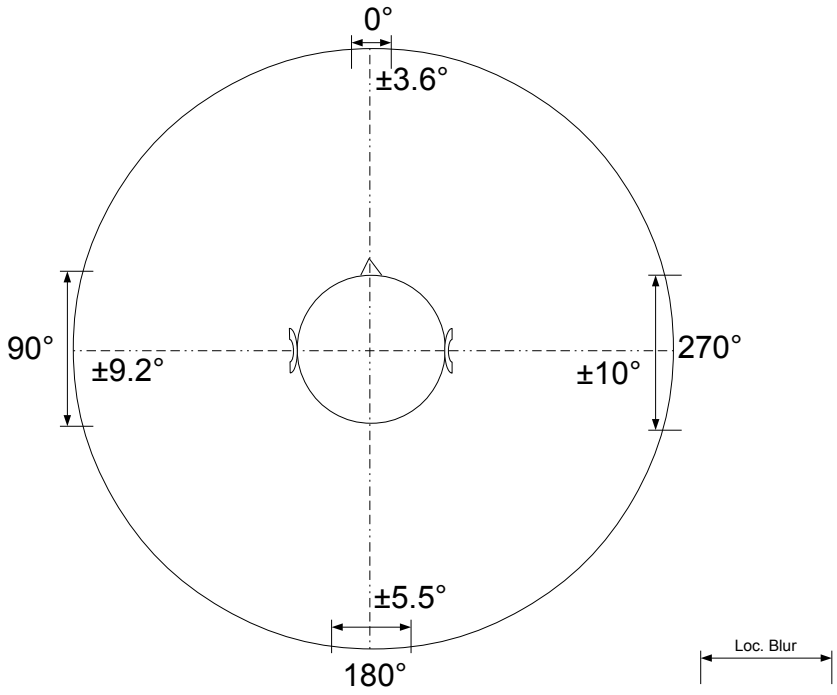


Figure 1: Localization blur in the horizontal plane, based on a diagram from [12]. Note: angles are not to scale.

2.1.2 Sound Localization in the Horizontal Plane

There are two main mechanisms that make sound source localization possible in the horizontal plane; both are caused by differences that exist between the two ears for a given signal. The first, and most important, are interaural time differences (ITD), which are summarized by Rumsey: "A sound source located off the 0º (centre front) axis will give rise to a time difference between the signals arriving at the ear of the listener that is related to its angle of incidence […] This rises to a maximum for sources at the side of the head, and enables the brain to localize sources in the direction of the earlier ear. [1]" An interaural time difference of approximately 650μs results in the full lateral displacement of an auditory event, i.e. the auditory event will be located at ±90°, depending on which ear the sound reaches first. ITDs are particularly useful for locating transient information such as the onsets and ends of sounds, but are ineffective for locating steady-state pure tones above 1.6kHz.

Interaural level differences (ILD) take place in higher frequency ranges, at which point the head acts as a physical sound barrier. As a sound source moves laterally from 0º, there will be an increase in sound pressure for one ear over the other, resulting in a corresponding shift in auditory event localization. This is true for sound pressure level (SPL) differences up to 15-20dB, at which point the auditory event will be located at ±90°. ILD's are important for signals that contain components above 1.6kHz and are of relatively low SPL [12].

2.1.3 Sound Localization in the Vertical Plane

Localization becomes somewhat more complicated in the vertical plane, particularly the median vertical plane (MVP), where ITDs and ILDs are, by definition, not present. Not surprisingly then, localization blur is at its worst in the MVP, which can be seen in Figure 2, from Blauert [12], based on research by Damaske and Wagener [14].

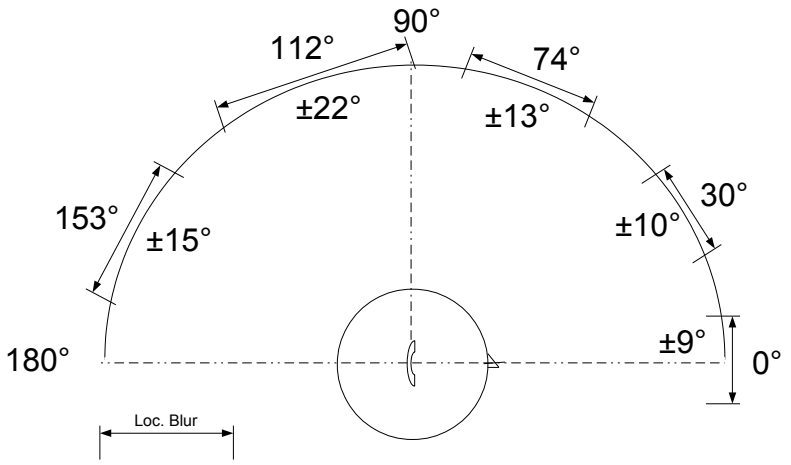


Figure 2: Localization blur in the median plane for continuous speech by a familiar person, based on a diagram from [12]. Note: angles are not to scale.

Figure 2 illustrates how as the sound source leaves 0º elevation localization deteriorates, and is particularly poor behind and above the head. In the median plane, monaural mechanisms for localization become more important, particularly those related to the spectral components of the sound source. Blauert states: "The pinna, along with the ear canal, forms a system of acoustic resonators. The degree to which individual resonances of this system are excited depends on the direction and distance of the sound source [12]." Roffler and Butler [15] performed extensive research in sound localization in the vertical plane, and concluded that for sound localization to be possible in the MVP, three factors must be present:

1. The sound must be complex
2. The sound must contain components above 7kHz
3. The pinna must be present

Butler and Humanski [16] examined the existence and importance of binaural cues (ILDs and ITDs) for sound localization in both the lateral and median vertical planes. They found that for the lateral vertical plane (LVP), binaural localization can be quite accurate, regardless of the presence of high frequency components. For the MVP, however, they confirmed that

spectral components above 7kHz must be present for localization accuracy above the level of chance. Butler and Humanski concluded that for the LVP, binaural cues (ITDs and ITDs) are by far the most important cues for sound localization, but that the presence of monaural spectral cues (e.g., pinna resonances) adds some improvement to accuracy [16].

The torso, as a large reflective surface, has also been examined as a contributing factor to sound localization. Algazi et al. [17] showed the existence of low-frequency cues for elevation that are significant away from the median plane; these cues are likely the result of sound reflecting off the torso and shoulders. Lee [18] has recently theorized that low-frequency spectral notches caused by torso reflections may be exploited within the context of creating elevated phantom images for 3D audio reproduction.

2.1.4 Other Monaural Cues

An interesting aspect of human hearing in the vertical plane is the so called “pitch-height” effect. This refers to the phenomenon wherein humans tend to perceive higher frequency tones or filtered noise as coming from higher up in physical space, and lower frequency tones or filtered noise as originating from positions lower in space. Pratt [19] first reported on this phenomenon in 1930. In that study, subjects heard a randomly presented series of tones, ranging from 256Hz to 4096Hz, being reproduced by a telephone receiver behind an acoustic screen. All subjects consistently perceived higher-frequency tones as emanating from positions higher in physical space, and lower-frequency tones from positions lower in space, regardless of the true origin point of the tone [19]. Roffler and Butler [20] examined this effect extensively using pulsed sine tones as stimuli. They found that for several different experimental conditions, including one using young children as subjects who had not yet developed an association between height and pitch, “the elevation angle of the auditory event was described as varying as a function of the frequency of the sound event [12]”, [20]. This effect was confirmed for band-passed noise signals as well [15]. The pitch-height effect has

also been investigated and confirmed by Blauert [12], Cabrera and Tilley [21], Lee [22] and others. Cabrera and Tilley [21] theorized that the pitch-height effect may be the result of spectral notches in signals caused by early floor reflections: a learned reflection cue similar to pinna cues for MPV localization.

Blauert undertook a series of investigations into the relationship between the perceived origin location of an auditory event, and the spectral content of the sound source. The results of these studies are summarized in [12], and indicate a correspondence between the centre frequency of a given test signal and its perceived location in the median plane. Blauert then introduces the concept of "directional bands": certain frequency bands that when boosted or attenuated correspond to a particular auditory event location in front, behind, or above the head. For example, the 8kHz band is reported as having a strong correspondence with "above the head" localization [12], which was confirmed in a subsequent study by Lee [23]. Similar to Blauert's findings, Hebrank and Wright [24] show that "median plane localization of white noise is based on simple but deep spectral cues generated by the directional filtering of the external ear." Hebrank and Wright categorize several broad directional bands that are in agreement with Blauert's findings. In a separate study, using 1/3-octave band noise bursts, Wallis and Lee [25] also confirmed the existence of directional bands at 1kHz, 4kHz, and 8kHz.

Distance and depth perception are also key components to spatial hearing, though they exhibit less accuracy than localization. Rumsey [1] and Blauert [12] both summarize the ways in which an auditory event will change as the sound source moves farther away from the listener:

1. For intermediate distances travelled (3-15m) the sound will become quieter
2. For distances greater than 15m, high frequencies are attenuated, due to air absorption
3. The sound becomes more reverberant (in reflective spaces)
4. There is less difference in time between direct sound incident and the first floor reflection
5. Ground reflections become attenuated

2.1.5 Inter-channel Differences in Loudspeaker-Based Sound Reproduction

For stereophonic sound reproduction, an almost universally accepted optimal layout exists for loudspeaker and listener positions, based on many years of research and practical observations [1]. The Left and Right loudspeakers should form an equilateral triangle with the listener, thereby providing a frontal sound reproduction angle of ±30°. If both loudspeakers output coherent, i.e. identical, signals that do not differ in terms of level or time, the signals perceptually fuse into a single auditory event, which will be localized at the centre point (0°) between the two loudspeakers. This auditory event is known as a “phantom image” [26].

2.1.5.1 Inter-channel Timing Differences

When an inter-channel delay is introduced between two coherent signals, the auditory event, i.e. phantom image will gradually shift in direction towards the location of the non-delayed sound source (loudspeaker). A delay of approximately 1.2ms (note: this figure varies depending on the author) will result in a phantom image that is fully localized on or at a given loudspeaker. The relationship between delay time and angular location of the auditory event remains mostly linear from 0° to ±20°; past ±20°, a greater amount of inter-channel delay is

required to maintain the same amount of angular displacement [26]. For inter-channel delays in the order of 30-40ms, the auditory event remains singular, though modified in terms of timbre and image spread. For inter-channel delays above 40-50ms, two auditory events become perceptible: the original event and an echo [26]. For stereophony based on inter-channel time differences, localization is more accurate for impulsive sounds and sounds with well-defined transient information [26]; continuous sounds may appear to shift in position [1].

2.1.5.2 Inter-channel level differences

When a difference in amplitude is introduced between coherent loudspeaker signals, the location of the phantom image will shift towards the louder of the two signals. The amount of level difference required for the sound source to localize on or at a given loudspeaker is reported at anywhere between 14 to 24 dB, depending on the experimental conditions and type of sound source used [26]. The relationship between angle of auditory event localization and inter-channel level difference follows a more linear trajectory than for inter-channel timing differences [26]. For frontal sound reproduction, inter-channel level differences tend to produce more accurate auditory event localization than inter-channel time difference [27].

2.1.5.3 Combined time and level differences

When combined, inter-channel time and level differences will have either an additive or subtractive effect on the lateralization of the phantom image. If the effects add to each other, i.e. the amplitude of one channel is increased while the delay of the other channel is increased, the lateralization of the auditory event towards a given loudspeaker will be greater than for only one type of inter-channel difference. The opposite is also true: the amount of perceived angular displacement of the phantom image from 0° can be reduced or even nullified by trading time and level differences against each other; this is known as the "compensation phenomenon" [26]. Rumsey notes that "the exact relationship between time

and level differences needed to place a source in a certain position is disputed by different authors and seems to depend to some extent on the source characteristics. [1]”

2.1.5.4 Some considerations for multi-channel audio

It should be noted that most of the above findings are based on experimental conditions using frontal, 2-channel stereophonic sound reproduction. Martin et al. [27] showed that for loudspeakers placed behind the listener at ±60°, i.e. for 5-channel surround sound [28], an inter-channel delay of only 0.6ms was required to produce a phantom image at the location of the non-delayed loudspeaker: approximately half the delay required to achieve the same effect for frontal 2-channel stereophony. Corey and Woszczyk [29] found that for a 5-channel reproduction system, localization blur for phantom images produced between side speakers was very similar to Blauert’s findings for real sources in the lateral area. The same study also found that type of sound source had significant effects on the perceived location of lateral phantom images [29].

2.1.5.5 Varying inter-channel correlation

Kurozumi and Ohgushi [31] investigated the relationship between inter-channel correlation of two-channel acoustic signals and perceived sound image quality. They found that the perceived width of the sound source is strongly dependent on the degree of correlation between the two loudspeaker channels: a decrease in inter-channel correlation results in a dimensional broadening of the sound image [31]. A complete lack of correlation between channels will typically result in a perceptual “hole” in the middle of the stereophonic sound image [32].

2.2 Stereo and 5-Channel Acoustic Music Recording Techniques

Many techniques have been developed for acoustic music recording, optimized for stereo, 5.1 multichannel audio, or both. This review will focus on techniques suited to classical music recording in acoustic spaces, mainly drawing from Rumsey [1], Dickreiter [31], Hugonnet

and Walder [26], and Imirajiri [32]. The term "surround" will always refer to a 5.1 surround sound speaker layout, as per ITU-R BS.775-3 [28]. Binaural "dummy head" recording techniques will not be discussed, as it is a recording system optimized for headphone reproduction, which falls outside the scope of this thesis.

2.2.1 Useful Acceptance Angle of Microphone Systems

Known alternately as the Stereophonic Recording Angle (SRA) or Useful Acceptance Angle, this refers to the sector of the sound field in front of a given stereo microphone system within which sound sources must be located to be reproduced within the stereo image, i.e. between the two loudspeakers of a stereo playback environment [26], [33]. (Here, we are assuming a listener position that forms an equilateral triangle with the loudspeakers.) Any sound sources located outside of the SRA will be reproduced "on" or "at" either the left or right loudspeaker, depending on which side of the SRA the sound source is located. This is an important consideration for the design of any microphone technique for which accurate sound source localization is a priority, and will be discussed in further detail for each type of stereophonic recording technique.

2.2.2 Intensity-Based Stereo Techniques

Stereophonic microphone techniques are typically divided into two types: those that rely on level differences between microphone signals, and those that rely on time differences. For intensity stereophony (also known as "coincident" techniques), two microphone capsules are arranged in a coincident manner, either housed as a single stereo microphone (e.g., Neumann SM69) or by placing the capsules of two separate microphones as close together as possible. In this technique, the acoustic intensity difference of a given sound source captured by the two capsules will determine the position of the corresponding auditory event in the stereo reproduction image [31]. M/S, X/Y, and "Blumlein" are all examples of intensity-based techniques, all of which stem from Alan Blumlein's 1931 patent [1]. Intensity-based

techniques tend to yield very well defined, stable stereo images. Here, the SRA changes as a function of microphone angle only: decreasing the physical angle between microphones increases the SRA [26].

2.2.2.1 X/Y and Blumlein

For X/Y, two matched cardioid capsules are placed coincidently in space facing the sound source, at a physical angle that can be adjusted depending on the desired stereo image: 90º is a typical starting point. The X and Y signals are routed directly to the left and right channels of a stereo mix. A variation on X/Y using bi-directional microphones set at an angle of 90º is often referred to as "Blumlein". The Blumlein technique has the added advantage of more reverberant sound capture due to the rear microphone pickup lobes.

2.2.2.2 M/S

M/S (mid/side) functions somewhat differently. The M (mid) microphone is placed on-axis to the sound source. The polar pattern of the M microphone is variable, though cardioid or omnidirectional are most often used in practice. A bi-directional microphone is placed coincident, perpendicular to the M microphone: this is the S or side microphone. Signal from the S microphone is split, either pre or post-preamp. The two resultant channels are set to equal gain and hard panned left and right, with the right channel being polarity reversed. When combined with the M microphone signal a stereo image results: $L = (M + S) / \sqrt{2}$, $R = (M - S) / \sqrt{2}$. The stereo image width and amount of ambient information can be changed by adjusting the level of the S signal. This technique is highly mono-compatible.

2.2.3 Time-of-Arrival Stereo Techniques

Time-of-arrival stereophony (known as "spaced" techniques) relies on the time delay between two microphones spaced apart (typically 20cm–100cm or greater). As a sound source moves laterally away from the centre of the sound stage, the increasing time delay between the two microphone signals results in a phantom image that moves farther towards

the left or right speakers, for a delay of sound arrival at the ears up to 1.1ms. For this technique, two forward facing omnidirectional microphones are typically used, though any polar pattern is possible. Time-of-arrival techniques tend to reproduce a greater sense of spaciousness than coincident techniques [31] due to a lack of phase coherence, and signal decorrelation between the two microphone signals [1]. As seen in section 2.1.5, stereophony based on inter-channel timing differences will have less accurate sound source localization than with inter-channel level difference. This trade-off is an important consideration for recording engineers, and must be weighed within the context of the overall aesthetics of the desired reproduced sound scene. For time-of-arrival techniques, the SRA changes as a function of microphone distance: decreasing the spacing between microphones increases the SRA, though any "linearity" in this relationship breaks down outside of a spacing less than 25cm or greater than 50cm [26], [33].

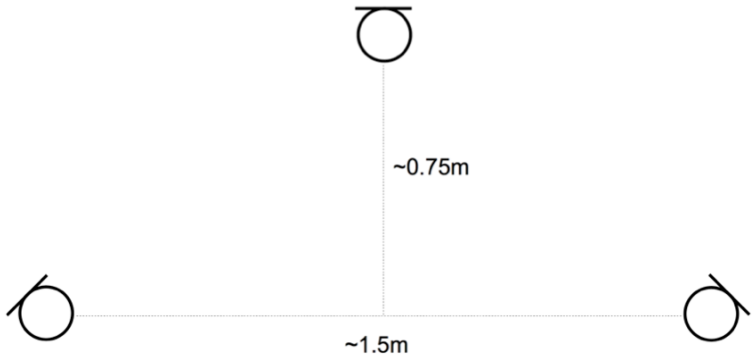


Figure 3: Decca Tree, with approximate microphone spacing

A popular variation of the spaced technique for orchestral music recording is the "Decca Tree", which uses three omnidirectional microphones (Figure 3). Spacing between microphones varies greatly depending on the recording engineer's taste. The currently known "Decca Tree" is one of several variations of "tree" shape-based microphone arrays developed by recording engineers working at Decca Records in the mid to late 1950s [34]. The technique was originally developed using three Neumann M50 omnidirectional microphones,

whose spherical capsule construction results in a polar pattern that becomes increasingly directional at higher frequencies [35]. For stereo reproduction, the centre microphone is panned centre, and the left and right microphones accordingly. A pair of widely spaced left and right outrigger omnidirectional microphones are typically added to this technique for better coverage of the entire ensemble. As Rumsey [1] points out, the use of a centre microphone to stabilize the sound image introduces further complications in terms of inter-channel phase relationships, as would the addition of outrigger microphones. These additional microphone signals, however, also introduce a great deal of flexibility at the mix stage, which may explain why the technique has become so popular for orchestral music recording, particularly for commercial film-scoring sessions where soundcheck times are typically very limited. Kamekawa et al. [36] compared a number of different techniques for recording orchestral music for 5.1 multichannel reproduction (see 2.2.6 – 2.2.8), examining various array-based techniques and combinations of frontal-sound and ambience arrays. Listeners consistently rated techniques incorporating a Decca Tree in their design highly for most subjective attributes under investigation, particularly those related to spatial impression.

2.2.4 Mixed Stereo Techniques

Mixed, or "near-coincident" or "semi-coincident" stereo microphone techniques are designed to combine both timing and level differences between microphone signals to create a stereo image. This is achieved by using a smaller spacing between microphones than typically seen with "spaced" techniques. Ideally, the resultant technique will combine the precise localization of intensity-based techniques with the superior spatial impression and low frequency transient localization of time-of-arrival techniques [31]. A popular example used extensively in broadcast is the ORTF method, developed at the Office de Radiodiffusion Télévision Française (ORTF) at Radio France. ORTF uses two cardioid capsules spaced 17cm apart, angled ±55º off the 0º centre axis (Figure 4). Based on the average physical

distance between our ears, the capsule spacing of 17cm ensures good headphone compatibility [1]. For near-coincident stereo recording systems, the SRA is changed by a combination of distance and angle between microphones [33].

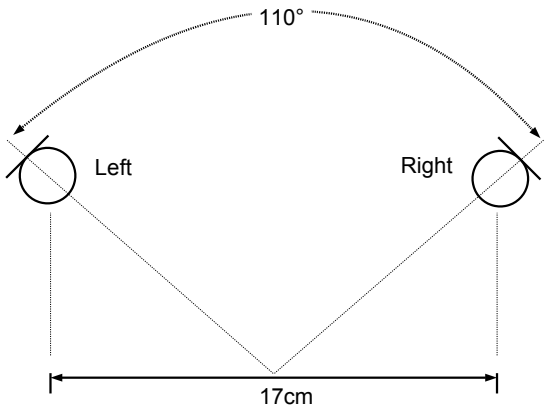


Figure 4: ORTF microphone arrangement, with cardioid capsules

2.2.5 Tools for Calculating Stereophonic Recording Systems

As seen above, the SRA for a given stereo microphone system is based on the relationships between microphone type, angle, and spacing: all of which affect inter-channel timing and level differences. Michael Williams has written extensively on this subject, publishing a series of diagrams that show the relationship between microphone spacing and angle, and the resulting SRA for pairs of microphones of a given polar pattern type (e.g. Cardioid, Hypercardioid, Omnidirectional, etc.) [33]. These diagrams, often referred to as the "Williams Curves", are collected in "The Stereophonic Zoom" [33], and can be used as a guide for designing near-coincident microphone systems, for stereo reproduction, that optimize sound source localization that is free of any angular distortion within the stereo image, as well as a consistent direct-to-reverb ratio within the reproduced sound field. Based on Wittek and Theile's [37] research into recording angles focusing on localisation curves, Wittek introduced the "Image Assistant" application in 2000, which allows the user to calculate the localization curve of any 2 or 3ch stereo microphone array [38]. More recently,

Lee et al. [39] introduced the MARRS (microphone array recording and reproduction simulator) tool for stereo microphone array design. MARRS incorporates linear image shift factors that are adaptively applied to two separate regions of 0 to 66.7% and 66 to 100% within a 60° loudspeaker base angle [39], an approach based on previous research into phantom image localization by Lee [40] and Lee and Rumsey [41]. As these tools are all optimized for stereophonic sound reproduction, however, recording engineers seeking to develop capture techniques for three-dimensional audio must still rely primarily on experimentation, previous research, and learned best practices.

2.2.6 Near-Coincident Surround Techniques

Like the techniques discussed in Section 2.2.4, near-coincident surround recording techniques are typically based on an array of closely spaced microphones, making use of both time-of-arrival and level differences between microphone signals. These techniques attempt to generate 360º accurate phantom source imaging around the listener, often using the afore mentioned "Williams Curves" [42] to define the spacing, angles, and polar patterns of microphones [1]. By relying primarily on directional microphones with relatively small spacing between them, these techniques will inherently lack the low-frequency signal decorrelation thought to be important for generating strong levels of listener envelopment [43]. One could also question the need to create reliable phantom images between side and rear pairs of loudspeakers, given that a "concert" perspective is used in most surround music recordings; i.e., music in front, ambience behind. Two typical examples are the INA-5 array, created by Hermann and Henkels, which combines three front facing cardioid microphones with two rear facing cardioids (based on the Williams curves), and OCT Surround, created by Gunter Theile, which combines cardioid and supercardioid microphones in an attempt to reduce inter-channel cross-talk [44] (Figure 5).

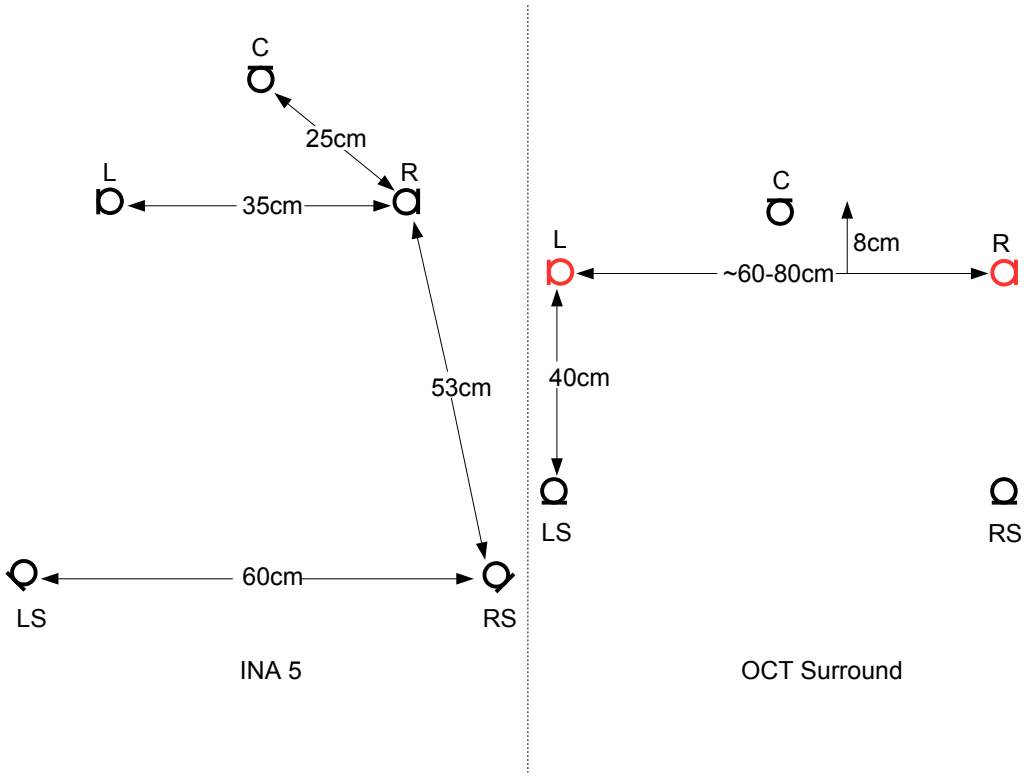


Figure 5: INA 5 and OCT Surround (black = cardioid, red = supercardioid)

2.2.7 Spaced (Front/Rear Separation) Surround Techniques

Whereas the techniques discussed in section 2.2.5 attempt to deliver 360º discreet phantom images of sources and reflections, "spaced" surround techniques focus more on creating a stable frontal sound image, and capturing a separate ambience component. In that respect, spaced techniques are optimized for "concert" perspective sound reproduction. Frontal sound capture arrays are often inspired by or literal translations of existing stereo techniques, both near-coincident and spaced, and may be optimized to capture only direct sound or some combination of direct and ambient sound information. Separate ambience arrays typically consist of several directional or omnidirectional microphones facing away from the sound stage, at a distance far enough to ensure some degree of decorrelation between front and rear microphone signals. There are an abundance of spaced arrays in use today, many of which are discussed by Imirajiri et al. [32] and Kamekawa [44]. Two popular and representative techniques are the Fukada Tree and Hamasaki Square.

2.2.7.1 Fukada Tree

Introduced by Akira Fukada, the Fukada Tree was designed to address the following requirements for multichannel sound reproduction: 1) breadth, 2) localization, 3) depth, 4) transparency, 5) spatial impression [45]. For orchestral music recording, cardioid microphones are used for the Left, Centre, Right, Left Surround, and Right Surround channels (Figure 6). The spacing and orientation of these microphones ensure primarily direct sound is reproduced from the front channels, and primarily ambient sound from the rear channels. To help coalesce the total sound image, as well as improve low frequency capture, two omnidirectional outrigger microphones are added, spaced wider than the L and R microphones, and panned somewhat wider than Left and Right [45]. To avoid significant delays between front and rear signals, the distance between front and rear microphones is kept to within 2m. A variation on this technique uses omnidirectional microphones fitted with acoustic pressure equalizers for the L, C and R channels [44].

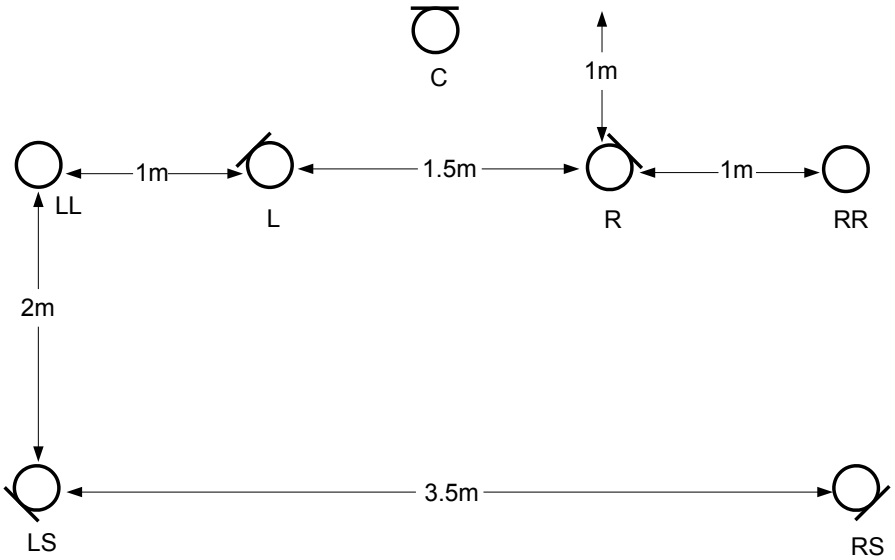


Figure 6: Fukada Tree. Spacing between microphones is based on [45] and [44].

2.2.7.2 Hamasaki Square

In [46] and [47], Hamasaki and his co-authors discuss various strategies for optimal reproduction of acoustic music and spatial impression in multichannel audio. [47] details the creation of a unique ambience capture array: the Hamasaki square, which is comprised of four laterally oriented bi-directional microphones assigned to the L, R, LS, and RS channels. Facing the null of the bi-directional microphones towards the sound stage minimizes the capture of direct sound and rear wall reflections. The spacing of the microphones in the Hamasaki square, 2-3m apart, is based on the results of subjective listening evaluations, as well as objective measurements of the minimum distance required for decorrelation between microphone signals above 100Hz [47]. The Hamasaki square can be combined with any number of frontal sound capture arrays. The square is typically placed anywhere from 2-10m further back from the main microphone array (Figure 7). This spacing is designed to ensure that minimal direct sound components are reproduced from the rear loudspeakers [47]. For orchestral music recording, Hamasaki and Hiyama [47] suggest the combination of a direct sound capture array of 5 spaced hypercardioid microphones with the Hamasaki square, which ensures a high degree of flexibility in the mix stage, especially for live recordings where the sound of the hall may change once the audience is seated. To ensure adequate low frequency capture, 2 spaced omnidirectional microphones, low-pass filtered at 250Hz, are added to the frontal sound array (Figure 7).

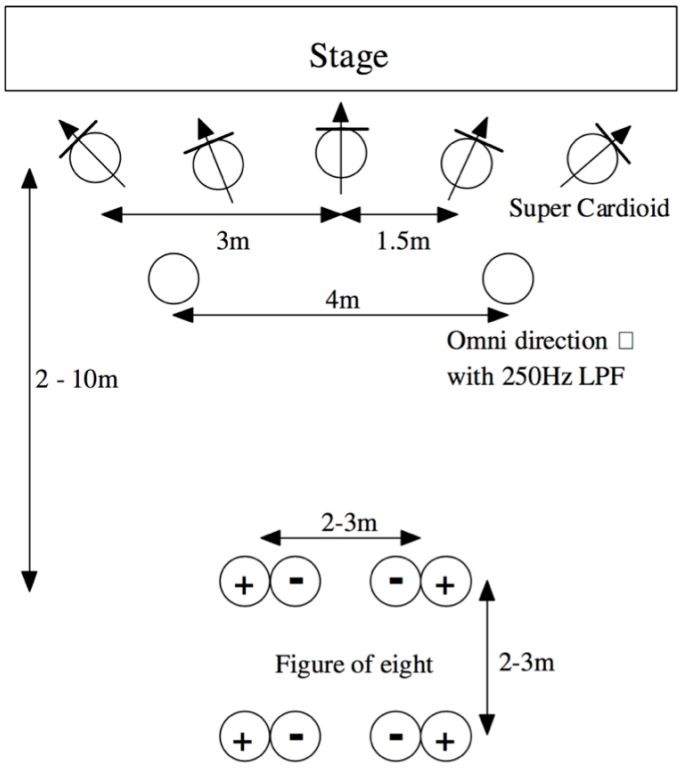


Figure 7: Orchestral recording using Hamasaki square, reproduced with permission from [47]

2.2.8 Coincident Surround Techniques

While the techniques described in Sections 2.2.5 and 2.2.6 are based on capturing discreet signals for each loudspeaker output, most coincident surround recording techniques are based on Blumlein's M/S stereophonic technique, and thus require some form of signal decoding to achieve a 5-channel surround mix. The most direct extension of the M/S technique would be "Double M/S", which has been developed extensively by Helmut Wittek and Schoeps Microphones [48]. The system uses three microphones: two back-to-back cardioids facing towards and away from the sound source and a laterally oriented bi-directional microphone, all placed coincidently in space. The simplest decoding for Double M/S would be:

Frontal Sound: L = Front M + S; C = Front M; R = Front M – S

Rear Sound: LS = Rear M + S; RS = Rear M – S

In practice, however, a simple decoding like this is discouraged by Wittek, as it would result in a great deal of inter-channel crosstalk, thereby creating a displeasing sound scene [48]. Instead, it is suggested that recording engineers make use of decoding software that can render "virtual microphone" polar patterns for each loudspeaker output. 5 hypercardioid signals are recommended for 5-channel surround [48].

2.2.8.1 Ambisonics and B-format capture

Developed in the 1970s by researchers at the British National Research Development Corporation, Ambisonics extends the M/S principle even further, adding a third bi-directional component to capture height information. The resultant signal set is called "B-Format" and is designed to capture the entire sound field at a given point in space by encoding an omnidirectional (pressure) signal (W), and three coincident bi-directional (pressure-gradient) signals (X, Y, and Z) [49]. The X, Y and Z signals represent depth, width, and height, respectively (Figure 8).

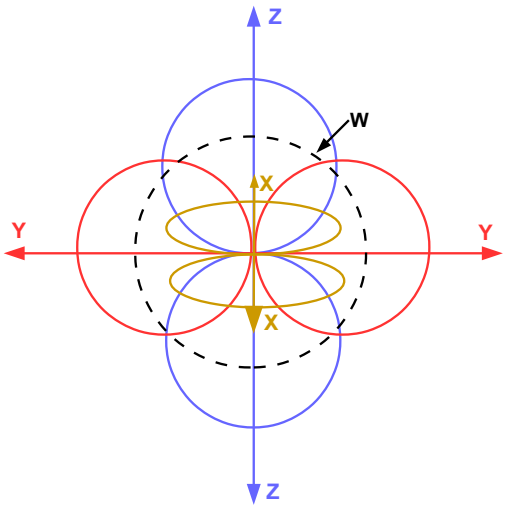


Figure 8: B-format W, X, Y, and Z signals

In theory, B-format signals can be decoded to match any number of two-dimensional or three-dimensional loudspeaker formats; Michael Gerzon's paper "Ambisonics in Multichannel Broadcasting and Video" [50] explains this process in detail. Gerzon [50]

describes a psychoacoustically optimized ambisonics decoder for broadcast that, put rather simply, first employs a phase-amplitude matrix to convert the transmission signals from C-format (a transmission compatible format) to B-format, then applies a series of corrective shelf filters to these signals that are based on human directional hearing, i.e. the head-related transfer function. The filtered signals are then fed to an amplitude matrix, which is used to derive the signals needed for whichever loudspeaker layout has been selected for playback [50]. Another possible workflow is to use a hardware or software decoder to derive "virtual microphones" with any first order polar pattern from the B-format signal set, allowing for the creation of various virtual stereo or surround coincident microphone signal sets or "arrays" at the mix stage [51].

B-format signals can be captured natively using a combination of coincidently spaced omnidirectional and bi-directional microphone capsules. In practice, "A-format" capture systems as described by Gerzon in [52] tend to be more popular. The "Soundfield ST450" microphone is a good example: it houses a coincident array of four sub-cardioid microphones, arranged as a tetrahedron [53]. A hardware processor converts sound captured as A-format signals to B-format, which can then be recorded to any multitrack system. Though widely used for location sound capture and acoustical research, Ambisonics has not proven to be a popular methodology for commercial multichannel music recording (see Section 5.1.2). One area where ambisonics is currently seeing renewed interested, however, is in sound scene capture for virtual reality, especially in conjunction with 360° video, owing to the physical convenience and positional simplicity of compact tetrahedral microphone arrays. Higher Order Ambisonics (HOA) attempts to improve the spatial resolution of first order ambisonics, theoretically increasing the size of the listener "sweet spot" while improving sound source localization [54]. This is theoretically accomplished by increasing the number of Ambisonic channels captured, which will possess increased directivity as compared with

first order B-format signals [54]. A more detailed discussion of HOA theory is beyond the scope of this thesis.

2.3 Channel-Based, Loudspeaker-Based 3D Audio Systems

Since the early 2000s, there has been a gradual proliferation of new 3D audio formats, as well as a growing interest from researchers, professionals, corporations, and consumers concerning the various applications of three-dimensional sound. Traditional surround sound formats, such as 5.1 or 7.1, only reproduce sound at ear level, whereas real-life listening experiences include sound from above and below. An ideal audio playback format would be optimized for reproduction of captured or manufactured sound in all three axes, giving a more realistic and immersive listening experience (Figure 9).

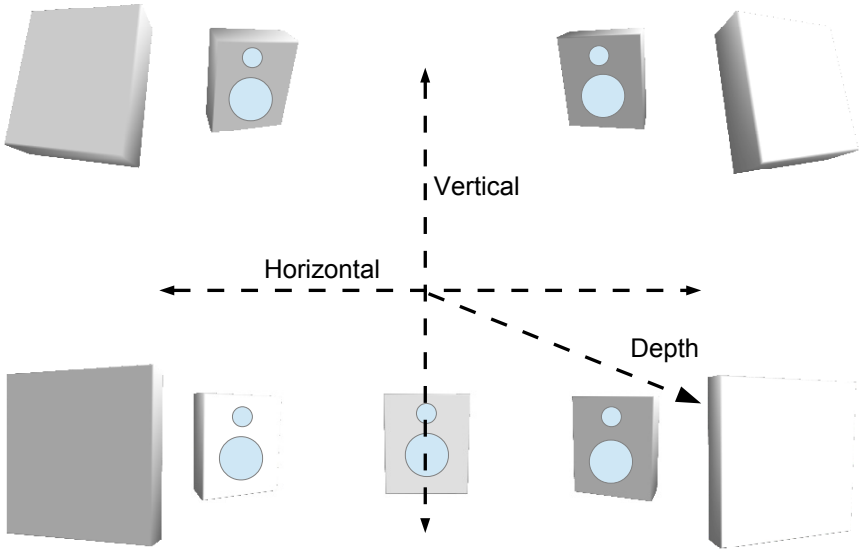


Figure 9: Simplified view of a 3D audio speaker layout

Hamasaki et al [2] compared a 3D audio format with 5.1 and 2-channel stereo for a number of different types of audio stimuli, showing that the addition of height information improves

audio reproduction for listener perception of "envelopment", "depth", "powerfulness", "presence", "realism" and "transparency." Additional studies by Hamasaki et al. [9] and Shim et al. [55] confirm that 3D audio playback increases the sensations of "presence" and "envelopment" respectively, as compared to stereo or 5.1. Kamekawa et al. [3] compared various combinations of audio playback formats (2 channel, 5 channel, and 7 channel 3D) with 2D or 3D video. Results showed that the combination of 3D video with 3D audio yielded the best listener evaluations in terms of "depth" and compatibility between visual and audio stimuli [3]. In a study comparing different height channel speaker positions, Kim et al. [4] found that the perceptual attributes most associated with the highest ranked reproduction conditions were "naturalness", "spaciousness", "continuity", "immersiveness", and "image dimension." Clearly, the expansion of sound reproduction systems to include all three axes greatly improves the overall experience of listening to recorded music, and has the potential to deliver a captured auditory scene far more immersive than established 2 and 5ch reproduction formats.

Although there are a number of different ways to capture and reproduce height information for music recordings, this thesis will focus on channel-based, loudspeaker-based three-dimensional audio formats. Binaural recording and reproduction via headphones or loudspeakers, object-based systems such as Dolby Atmos [6], and Wave Field Synthesis [56] will not be discussed, as they are not pertinent to this thesis research. Also outside the scope of this thesis are techniques that aim to generate virtual or phantom height channels, such as those using signal processing techniques based on binaural crosstalk cancellation [57] or perceptual band allocation [22].

2.3.1 Early Commercial Formats

Channel-based audio formats rely on a linear relationship between signals and loudspeakers: a stereo mix will contain two discreet channels for the left and right speakers, a 5.1 mix will

contain six discreet channels for the left, right, centre, LFE, left surround and right surround speakers, and so on. In 2001, Telarc Records released a recording of Tchaikovsky's *1812 Overture* on SACD that featured a mono centre height channel [58]. Signal from the height channel microphone (a bi-directional Sennheiser MKH-30) was high-passed with a steep filter at 180kHz. This height signal was then assigned to the LFE channel of the SACD production master, mixed together with low-passed (80Hz) signal from the left, centre, and right channels of the recording. For playback from an SACD, the LFE channel would have to be split, with signal feeding both a subwoofer and mono amp for the height speaker(s) [59]. Telarc also proposed an alternate 3D format, adding two height channels positioned at ±90º to the existing 5.1 standard [58]. Chesky Records and MDG Records (Musikproduktion Dabringhaus und Grimm) both released 3D music recordings on the DVD-A format for 2+2+2 speaker layouts: left, right, left height, right height, left surround and right surround (Figure 10) [58]. In both cases, the LFE and centre channels are repurposed as height channels. Another early attempt at a commercial 3D format was Tomlinson Holman's 10.2, which increased the number of playback channels at ear level to eight, whilst also adding two front height channels [60].

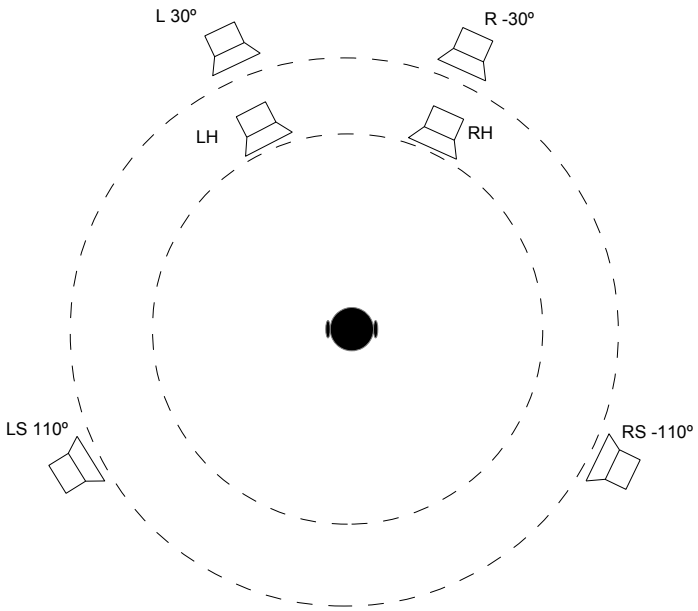


Figure 10: Overhead view of MDG 2+2+2

2.3.2 NHK 22.2 Multichannel Sound

As part of their research into ultra-high definition resolution video for broadcast (Super Hi-Vision), Japan Broadcasting Corp. (NHK) has developed and standardized 22.2 Multichannel Sound, a channel-based 3D audio format with 24 discreet speaker channels arranged in three layers: ear level, height level, and floor level (Figure 37) [8]. This format has been further standardized by SMPTE [61] and the International Telecommunications Union (ITU) [5]. 22.2 was designed with the following goals in mind:

"1) Matching of sound with on-screen picture, 2) Maintaining realism over a wide viewing area, 3) Achieving a sense of sound approach from above or below, 4) Ensuring compatibility with current theatre sound formats. [8]"

There is a great deal of published research on 22.2 Multichannel Sound, covering areas such as improved spatial impression and listener experience [9] [2], audio production techniques [62], [63], construction of 22.2 production facilities [64], flexibility of speaker

placement [65], and transmission concerns, such as downmixing to reduced playback formats [66] and the subjective loudness of 22.2 program material [67]. These areas of research, as they pertain to this manuscript, will be discussed in greater detail in later chapters, as necessary.

2.3.3 KBS 10.2 and ATSC 11.1

Samsung developed a 3D audio format also referred to as 10.2, thought with a different channel layout then Holman's [68]. The format has seven speakers at ear level at ± 0º, 30º, 90º and 135º, with two height channels at ±30º and a single above-head height channel (Figure 53). 10.2 has been adopted by Korean Broadcasting Systems (KBS) as their next-generation audio format for ultra-high definition video broadcast [5]. The Advanced Television Systems Committee's ATSC 3.0 standard recommends a similar, 11 channel format (Figure 52) for multichannel program material for immersive audio broadcast [69].

2.3.4 Auro 3D

Introduced by Wilfred Van Baelen in 2006, "Auro 3D" offers several different three-dimensional audio formats, though the most common is Auro 9.1, which combines the standard ITU 5.1 surround sound format [28] with four height channels positioned at ±30º and ±110º (Figure 54) [7]. Dozens of feature films and commercial music releases have already been mixed for Auro 9.1 and Auro 11.1 [70]. There are currently numerous home theatre receivers that support 9.1 audio playback.

2.4 Acoustic Music Recording Techniques for Three-Dimensional Audio

As the 3D audio formats discussed in Section 2.3 have been introduced and become more commonplace in audio production, various researchers and recording engineers have developed acoustic music recording techniques optimized for said playback formats. Most

techniques discussed in the literature tend to be optimized for Auro 9.1 [71], [72], [73], [74], [75], [76] . At the same time, considerable work has gone into developing capture techniques and concepts for 22.2, techniques that can often be scaled to 3D formats with less channels [9], [62], [63], [77]. As with the stereo and surround recording techniques discussed in Section 2.2, 3D acoustic music recording techniques can be largely divided into three groups: spaced, near-coincident, and coincident.

2.4.1 Spaced 3D recording techniques

As with previously discussed stereo and surround techniques, spaced three-dimensional microphone techniques capture and reproduce spatial sound information through time of arrival differences between microphones. A linear, one-to-one microphone signal to loudspeaker relationship is typically maintained, adding support microphones as necessary. Another common feature in many proposed techniques is an emphasis on distant spacing between microphones to prioritize decorrelation between microphone signals. Hamasaki and his co-authors have proposed a technique for 3D orchestral music recording, building on previously established recording concepts from 5.1 surround [9], [10]. The technique is discussed in detail in Section 5.2.

Several authors have commented on the importance of minimizing direct sound components in the height channels in order to ensure instrument or ensemble image stability at ear level while maintaining a traditional “concert” perspective. Lee [78], and Wallis and Lee [79], [80] have discussed methods for reducing or minimizing inter-channel crosstalk between main layer and height layer playback components. For height channel microphones spaced less than 2m above main layer microphones, both authors suggest the use of directional microphones, set at angles of at least 90° [78] or 105° [80]. King et al. [76] suggest the addition of acoustic pressure equalizers to omnidirectional microphones for surround and height channels, ensuring increased channel separation at frequencies above

1kHz, but maintaining an even capture of low frequency information [76], [81]. Bowles [71], on the other hand, suggests that to minimize direct sound in the height channels, hypercardioid microphones should be used, angled such that the nulls of the microphones are facing the soundstage. Hamasaki and Van Baelen describe a similar approach, suggesting upward facing hypercardioid microphones for height channel capture, placed very high above the main microphones [10]. Lee and Gribben [82] investigated the effect of height layer microphone spacing on spatial impression and listener preference. Listeners in that study showed little to no perceived difference in spatial impression or preference for microphone layer spacing ranging from 0m to 1.5m. It should be noted, however, that the stimuli in that study were created using anechoic recordings of musical instruments, reproduced either through loudspeakers and captured with a 3D microphone array, or convolved using 3D multichannel room impulse responses [82]. As such, these results may not be strongly applicable to the experience of recording actual musical instruments in acoustic spaces.

Morten Lindberg of 2L records [83] and Mick Sawaguchi of Unamas records [84] have both developed spaced 3D microphone techniques for acoustic music recording, optimized for Auro 3D 9.1, which have been implemented on multiple commercial recording sessions. Both engineers use primarily omnidirectional microphones for main and height layers, and both have experimented extensively with non-traditional, “wrap-around” ensemble reproduction aesthetics. Though Sawaguchi tends to change his approach for each session [85], Lindberg has found consistent results using a cube-like microphone array based on the speaker positions in an Auro 3D reproduction environment [83]. Hinata et al. [62], and Irie and Miura [63] have discussed sound capture workflows for 22.2, describing case studies involving live sporting events [62] and outdoor ambience capture [63]. Though not specifically related to acoustic music production, both studies describe large-scale spaced microphone methodologies, and are valuable in terms of providing practitioners and

researchers with an understanding of the complexity associated with recording content for 22.2.

Figure 11 shows a typical spaced 3D microphone array applied to a solo piano recording. For this technique, omnidirectional microphones fitted with acoustic pressure equalizers are used for the front Left, Centre, and Right channels, ensuring even sound capture across the audible frequency spectrum. Directional microphones are used for the surround and height channels, with spacing and placement angles designed to ensure decorrelation of room reflections, and minimal direct sound capture in the height channels. This technique uses simplified versions of the complex microphone arrays introduced in Chapter 4.

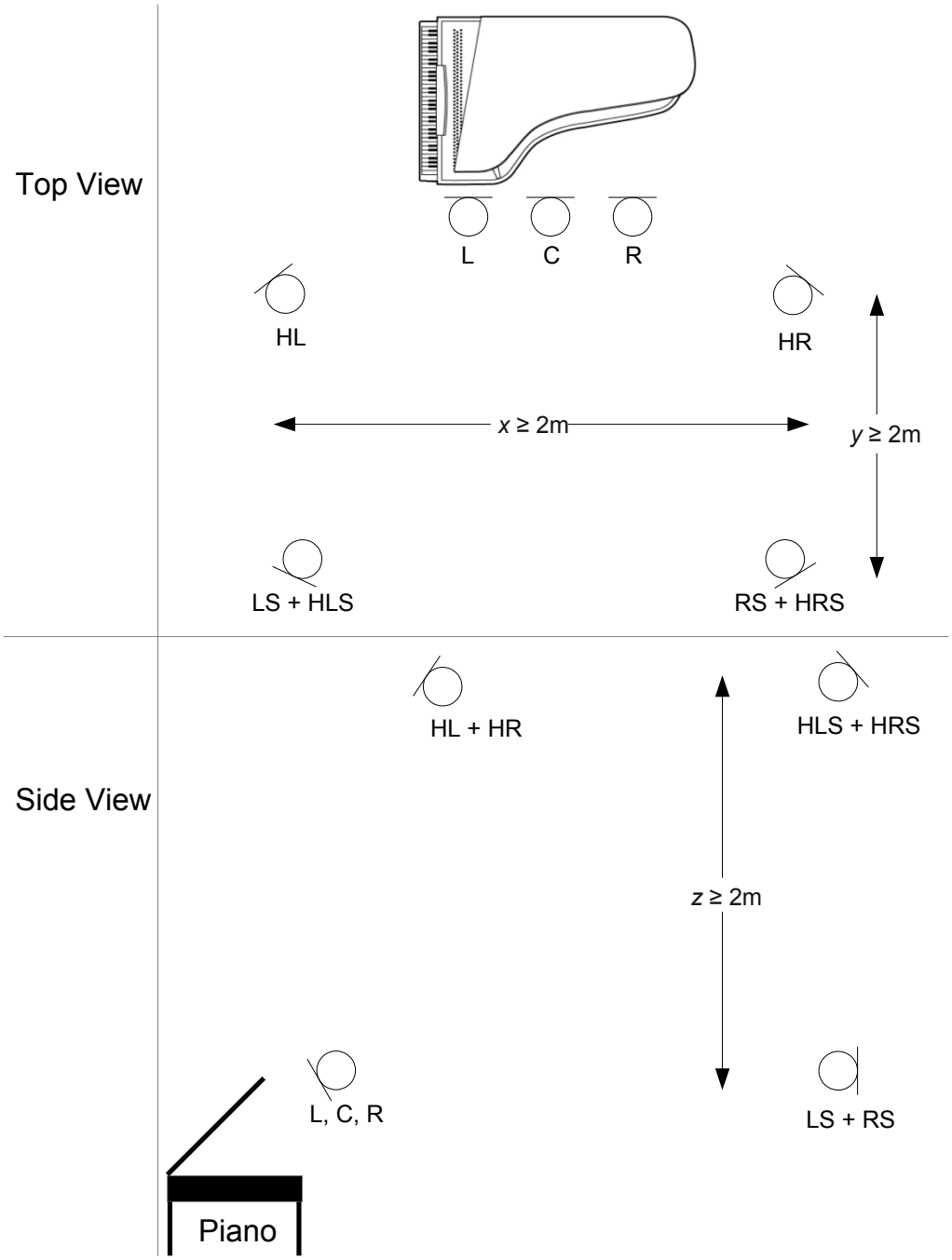


Figure 11: Example 9 channel 3D recording array. Note spacing equal to or greater then 2m for ambience capture microphones, ensuring signal decorrelation.

2.4.2 Near-Coincident 3D Recording Techniques

Several authors have proposed near-coincident three-dimensional music recording techniques, often building on previous work from 5.1 surround sound recording. Michael Williams has written extensively on his "3D Multiformat Microphone Array Design" [86], [87], [88], [89]. Unlike the arrays discussed in Section 2.4.1, Williams' array is designed to prioritize localization of direct sounds in the horizontal and vertical plane, while minimizing interaction effects between the two loudspeaker layers [86]. Williams claims that localization of phantom images between main and height layer speakers placed at the same azimuth is not possible, but *is* possible when main and height layer speakers are positioned to form an isosceles triangle [87]. Williams' array uses relatively closely spaced microphones, with a mixture of different directional polar patterns based on the author's own "William's Curves" [42] and the segmented sound field coverage schemes alluded to in Section 2.2.4 [1].

Zielinsky's "Twins Square" makes use of four Sennheiser MKH800 Twin microphones, which have a back-to-back capsule design with a discreet signal output for each microphone capsule. By placing the four microphones in a vertical square shape, 8 channels of audio are captured: front facing Left, Right, Left Height and Right Height, and rear facing Left Surround, Right Surround, Left Height Surround, and Right Height Surround. A single cardioid is added for the Centre channel [74]. Theile and Wittek have expanded "OCT Surround" (Section 2.2.4) for 3D audio, adding four upward facing hypercardioid microphones, placed either 1m above or coincident to the main layer microphones [90]. As with OCT Surround, OCT 9 (Figure 12) is designed to prioritize channel separation while maintaining a semi-spaced microphone arrangement, and uses a mixture of cardioid and hypercardioid polar patterns.

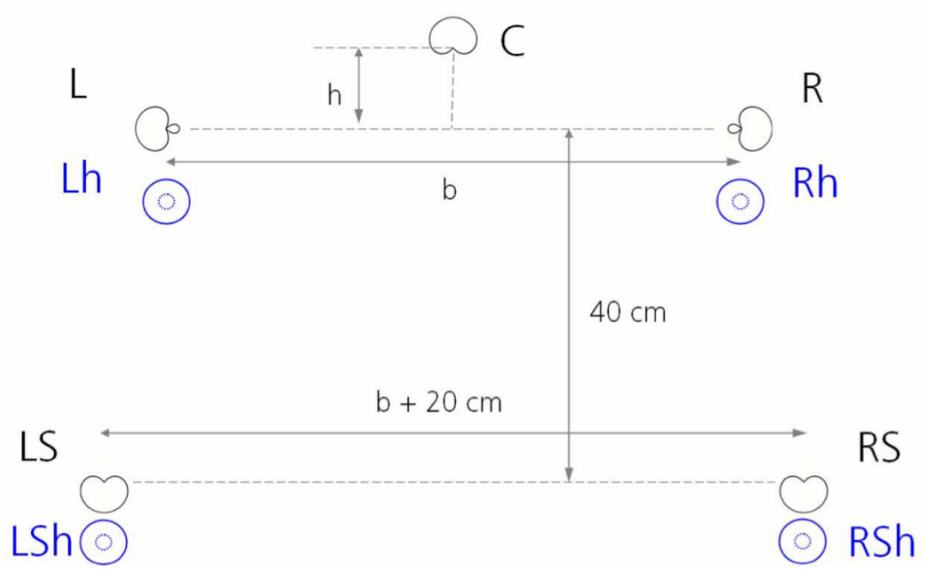


Figure 12: OCT 9. Spacing between height layer and main layer is typically 1m. Typical values for *b* and *h* are 70cm and 8cm respectively [75]. Reproduced with permission from [75].

Wittek and Theile also recently introduced "3D ORTF" [31], an ambience capture system for 3D audio and VR applications that, as the name implies, is comprised of four closely spaced ORTF pairs. Another recording system primarily aimed at ambience capture is Ono et al.'s Portable Spherical Microphone [91]. Designed for 22.2, the array consists of 22 near-coincident omnidirectional microphones, separated by acoustic baffling, and arranged to mirror the 22.2 loudspeaker layout. The use of acoustic baffling, as well as post-production signal processing, improves microphone directionality above 500Hz. A 3D recording technique that somewhat defies classification is the "Ellis-Geiger Triangle" [92]. This technique combines three spaced, coincident Double M/S arrays, where the S microphones are oriented to capture horizontal (left and right) and vertical (up and down) information, but no rear sound components. By placing the musician(s) within the spaced array triangle, the technique aims to create a more "inside the music" or perhaps "holographic" perspective.

2.4.3 Coincident 3D Recording Techniques

Most publications addressing coincident microphone techniques for three-dimensional acoustic music recording have focused on ambisonics-based recording techniques [72], [74], [93]. A notable exception is Martin et al.'s single instrument capture arrays for 3D audio [94]. "Double-XY" (Figure 13) combines a traditional XY cardioid pair with a 2nd, vertically oriented cardioid pair. "M/S XYZ" (Figure 13), combines a standard M/S array with a vertically oriented bi-directional microphone [94]. In both cases, all microphone capsules are placed as close to coincidently as is physically possible. Though not designed to capture a complete sound scene, as there is no information captured for the rear channels, Martin et al. have shown that these techniques create sonic images with well-defined horizontal *and* vertical extent, which is highly valuable for achieving realistic or hyper-realistic recreations of acoustic instruments [95].

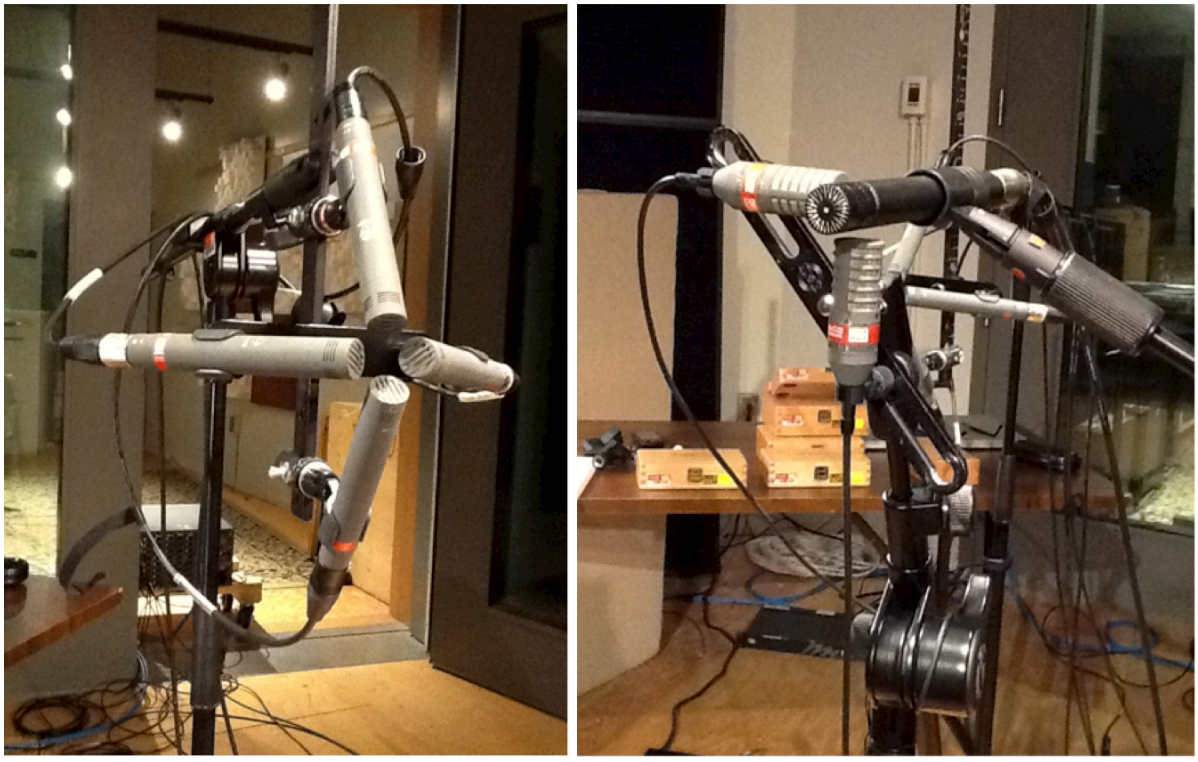


Figure 13: Double XY (Left) and M/S XYZ (Right) [95]. Reproduced with permission.

2.4.3.1 Ambisonics for 3D Acoustic Music Recording

Ambisonics sound capture has already been discussed in Section 2.2.7. Here we simply extend the principles to include sound reproduction in the vertical plane. Geluso [72] and Ryaboy [74] both discuss a native B-format approach for acoustic music recording: Double MS+Z. As the name implies, a vertically oriented coincident bi-directional microphone is added to a standard Double MS array. For ease of coincident spacing, both authors suggest the use of a Sennheiser MHK800 Twin microphone to capture both front and rear M components [72], [74]. As with coincident surround techniques, 1st order A-format capture systems such as a Soundfield microphone, or higher order spherical microphones such as the Eigenmike can also be used for coincident 3D sound capture. Bates et al. [96] provide a useful overview of several different commercially available 1st order and higher order ambisonics-based microphone systems, comparing them in terms of timbre and directionality. Ikeda et al. discuss different sound capture methods for orchestral music with 3D video, including several spherical HOA microphones in their recording, all placed within the orchestra, giving the listener a non-traditional perspective [97].

2.5 Subjective Evaluation and Analysis of Reproduced Sound

This review will focus on methodologies for subjective audio evaluation and analysis, which tend to be the norm for research focused on evaluating music recording and reproduction techniques. Most work concerning objective methodologies for multichannel audio evaluation has focused on 5.1 surround sound [98], [99], [100], [101], [102], [103], [104], and as such, may not be entirely applicable to 3D audio. With a few exceptions, the objective measures referenced above tend to focus on detecting impairments in audio created by multichannel audio codecs or consumer playback systems, and have not been used to evaluate recording techniques.

2.5.1 Subjective Evaluation of Multichannel Audio Stimuli

2.5.1.1 ITU-R BS.1116-1

ITU-R recommendation BS.1116-1 describes a methodology for subjective listening tests designed to detect subtle differences between audio stimuli [105]. ITU-R BS.1116-1 recommends using the "double-blind triple-stimulus with hidden reference", also known as a "triad" test. As an example, in the experiment described in Section 6.4, subjects are asked to compare stimuli randomly assigned labels "A", "B", and "C", and to determine which two are the same. ITU-R BS.1116-1 contains useful recommendations in terms of stimulus and test length, type and selection of subjects, suitable sound reproduction environments, and statistical models that can be applied to acquired data. Another key point from this set of recommendations is the use of "critical testing material": program material that stresses the playback system under test:

"It must be empirically and statistically shown that any failure to find differences among systems is not due to experimental insensitivity because of poor choices of audio material, or any other weak aspects of the experiment, before a "null" finding can be accepted as valid. […] The artistic or intellectual content of a programme sequence should be neither so attractive nor so disagreeable or wearisome that the subject is distracted from focusing on the detection of impairments. [105]"

2.5.1.2 Evaluating Multidimensional Perceptual Spaces

ITU-R BS.1116-1 recommends the following subjective sound attributes for evaluation of multichannel audio: "Basic audio quality", "Front image quality", and "Impression of surround" [105]. Multidimensional terms such as these may be adequate for making general statements or impressions about the overall performance of a given stimulus, but they tell us little about what specific aspects of the sound are positive or negative, or which of these aspects contribute most to perceived overall quality or general preference. Bech introduced

"to the audio community, the basic principles of an experimental method for evaluation of multidimensional auditory stimuli. [106]" Bech and Zacharov expand on these principals in their authoritative *Perceptual Audio Evaluation – Theory, Method and Application* [107]. Bech [106] presents a methodology that is now common practice for evaluating multichannel audio recordings and systems, emphasizing the multidimensional nature of reproduced sound; similarities or differences between different audio stimuli can rarely be defined by a single attribute. Bech details a "conceptual model of human perception of multidimensional auditory stimuli" [106], roughly summarized below:

1. A sound field exists, which has a certain number of physical variables or objective dimensions
2. This sound field is "perceived and transformed by the human hearing system […] into auditory attributes. [106]"
3. The subject assigns each auditory attribute a certain magnitude of impression, based on the sensorial strength of the stimuli, as well as contextual and cognitive effects
4. These individual impressions combine to create a total auditory impression

[106] and [107] provide methodologies related to descriptive analysis, selection and training of experimental subjects, experimental design, appropriate statistical models for acquired data, presentation of results, and techniques for exploring the relationships between individual auditory attributes and overall preference. All this combines to create an experimental strategy designed to:

1. Identify individual auditory attributes
2. Devise methods to measure the magnitude of each attribute
3. Establish a relationship between auditory attributes and total auditory impression [106]

2.5.2 Audio Attributes for Spatial Sound Evaluation

Ideally, perceptual audio attributes should be well defined, unambiguous and commonly agreed upon, broadly applicable to a wide range of applications, as close to unidimensional as

possible, and relate to physical measures when possible [106], [108]. Rumsey [109], [110], Berg and Rumsey [111], [112], [113], [114], Zacharov and Pederson [108], Zacharov and Koivuniemi [115], [116], [117], Choisel and Wickelmaier [118], Kamekawa and Marui [119] and others have aimed to collect, define, evaluate, verify, and codify perceptual attributes specific to spatial audio evaluation, as well as examine these attributes for correlation between themselves, and more multidimensional terms, such as "overall quality" or "preference". Le Bagousse et al. provide an excellent summary of the key work in this area, reviewing different techniques used by various researchers to elicit spatial audio attributes [120]. Their lexicon study [120] classifies sound attributes for audio quality assessment by four broad categories: "Defects" (noise, distortion, hum, hiss, disruption, etc.), "Space" (depth, reverberation, width, distance, localization, spatial distribution, envelopment, immersion, etc.), "Timbre" (brightness, tone colour, colouration, clarity, richness, etc.), and "Quality" (homogeneity, stability, fidelity, naturalness, etc.). Zacharov and Pederson used semantic text mining to arrive at the extensive lexicon of common auditory attributes shown in Figure 14 [108].

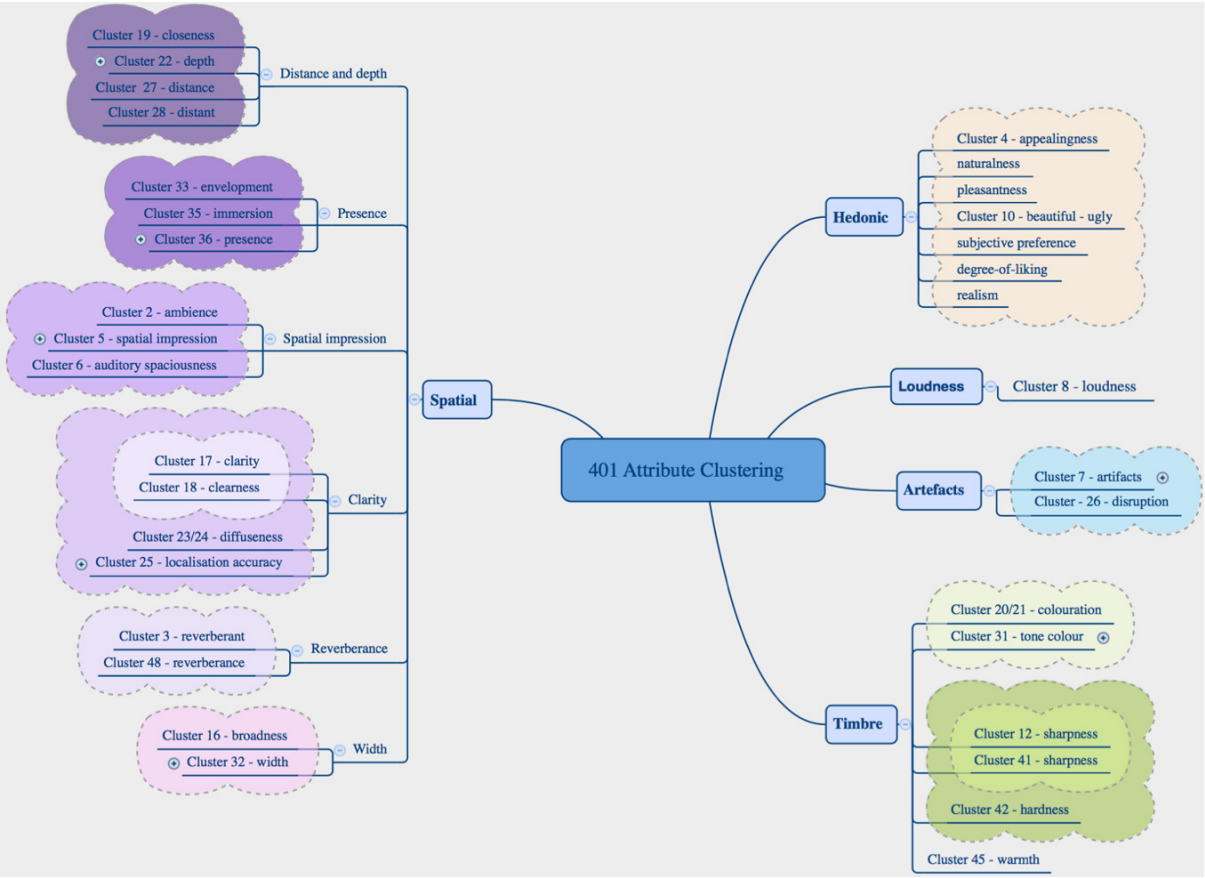


Figure 14: Overall structure of attribute clusters, reproduced with permission from [108]

2.5.2.1 Correlation of Attributes

The large number of studies designed to identify or elicit valid spatial audio attributes has resulted in an equally large body of terms to choose from when conducting studies in perceived quality of multichannel sound. Lexicon studies such as those seen in [108] and [120] show us how these terms group together, which helps narrow the selection process. Knowing what, if any, correlation exists between various audio attributes allows researchers to further eliminate similar or redundant terms when preparing an experiment. Also of value is knowing what perceptual attributes most relate to overall listener preference. Examining 12 different spatial audio attributes, Berg and Rumsey found the strongest correlations between "naturalness and presence", and "preference and enveloping", noting: "This analysis also verifies the relatively strong interrelation between *envelopment* and the attributes expressing

naturalness and a feeling of presence. [111]" This strong correlation between "naturalness" and "presence" agrees with findings from work by Guastavino and Katz [121]. Choisel and Wickelmaier found that both timbral and spatial attributes are important predictors of overall listener preference [118], echoing similar findings from Rumsey et al. [122]. Shim et al.'s study on perceived quality between different multichannel reproduction systems found a particularly strong correlation between "Listener Envelopment" and "Apparent Source Width", which is understandable, given both are known to be strongly tied to lateral reflected sound energy [123], [124]. Much of this suggests that for multichannel sound reproduction, achieving a sound scene with good timbre and spatial impression results in a more "natural" listening experience, which is more likely to be preferred by listeners.

2.5.3 Scene-Based Analysis of Multichannel Audio Reproduction

2.5.3.1 Auditory Streaming

In *Auditory Scene Analysis*, Bregman writes "The job of perception, then, is to take the sensory input and to derive a useful representation of reality from it. [125]" In the case of human hearing, a great deal of perceptual information is derived from complex and potentially confusing sensory information. The auditory system's only available information is the vibrations at our two ear drums [125], which is a combination of the spectral and temporal information of all sound events surrounding us. And yet, from this jumble of information we can discern and identify the various sound sources that make up a complete auditory scene. This is accomplished by the streaming of auditory information by the brain. As Bregman explains: "An auditory stream is our perceptual grouping of the parts of the neural spectrogram that go together. […] the goal of scene analysis is the recovery of separate descriptions of each separate thing in the environment. [125]" The brain takes the mass of sensory information processed by the inner ear, and clusters information that is temporally and timbrally related to create different streams that each represent a different auditory event,

or “sound image”. Auditory signal components combine to form a common object, which can be recognized and labeled by the brain for future reference [1]. Going from the micro to macro level, Letowski writes “Several coexisting images also can be merged together into a more generalized picture of the acoustical environment surrounding the listener. [126]”

Griesinger believes that perceptual streaming is key to understanding the multidimensional concept “spatial impression” [43]. He suggests that auditory information is split into multiple foreground streams and a single background stream [43]. When a direct sound is continuous and difficult to separate into multiple sound events, it perceptually combines with the reflected energy in a room, resulting in a sense of envelopment that is connected to the sound source. Griesinger calls this “continuous spatial impression”. When the sound events are separable, early reflections (within 50ms) perceptually fuse with the sound event, creating a single foreground stream, and a sense of dimensional broadening of the sound source that Griesinger calls “early spatial impression”. Reflected sound energy later than 50ms forms a separate, background stream. If this “background spatial impression”, which is perceived in the space between musical notes or spoken words, contains a high level of spatially diffuse sound energy, interaural time and level differences result in a strong sense of envelopment [43].

2.5.3.2 Scene-Based Paradigm for Spatial Audio Evaluation

Bregman’s perceptual model, wherein the brain is analyzing sound on a subconscious level, can be extended to the world of spatial audio evaluation. Here the listener is making conscious judgments about various aspects of the complete sound scene. Rumsey proposes a “scene-based” paradigm for subjective evaluation of spatial audio, which “requires that the elements of the reproduced scene be grouped according to their function within the scene, at levels appropriate to the task. [109]” This approach is concerned with evaluating the scene through specific descriptive attributes, rather than basic audio quality or preference [109].

Rumsey's proposed paradigm is hierarchical in nature, utilizing a set of auditory attributes that builds from the micro to the macro level. An example would be the commonly used attribute "width" (Figure 15). In a scene-based paradigm, we would want to define multiple types of width: "individual source width", "ensemble width", "environmental width", and perhaps even "total scene width" [109]. Rumsey's work [109] provides an extensive list of auditory stream-specific attributes that allow for the evaluation of very specific components of multichannel sound reproduction. Understanding how these components combine within different weighting and contextual schemes is key not only for evaluating reproduced sound, but also for creating content specific to two and three-dimensional audio systems.

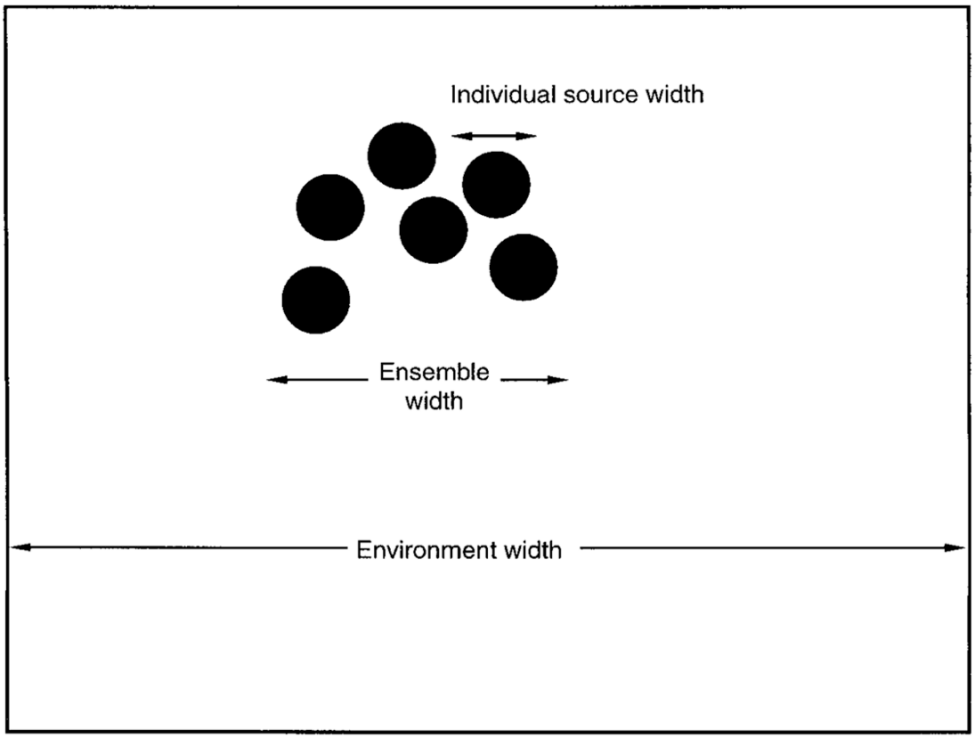


Figure 15: Width attributes, from micro to macro, reproduced with permission from [109]

2.6 Other Areas of Consideration

2.6.1 Room Acoustics

Architectural acoustics is an area of study with a long history of research, going back to Sabine's development of reverberation time measurement in the late 1800s [127]. A detailed summary or discussion of architectural acoustics is beyond the scope of this thesis. It is worth noting, however, that as the amount and distribution of acoustic information reproduced by audio systems increases, concepts from room acoustic measurement and evaluation become increasingly relevant to music recording and reproduction. When considering the number and location of points of sound reproduction in an audio format such as 22.2 as compared with stereo or 5.1 surround sound, it is implicitly understood that 3D audio systems have the potential to deliver a more detailed and nuanced reproduction of the sound field of a given space.

Orchestral music tends to be recorded in large acoustic spaces: either concert halls or rooms with similar acoustic qualities, such as scoring stages, ballrooms, or large churches [128]. A number of objective and subjective techniques have been developed for evaluating the acoustics of a room, most of which are summarized effectively by Gade [129] and Ando [130]. Objective measurements of room acoustics tend to be based on analysing a room's "impulse response" (IR), which Gade defines as: "The basic source of information regarding the audible properties of the sound field in a room… [129]"

"Spatial Impression" is a concept that features prominently in Chapters 3 – 6 as a means of evaluating or distinguishing between different three-dimensional audio stimuli. In the study of concert hall acoustics, spatial impression is normally divided into two different but related areas: "apparent source width" (ASW) and "listener envelopment" (LEV). ASW refers to the perceived width of the sound source, thought to be primarily influenced by the level of lateral reflected sound energy arriving within the first 80 ms, i.e. early

reflections [129]. The Lateral Energy Fraction (LEF) has been found to be a good measure of ASW in concert hall acoustics. An IR is captured using both an omnidirectional microphone, representing total sound energy, and a bi-directional microphone positioned perpendicular to the sound source, representing lateral energy. The ratio of these two signals, within the first 80ms, and averaged across four octave bands (125Hz – 1000Hz) correlates well with perceived ASW in a room: higher LEF values indicate a wider ASW [129].

More germane to the research in this manuscript is listener envelopment, which in concert halls is mainly determined by the spatial distribution and level of late reflections [129]. Hanyu and Kimura [124] have conducted numerous experiments investigating LEV. Based on their own work, as well as previous work by Bradley and Soulodre [131], Furuya et al. [132] and Morimoto et al. [133], Hanyu and Kimura have arrived at several conclusions regarding listener envelopment:

i. LEV increases as lateral reflections increase

ii. The influence of reflections arriving from the front is not zero

iii. The contribution of individual reflections to LEV depends on the arrival direction of other reflections

iv. LEV increases if there is adequate spatial balance in the energy of arriving reflections. [124]

Bradley and Soulodre [131] have proposed Late Lateral Sound Level (LG) as being an effective objective measure of LEV in concert halls. Hanyu and Kimura have proposed an alternative measure: SBT_s, which aims to quantify the spatial distribution of reflections using the Centre time T_s for each direction [124].

Another objective measure of concert hall acoustics worth noting here is Interaural Cross Correlation (IACC), which has been shown to correlate with both ASW and LEV.

Using a binaural "dummy head" microphone to capture the IR of a given space, IACC aims to quantify the dissimilarity of signals at the two ears [129]. In that respect, it can be seen as an objective measure of the fluctuations in the interaural time delay and interaural intensity difference at the two ears, which Griesinger believes to be a key factor in contributing to strong levels of LEV [43], [134]. Studies by Power et al. [135], Choisel and Wickelmaier [99], and Mason and Rumsey [136] have all shown IACC or modified IACC measurements to be a good predictor of spatial impression in multichannel audio reproduction.

2.6.2 Directional Characteristics of Musical Instruments

Another area of acoustic scholarship worth noting here is Meyer's work cataloguing the tonal and directional characteristics of musical instruments [137]. *Acoustics and the Performance of Music* shows how most musical instruments do not radiate sound in all directions with equal intensity, but instead exhibit more or less pronounced directional effects. Having measured the various instruments of the symphony orchestra in an anechoic chamber, Meyer finds that for most, overall sound strength changes with direction, but also spectrum and thus the tone colour [137]. This knowledge is important for recording engineers creating content for *any* reproduction medium, serving as a guide to identify optimal points of sound capture for a given instrument, i.e. areas where an aesthetically desirable range of tone colour and timbre can be captured by the microphones. This becomes especially important when implementing complex, closely positioned microphone arrays designed to capture many facets of an instrument's tonal and timbral characteristics, such as the recording techniques described in Chapter 6.

3 PRELIMINARY EXPERIMENTS

This chapter is comprised of two short publications that detail preliminary three-dimensional music recording experiments. The 2nd paper, an exploration of microphone polar patterns for height channels, follows directly from the results of the first paper. It should be noted that at the time the studies detailed in Chapter 3 were undertaken (November 2014 – April 2015), a number of the papers discussing the practical implementation of three-dimensional music recording techniques discussed in sections 2.2.6 – 2.2.8 had yet to be published. This is reflected in the somewhat conservative language regarding the progress of research into 3D recording found in the introductory sections in Chapters 3.1 and 3.2. For Chapters 3 – 6, all loudspeaker positions and corresponding microphones follow the SMPTE [61] naming convention for 22.2, which are described below in Table 1, and in Figure 37 (Chapter 4).

Table 1: Channel Naming and Abbreviations for 22.2 Multichannel Sound, as per [61]

Channel	Full Name	Abbreviation	Azimuth
1	Front Left	FL	-60°
2	Front Right	FR	60°
3	Front Centre	FC	0°
4	Low Frequency Effects 1	LFE 1	n/a
5	Back Left	BL	-135°
6	Back Right	BR	135°
7	Front Left Centre	FLc	-30°
8	Front Right Centre	FRc	30°
9	Back Centre	BC	180°
10	Low Frequency Effects 2	LFE 2	n/a
11	Side Left	SiL	-90°
12	Side Right	SiR	90°
13	Top Front Left	TpFL	-60°
14	Top Front Right	TpFR	60°
15	Top Front Centre	TpFC	0°
16	Top Centre	TpC	0°
17	Top Back Left	TpBL	-135°
18	Top Back Right	TpBR	135°
19	Top Side Left	TpSiL	-90°
20	Top Side Right	TpSiR	90°
21	Top Back Centre	TpBC	180°
22	Bottom Front Centre	BtFC	0°
23	Bottom Front Left	BtFL	-60°
24	Bottom Front Right	BtFR	60°

3.1 Exploratory microphone techniques for three-dimensional classical music recording

Abstract

At McGill University's Redpath Hall, a conventional stereo recording array was augmented with additional microphones in both the horizontal and vertical planes, yielding a fourteen-channel, three-dimensional sound recording, featuring seven discrete height channels. Based on existing multichannel recording models, microphone placement was designed to prioritize listener envelopment. Preliminary evaluations of the recordings by researchers at the Graduate Program in Sound Recording at McGill University found that these 3D recordings have an increased sense of envelopment and realism as compared to traditional 5.1 surround sound. Several areas have been identified for further investigation through future recordings and listening tests.

3.1.1 Introduction

Common music playback formats, stereo and 5.1 surround, offer a decidedly two-dimensional listening experience, recreating sound in the horizontal plane only. Three-dimensional audio formats, such as Japan Broadcasting Corp. (NHK)’s 22.2 Multichannel Sound (22.2) [8] offer the potential to recreate musical performances with an unprecedented sense of depth and realism (see: section 2.3).

3.1.1.1 Historical Context

Classical music and the concert hall are strongly linked: the venue creates sonic reflections and reverberation that envelop the listener, while also informing the musicians' performance [138], [139]. Numerous recording techniques have been developed to capture an ideal and realistic balance of music and reverberation for stereo and/or 5.1 surround (see: section 2.2). However, comparatively few such techniques have been developed and evaluated for 3D audio formats, leaving a large gap in the current knowledge base of music production.

Immersive content created for 22.2 can be downmixed or remixed for other common multichannel formats [8], [77], making it an ideal research tool for developing versatile 3D recording techniques. However, the small number of publications that specifically address recording for 22.2 have mostly dealt with topics such as live sports broadcast [62] or ambience capture for television specials [63].

3.1.1.2 Motivation

Previous research has shown that the addition of vertical "height" channels to multichannel audio improves listener impression for a number of subjective attributes, such as depth, presence, envelopment, intensity, and naturalness [2], [3], [4]. As a pilot study, an experimental fourteen-channel double layer microphone array was designed and implemented, optimized for three-dimensional music capture. The focus of this investigation was to determine what sonic information yields the best listening experience in terms of envelopment and realism.

3.1.2 Methodology

Recordings took place in McGill University's Redpath Hall (Figure 16). The hall measures 27.8m in length by 13m wide, with a height of 13.35m; the RT60 is approximately 1.7s. The musicians, a small baroque ensemble, were setup in a “quasi-concert" positioning that was determined ahead of time by the group leader and recording producer as being one that would deliver ideal scene depth and horizontal ensemble imaging.

3.1.2.1 Recording Methodology

Typical of many commercial recordings, the sessions were split over two days. For the first day, the recording team focused only on stereo capture (FLc and FRc). The main stereo array consisted of two DPA 4006TL omnidirectional microphones, spaced 60cm apart, 2.51m high (from the floor to capsules), approximately 1.15m from the ensemble leader (violin). Spot microphones were also placed near each instrument for additional detail: *Violin:* Royer SF-

24, *Cello:* Neumann m149, *Theorbro:* Schoeps MK4, *Harpsichord:* two Schoeps MK21, and *Portative Organ:* two Neumann U87i (Figure 17).

On the second day of recording, additional microphones were added as surround and height channels. At the same height as the main pair were added two cardioid DPA 4011s facing outward ±90° (SiL and SiR) and two more DPA 4006s fitted with 50mm Acoustic Pressure Equalizers as surrounds (BL and BR) (Figures 18 and 19). For the height array, a combination of four omnidirectional Neumann KM 183s (TpFL, TpFR, TpBL, TpBR), and two DPA 4011s (TpSiL, TpSiR) were used, all at a height of 3.72m. An additional DPA 4011 (TpC) was placed in the centre of the height array, facing upward, capsule height 4.07m (Figures 18 and 19). All microphones were routed to a Merging Horus audio interface. Audio was recorded to a Pyramix workstation at 96kHz/24bit resolution. Monitoring took place in a nearby control room outfitted with eight Focal Audio loudspeakers, arranged for playback as: FLc, FC, FRc, BL, BR, TpFL, TpFR and TpC.

3.1.2.2 Methodology of Microphone Placement

A great deal of literature exists on microphone techniques for 5.1 surround, many of which could easily be adapted to expanded 3D recording arrays (see: Sections 2.2, 2.4). Microphone positions for the surround array were chosen largely based on the recording producer's previous experience in recording classical music for 5.1 reproduction. Recording techniques by Hamasaki [47] and Fukada [45] were also re-examined. Microphone type and placement choices for the height array were largely experimental, focusing on capturing a variety of sonic information. The decision to include "lateral" ±90° microphones in both the main layer and height arrays was based primarily on previous research by Hanyu and Kimura [124] showing the importance of lateral reflections for listener envelopment.

Figure 16: Redpath hall during recording sessions. Ensemble in top left.

Figure 17: Baroque Ensemble with main stereo pair and spot microphones

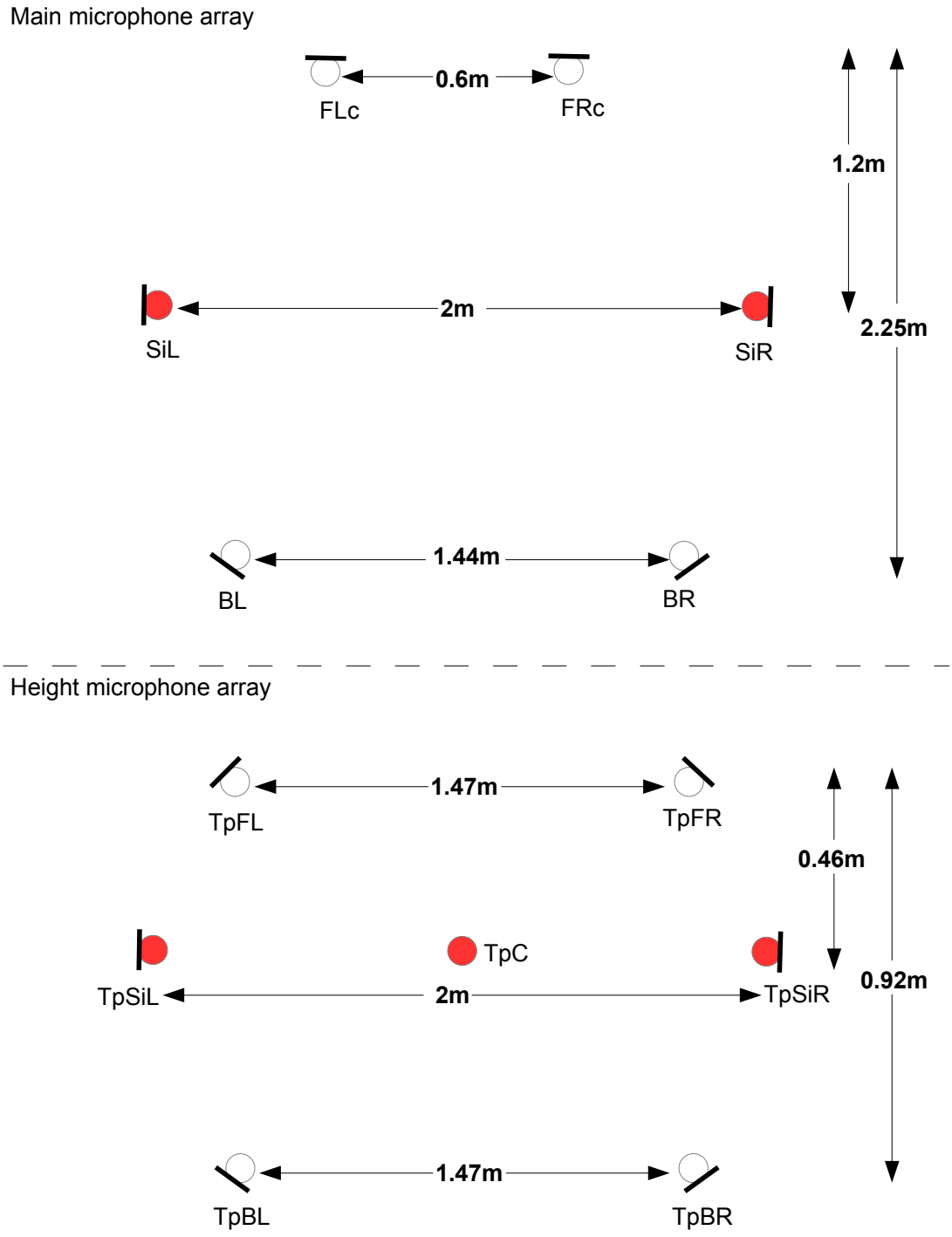


Figure 18: Microphone arrays, as seen from above. White microphones are omnidirectional, red microphones are cardioid. Main layer array height: 2.51m; height layer array height: 3.72m (4.07m for TpC).

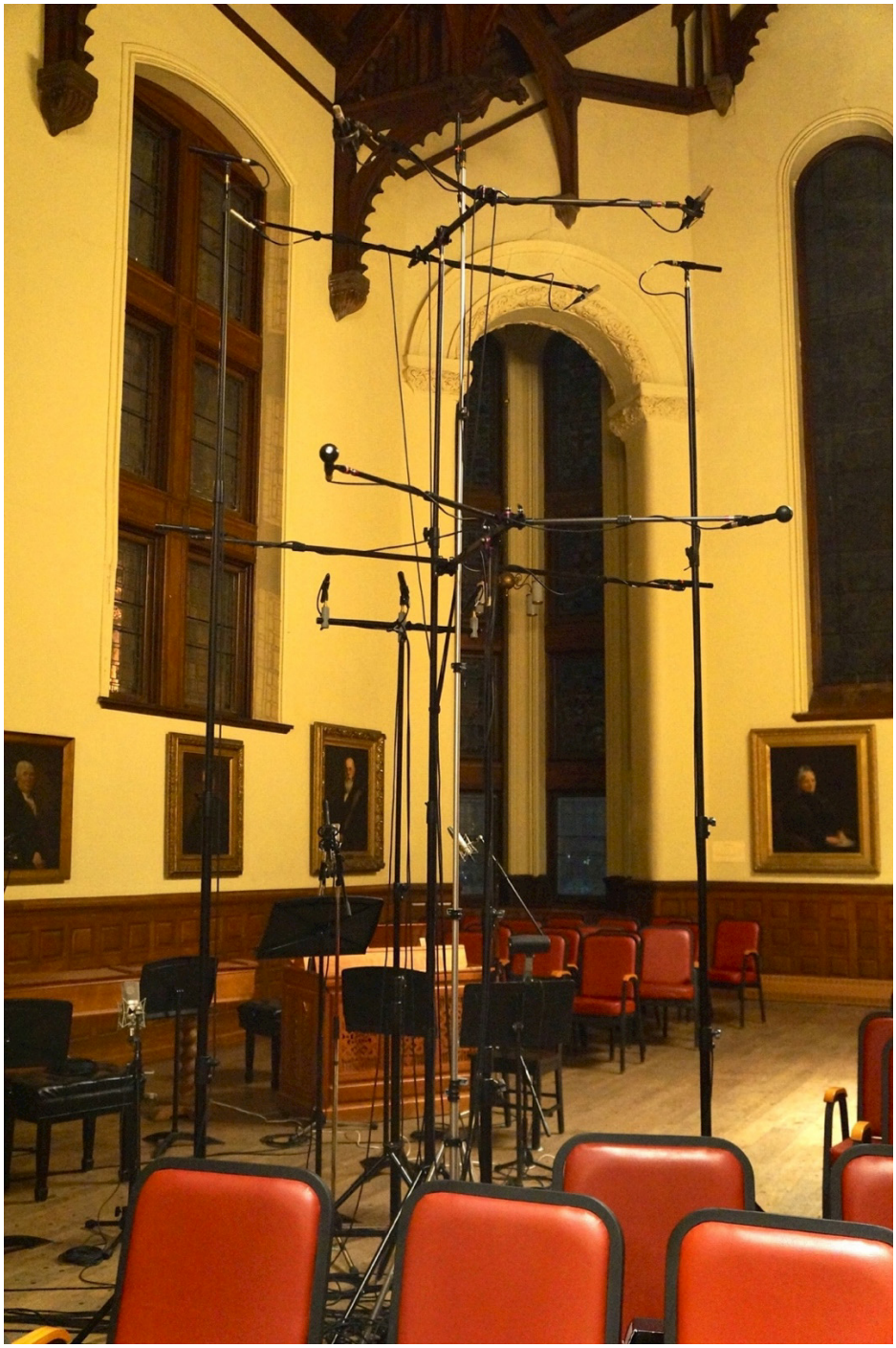


Figure 19: Microphone arrays setup for recording

3.1.2.3 Playback Environment

Mixing and listening sessions of the recordings took place in McGill University's Studio 22 (Figure 29), a purpose-built multi-channel listening room with 28 channels of discreet audio playback via Musikelectronic Geithain *M-25* two-way loudspeakers, powered by Flying Mole digital amplifiers. The loudspeakers are arranged for reproduction of both 22.2 Multichannel Sound and Auro 3D 9.1. The room's dimension ratios and reverb time fulfil ITU-R BS.1116 requirements [105], [140] (see Section 6.4 for additional details on Studio 22).

3.1.2.4 Informal Evaluation of Recordings

The recordings were evaluated by four faculty members and thirteen students from the Graduate Program in Sound Recording, during a series of informal listening sessions. Each participant was seated in Studio 22's listening position and presented with a Pyramix 9 session with which they could listen to completed 3D mixes of several different pieces performed by the baroque ensemble. Using a set of VCA faders within the Pyramix mixer, participants could also add, remove, rebalance, or solo various elements of the mix if they so desired. No time limit was set for this activity. General feedback and impressions were provided to the mixing engineer (the author) verbally, who was in the studio with the participants. The evaluated mixes featured the following playback channels:

Main Level: FLc, FC, FRc, SiL, SiR, BL, BR

Height Level: TpFL, TpFR, TpSiL, TpSiR, TpBL, TpBR, TpC

3.1.3 Results and Discussion

All listeners agreed that the addition of discreet height channels to the main playback layer yielded a significant increase in the envelopment and realism of the recordings. There was also a consensus that the four height channels that used omnidirectional microphones (TpFL, TpFR, TpBL, TpBR) likely contained too much direct sound and not enough decorrelated or

diffuse sonic information. This was especially noticeable in the 2 front height channels, which tended to "pull" or "smear" certain elements of the ensemble image upward. It was also observed that the height channels needed to be balanced strongly in the mix to create an adequate level of listener immersion. The side height channels were deemed to have the best mix of decorrelated and diffuse sound, which is understandable, given that those microphones were cardioid pattern and facing away from the ensemble. The main level side channels were also observed to increase envelopment, though they too suffered from the problem of containing not enough diffuse, decorrelated information, and were at times somewhat disruptive to the frontal sound image.

Oode et al. [141] investigated the effect of the number and arrangement of vertical loudspeakers on listener perception of spatial uniformity within a sound field. Results of that study demonstrated the importance of an "above the head" centre loudspeaker for maintaining spatially uniform sound, particularly when using a reduced number of height channels as compared with 22.2. In the current study, several listeners who took the time to perform a more detailed analysis of the individual components of the height channels observed that when only adding pairs of height channels (e.g., TpFL and TpFR, or TpSiL and TpSiR) to the main layer channels, the addition of the Top Centre channel increased the overall cohesion and envelopment of the sound scene. This adds credence to Oode et al.'s findings, and suggests that a Top Centre channel would be a valuable addition to smaller-scale 3D audio reproduction formats. Auro 3D [7], for example, already includes a TpC channel in all their expanded formats (10.1, 11.1, 13.1).

Results from previous research [142], [10], [2], and this pilot study suggest that the addition of discreet height channels to classical music recordings significantly increases listener envelopment. What now needs to be determined is exactly what components of height information are the most important to achieving increased envelopment and realism.

Future studies in this area should include formal subjective listening evaluations. Future topics for exploration could include: 1) techniques for creating 3D music recordings that achieve similar sonic results using a reduced number of height channels, 2) improving the spacing and placement of lateral microphones for the main playback layer, 3) the impact of height channel microphone polar patterns on overall listener preference.

3.2 Listener preference for height channel microphone polar patterns in three-dimensional recording

Abstract

A study was conducted to determine if listener preferences exists among three different height channel microphone polar patterns, for three-dimensional music production. Three-dimensional recordings were made of four different musical instruments, using five-channel surround microphone arrays augmented with two Sennheiser MKH 800 Twin microphones as height channels. In a double-blind listening test, subjects were asked to rate different mixes of the same recordings based on preference. The independent variable in these mixes was the polar pattern of the height channel microphones. Analysis of the results found that a clear majority of subjects showed no statistically significant preference for any one polar pattern.

3.2.1 Introduction

Chapter 3.1 discusses experimental microphone techniques for three-dimensional classical music recording [143]. Using a combination of omnidirectional and cardioid microphones, a fourteen-channel microphone array was designed to capture a three-dimensional sound scene of a baroque ensemble performance. Omnidirectional microphones assigned to front and rear height channels were observed to contain too much direct sound from the ensemble. This correlation of direct sound with the main layer microphones made it difficult to achieve an ideal recording balance, as increasing the level of the height channels past a certain point tended to destabilize the image of the ensemble, “smearing” the instruments upward. Based

on this, and the aesthetically superior sound captured by cardioid pattern "side" (±90˚) microphones [143], it was hypothesized that directional microphones would be the best choice for capturing height information in a way that yields both a strong focused ensemble image, and excellent listener immersion.

3.2.1.1 Capturing an Ideal Balance of Sound

Listeners of recorded classical music have become accustomed to an idealized, realistic recreation of a live performance in an acoustic space [46], [45]. Many acoustic music capture techniques have been developed for both stereo and 5.1 surround sound (see: section 2.2), typically optimized to reproduce an even balance of direct and diffuse sound. These one and two-dimensional capture techniques fall short of reproducing the fully immersive experience of listening to a live performance in a real acoustic environment. The addition of height channels allows the recording engineer to enhance the reproduction of musical performances by improving the depth, presence, envelopment, naturalness, and intensity of the sound scene [2], [3], [4]. As seen in comprehensive reviews by Dichreiter [31], and DPA Microphones [144], most stereo and five-channel acoustic music capture techniques specify the polar pattern of each microphone in their respective array designs.

3.2.1.2 3D audio for home listening

Japan Broadcasting Corp. (NHK) plans for Super Hi-Vision with 22.2 Multichannel Sound to begin broadcasting to consumers prior to the 2020 Tokyo Olympic Games [64]. Other three-dimensional audio formats, such as Auro 3D [7] and Dolby Atmos [6] are already available for cinema and consumer entertainment systems. Record labels such as 2L and UNAMAS are producing commercially available 9.1 channel music recordings using Pure Audio Blu-ray as a delivery format [83]. Given the growing availability and importance of three-dimensional audio for film and music production, surprisingly few published works have discussed the development *and* implementation of 3D acoustic music recording techniques in detail [72],

[87], [10], [82].

3.2.2 Test Recording

As a pilot study, a test recording was undertaken to capture height information using multiple microphone polar patterns simultaneously. The recording of a contrabass-recorder took place in a medium-large studio space, using a seven-channel three-dimensional microphone array (Table 2 and Figure 20):

Table 2: Microphones used for test recording

Channel	Microphone	Polar Pattern
FLc	Schoeps MK 21	Wide Cardioid
FRc	Schoeps MK 21	Wide Cardioid
FC	Neumann U87	Cardioid
BL	Schoeps MK4	Cardioid
BR	Schopes MK4	Cardioid
TpSiL (+90°)	Sennheiser MKH800 Twin	Variable
TpSiR (-90°)	Sennheiser MKH800 Twin	Variable

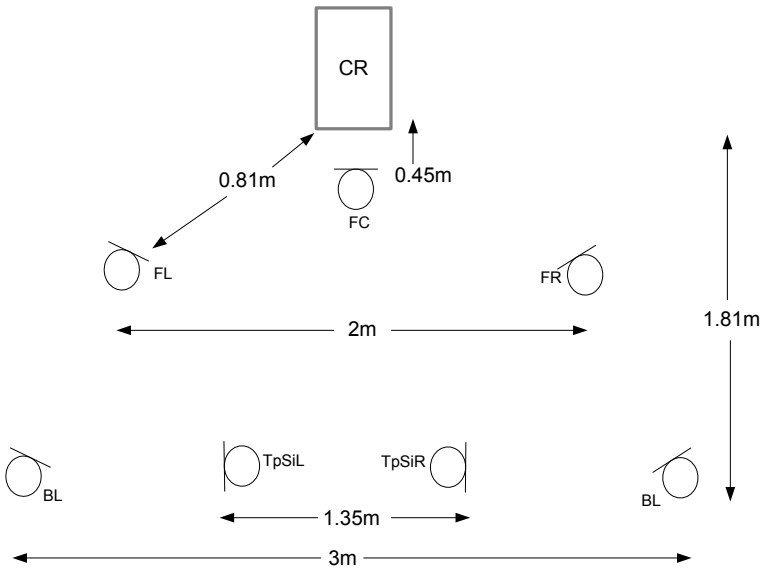


Figure 20: Overhead view of pilot test recording microphone layout

This small-scale array was designed to be simple to set up, and compatible with any current 3D audio system: the height microphone signals can be assigned to any pair of height channels. For this recording, the decision to use the TpSiL and TpSiR height channels (Figure 30) was based on a previous experimental recording [143], as well as previous research showing the importance of lateral reflected sound energy for achieving strong levels of listener envelopment [124]. Sennheiser MKH800 Twin microphones, which feature a back-to-back dual capsule design, were used to record height information. The microphone's dual outputs, one from each transducer, allows the recording engineer to derive any polar pattern by adjusting the balance between the two capsules. The recordings were monitored in McGill University's Studio 22 (see: Section 3.2.3.3).

After the test recordings were completed, the recording engineer, performer, and composer of the recorded repertoire spent time mixing and comparing the available polar patterns of the height channel microphones, focusing on cardioid, omnidirectional and bi-directional. All three listeners were surprised by the apparent differences in captured height information, and how greatly the overall sound of the recording was affected by changing the height channel polar patterns. It was observed that the cardioid height channels contributed to a strong, focused instrument image, while the omnidirectional height channels gave a less stable image, but a richer room sound. None of the listeners enjoyed the sound of the bi-directional height channels, which had an overly thin timbre, and seemed to promote a shrillness in the tone of the contrabass-recorder.

3.2.3 Listening Test

Based on the results of the pilot study, a test was designed to investigate possible preferences among listeners for height channel microphone polar patterns.

3.2.3.1 Test Stimuli Creation

Using the experimental recording technique developed for the pilot study (Section 3.2.2) as a guide, nine more seven-channel three-dimensional music recordings were made. Seven different solo instruments were recorded in three different acoustic spaces. All three acoustic spaces are located within the Schulich School of Music's Elizabeth Wirth Music Pavilion. The large scoring stage (Music Multimedia Room) measures 24.4m x 18.3m x 17m, and has little acoustical treatment, with an RT60 of 2.5s (Figures 25, 26). The Medium-large studio, measuring 11m x 7m x 6.1m, has a combination of absorptive and diffusive acoustical panels in the lower part of the room, while the upper walls remain untreated (Figure 27). The studio's asymmetrical layout results in a sound field that is quite reverberant for a room of its size: RT 60 is 1s. The isolation booth has similar acoustic treatment to the medium-large studio, and measures 5m x 3.2m x 6.1m (Figure 28), with an RT 60 of approximately 600ms. Table 3 shows which instruments were recorded in what spaces:

Table 3: List of stimuli recordings

Instrument	Acoustic Space
1. Harp	Large scoring stage
2. Piano (Yamaha C7)	Large scoring stage
3. Drum Kit	Large scoring stage
4. Cello	Medium-large studio
5. Trumpet	Medium-large studio
6. Drum Kit	Medium-large studio
7. Acoustic Guitar	Medium isolation booth
8. Male Vocal	Medium isolation booth
9. Drum Kit	Medium isolation booth

3.2.3.2 Microphone Choice and Placement

For all stimulus recordings, the spacing between and angle of the height channel microphones remained the same, though their height and distance from the sound source varied depending on the instrument and room. For the main layer microphones (L, C, R, BL, RB), microphone choice and placement varied depending on the instrument, acoustic space, and repertoire being performed. For the harp and cello, traditional spaced arrays were used, with a focus on achieving a strong centre image and diffuse surrounding ambience. For the drums and guitar, a more pop-based approach was taken, resulting in somewhat asymmetrical setups (Figures 21–28). For all recordings, the height channel microphones were Sennheiser MKH800 Twins, oriented perpendicular to the instrument. All microphones were routed to a Sony SIU-100 System Interface Unit, using the internal microphone preamps and analog to digital conversion. Recordings were made to a Pro Tools HD system, at 96kHz/24bit resolution. Table 3 shows microphone choice as per musical instrument.

Table 4: Microphones used for stimulus recording

Instrument	Microphone	Polar Pattern
Drums Overhead L	Neumann U87	Cardioid
Drums Overhead R	Neumann U87	Cardioid
Drums Kick Spot	Audio Technica AT4047	Cardioid
Drums Snare Spot	Shure SM57	Cardioid
Drums BL	Schoeps MK4	Cardioid
Drums RB	Schoeps MK4	Cardioid
Harp FL	Schoeps MK2	Omnidirectional
Harp FC	Schoeps MK4	Cardioid
Harp FR	Schoeps MK2	Omnidirectional
Harp BL	Schoeps MK4	Cardioid
Harp BR	Schoeps MK4	Cardioid
Acoustic Guitar FC	Schoeps MK4	Cardioid
Acoustic Guitar FL	Schoeps MK4	Cardioid
Acoustic Guitar FR	Schoeps MK4	Cardioid
Acoustic Guitar BL	Schoeps MK4	Cardioid
Acoustic Guitar BR	Schoeps MK4	Cardioid
Cello FL	Schoeps MK21	Wide Cardioid
Cello FC	Schoeps MK4	Cardioid
Cello FR	Schoeps MK21	Wide Cardioid
Cello BL	Schoeps MK4	Cardioid
Cello BR	Schoeps MK4	Cardioid

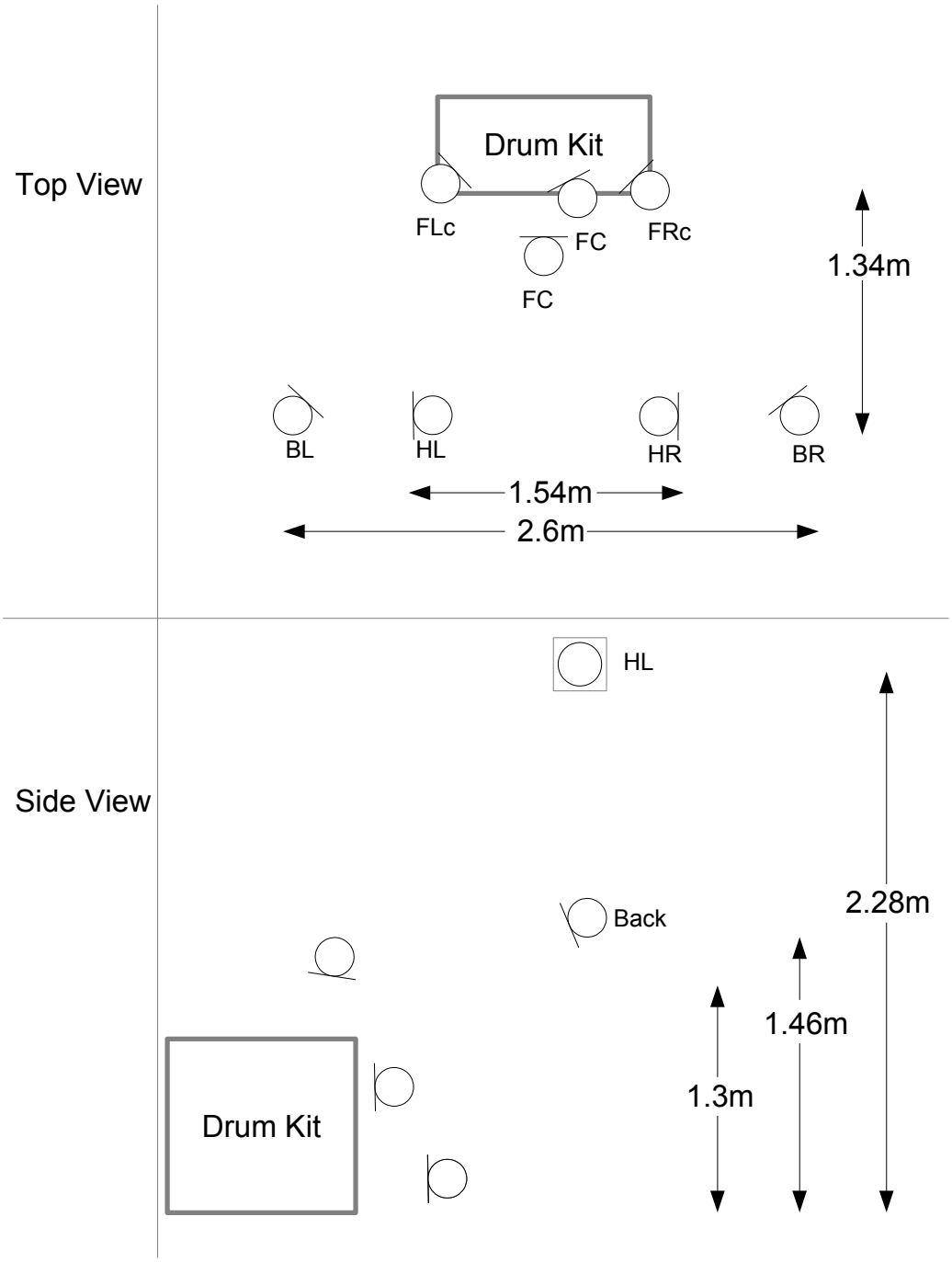


Figure 21: Drums in large scoring stage

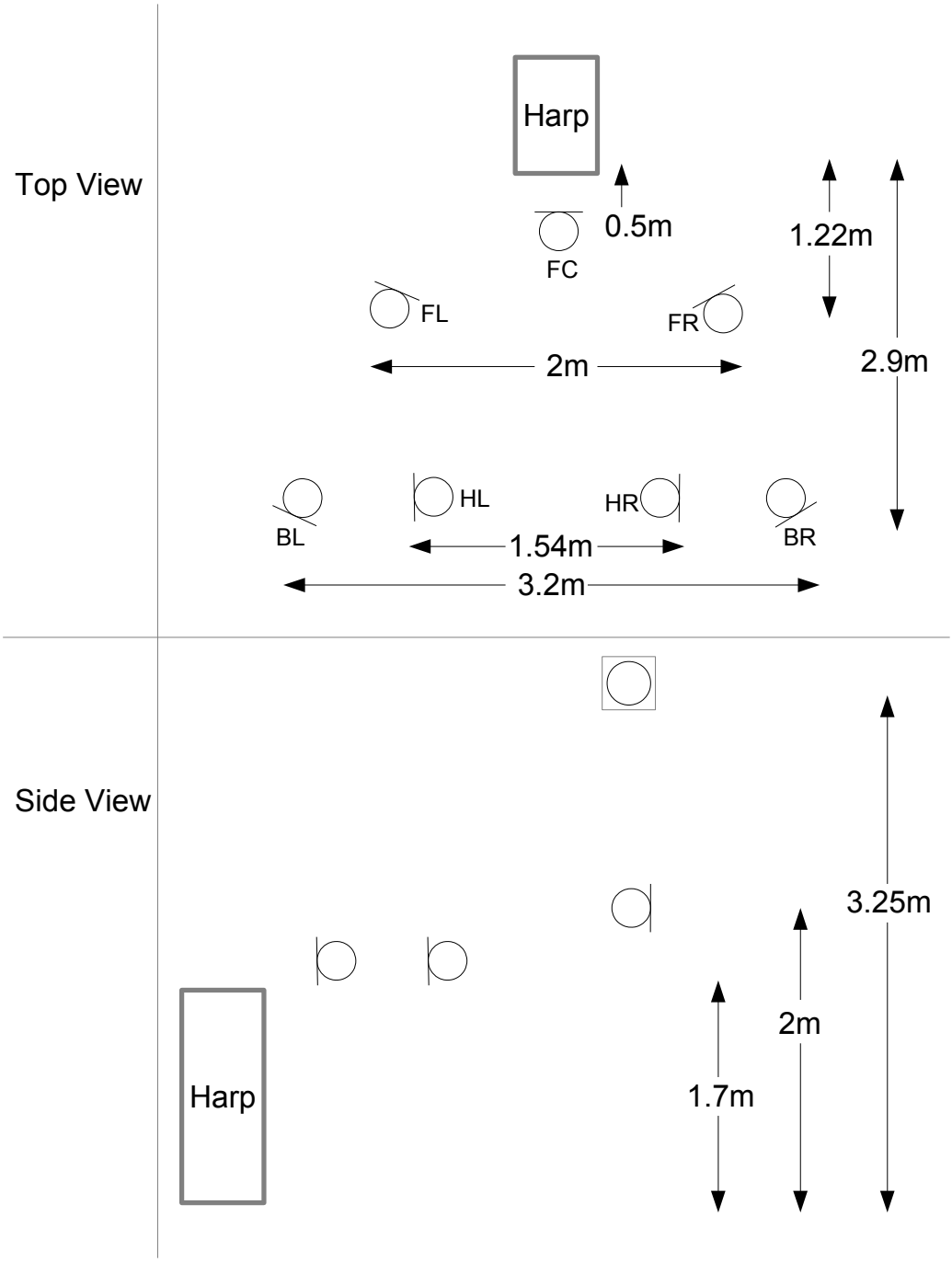


Figure 22: Harp in large scoring stage

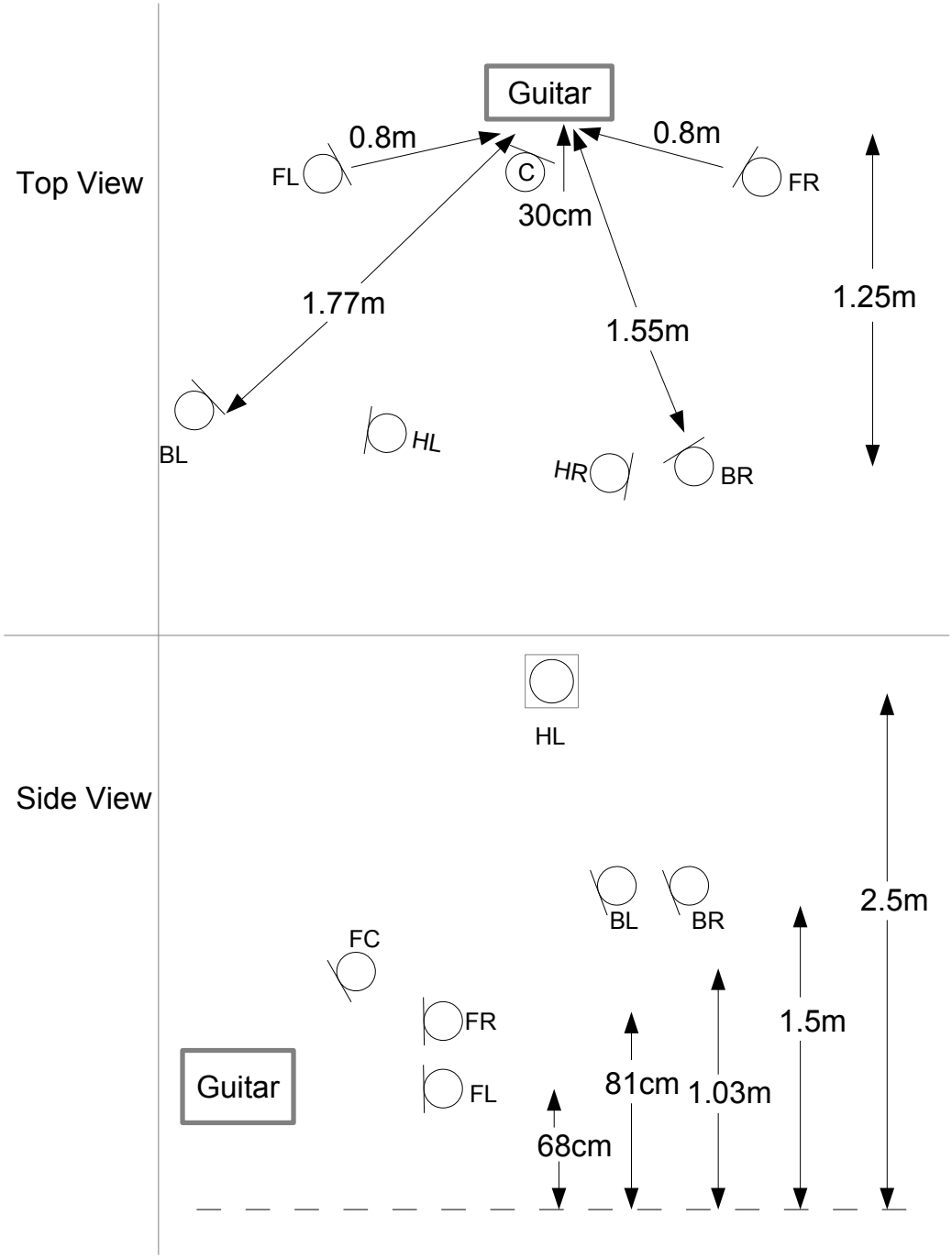


Figure 23: Guitar in isolation booth

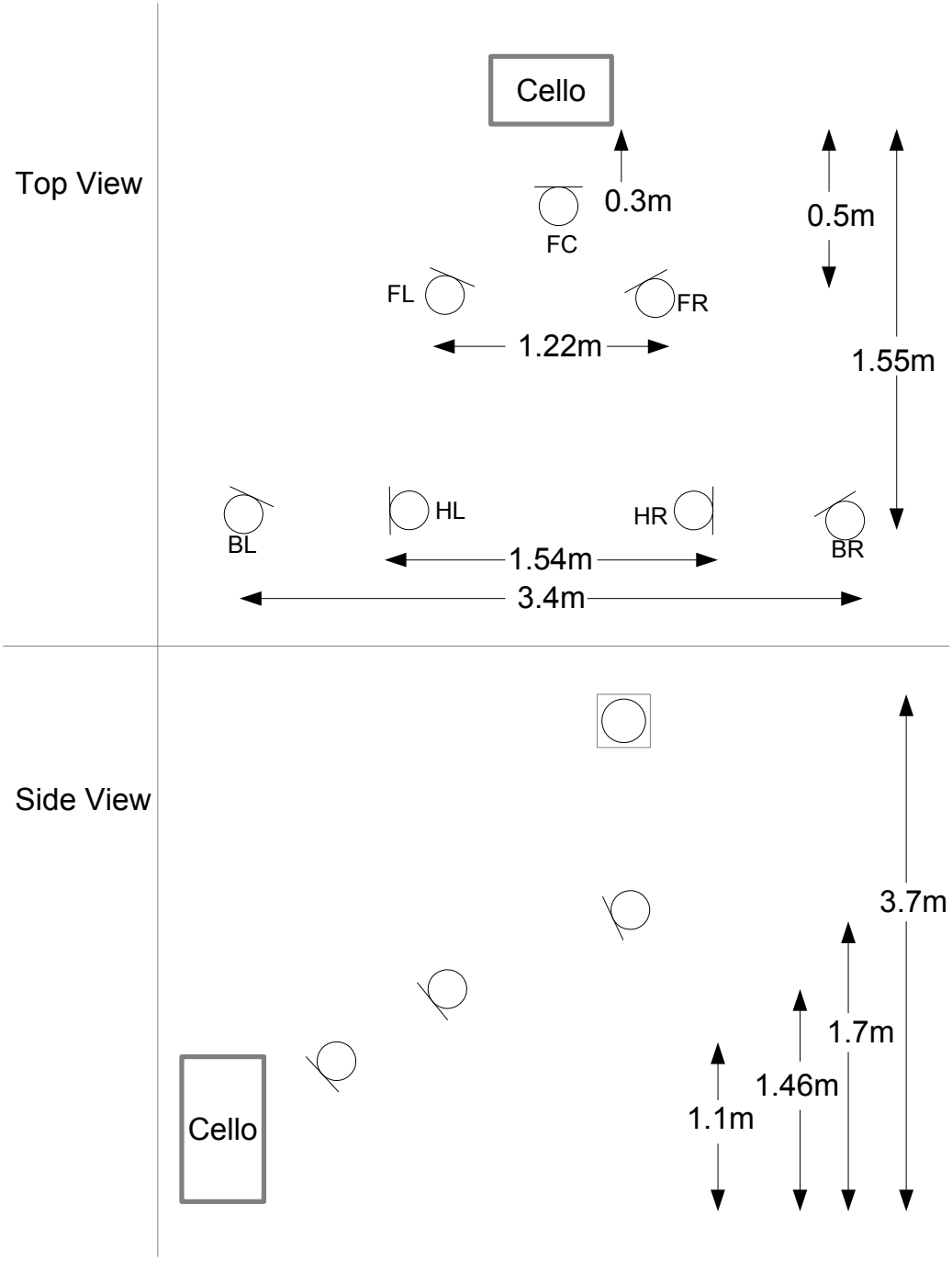


Figure 24: Cello in medium recording studio

Figure 25: Harp in large scoring stage

Figure 26: Drums in large scoring stage

Figure 27: Cello in medium recording studio

Figure 28: Acoustic guitar in isolation booth

3.2.3.3 3D Audio Control Room

All 3D audio playback (recording and mixing of stimuli, test administration) took place in McGill University's Studio 22 (Figure 29), an acoustically treated listening room with 28 channels of discreet audio playback via Musikelectronic Geithain GmbH *M-25* speakers. The 28 speakers are arranged to accommodate both 22.2 multichannel sound [61] and Auro 3D 9.1 [7]. Studio 22 fulfills ITU-R BS.1116 [105] requirements (see also: Section 6.4).

Figure 29: Studio 22, McGill University

3.2.3.4 Stimulus Mixing and Level Matching

Three height channel microphone polar patterns were chosen for comparison: cardioid, omnidirectional, and bi-directional. The four 3D stimulus recordings considered to have the highest sound quality and greatest contrast in acoustic and musical content were chosen for

the listening test: harp (scoring stage), drums (scoring stage), cello (medium studio), and acoustic guitar (isolation booth). Pairs of height channels for each of the three polar patterns under test were created for each of the four musical excerpts. Three audio engineers independently level-matched the different polar pattern height channel mixes for each stimulus. Listening only to the height channels, each engineer compared the different polar pattern pairs (TpSiL and TpSiR: Figure 30), and balanced these pairs until they were perceived as being of equal loudness. The mix volume levels for each author were recorded, and the averages of those levels were used to determine the final matched levels.

Each seven-channel stimulus recording was then balanced by a team of two professional recording engineers, both of whom had previous experience in 2D and 3D audio production. All mixes maintain a "concert" perspective: direct sound in front, ambience to the sides, behind and above. Martin et al. [142] examined listener perception of "immersion" as influenced by the ratio of height channel level to main layer channel level, for a 3D recording of an acoustic guitar. The authors concluded that in order for listeners to perceive a significant level of immersion, height channel microphone signals should not be mixed less than 10dB lower than main layer microphone signals [142]. Using this recommendation, as well their own professional experience in mixing multichannel music, the mixing engineers attempted to balance the microphone signals for each sound source in such a way as to contain enough height channel information to be pleasant, realistic, and enveloping, but not overly obtrusive or exaggerated. The goal with this approach was to create stimuli mixes that reflected the kind of balances typical of commercial recordings. In this way, the results would hopefully be more ecologically valid.

It was observed that for most musical excerpts, using the ±60˚ FL and FR channels contributed to a greater sense of width and spaciousness in the frontal sound image. Drum overhead microphones, however, were panned to the FLc and FRc speakers (Figure 30),

which gave a more realistic impression of instrument size and width. Seven-channel mixes were created for each of the three polar patterns under investigation, for each of the four musical excerpts, for a total of 12 stimuli. All stimuli were 30 seconds in duration. Given the obvious acoustic differences between the three recording spaces, as well as the stylistic differences between the four musical excerpts chosen as stimuli, a certain amount of change in the perceived direct-to-reverberant sound ratio was unavoidable between mixes. However, great care was given to ensure the relative level of height channel information remained consistent between the four musical excerpts/sound sources.

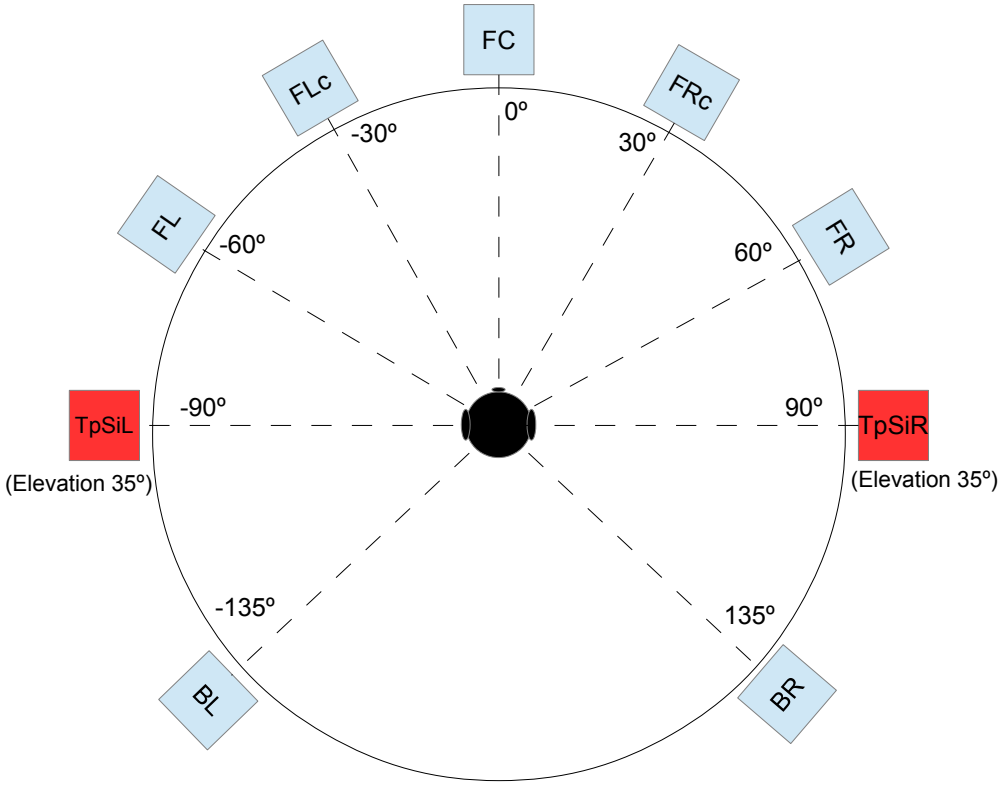


Figure 30: Speaker configuration for listening test

3.2.3.5 Test Design and Implementation

A double-blind listening test was implemented using Cycling 74's Max/MSP. Subjects were seated in Studio 22's central listening position and presented with an interactive GUI (Figure 31). For each trial, one of the four musical excerpts played on a repeating loop. Subjects were

asked to listen to mixes labelled “A”, “B” and “C” on the GUI. Subjects could switch between mixes at any point during playback, as many times as needed. All stimuli were time-aligned for seamless transition between listening selections. For each trial, subjects were instructed to “rate the three mixes in order of general preference”, using 100 point sliders. A comments box in the GUI allowed subjects the option to briefly explain why they made their decision.

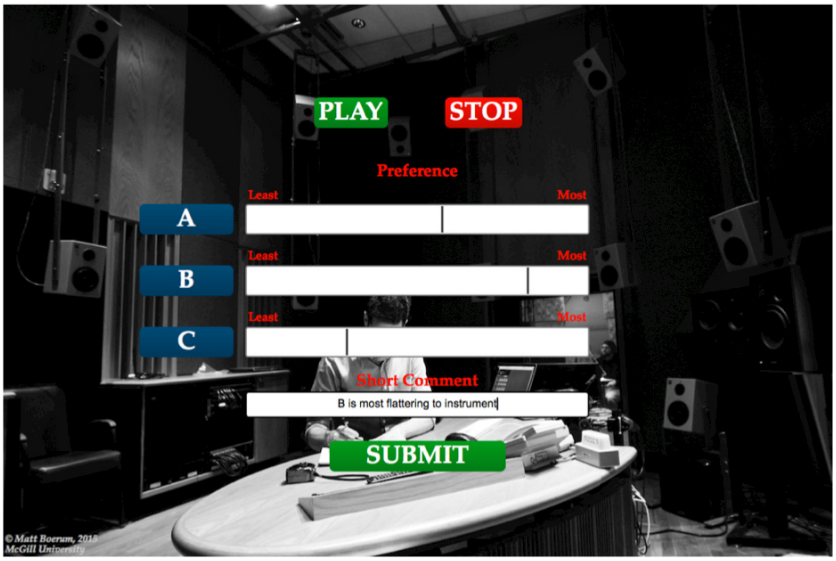


Figure 31: Listening Test GUI

Within the current literature, there are numerous examples of listening tests comparing different multichannel microphone techniques. Some tests have asked listeners to rate techniques based on specific attributes, such as spaciousness [145], [36], envelopment, depth, or localization [36]. Other tests have focused on general listener preference between recording techniques [146], [147]. This area of research is further discussed in Section 5.1.2. For this study, subjects were not given any specific subjective qualities or attributes to consider when making their preference choices.

Each subject completed four trials for each musical excerpt, for a total of sixteen trials. The presentation order of the different excerpts was randomized. The order in which

the different stimuli were assigned as letters “A” “B” and “C” for each trial was also randomized. A total of 29 subjects performed the listening test. The subjects ranged greatly in terms of age and audio production experience (Table 5). Many of the subjects had limited or no prior experience listening to three-dimensional audio reproduction of acoustic music. Upon completing the test, subjects were asked to fill out a brief demographic survey, which included space for general comments about the test experience.

Table 5: Subject demographics

Subject Age (in years)	Number of Subjects
18-25	14
26-32	10
33-39	3
40-50	1
51+	2
Subject Identification	Number of Subjects
Professional Engineer/Producer	7
Recent McGill Sound Recording Masters Graduate	4
Current McGill Sound Recording Masters Student	4
McGill Sound Recording/Music Undergrad	7
Other McGill Students	8

3.2.4 Results

Prior to the experiment, plans were made to analyse the preference scores for the three microphone polar patterns in two ways: with all subjects pooled together, and with each subject considered separately. The first analysis would reveal general trends valid for the entire population of subjects, while the second would reveal individual differences in preference. For most of the statistical tests performed (sections 3.2.4.1 to 3.2.4.4), the data for the four different musical excerpts were pooled together. For subsequent analyses investigating a main effect of instrument (section 3.2.4.5) and instrument-specific effects

(section 3.2.4.6), the collected preference ratings were divided into four data sets, one for each instrument, with each being analyzed separately.

3.2.4.1 Normality Tests: Pooled Results

As a first step in the analysis, the pooled preference scores were tested for normality using a Shapiro-Wilk test executed in R. All three were significantly non-normal (Cardioid: $W = 0.979$, $p < .001$; Figure-8: $W = 0.984$, $p < .001$; Omnidirectional: $W = 0.983$, $p < .001$). Histograms of the data showed two visual features that deviated from a bell-curve shape (Figure 32).

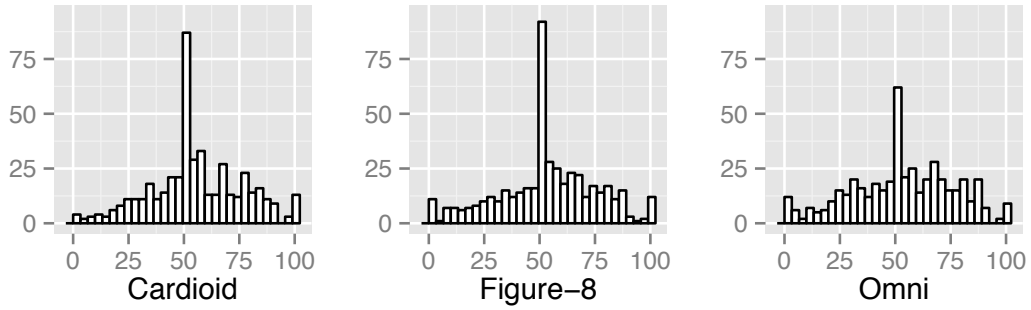


Figure 32: Histogram of preference scores by polar patterns

There were a large number of responses at the centre of the scale, with a value of exactly 50. This can be attributed to the fact that the sliders were reset to this value at the start of each trial. It seems that, in many cases, subjects left the sliders at this initial value rather than moving them. There were also a large number of responses at the ends of the scale. This excess of extreme scores resulted from a small number of subjects who gave highly polarized ratings.

3.2.4.2 Normality Tests: Individual Results

The data from individual subjects were also tested for normality. While some subjects produced normally distributed scores, many gave responses exhibiting the features described above. These non-normalities precluded the use of parametric tests, such as analysis of

variance (ANOVA), to check for differences between groups. Instead, the Kruskal-Wallace test was used. Kruskal-Wallace is a non-parametric alternative to a one-way ANOVA that operates on ranked data.

3.2.4.3 Pooled Preferences

When the preference scores of all subjects were pooled together, no significant differences between the polar patterns were found, $H(2) = 1.58$, $p = .45$. (Figure 33)

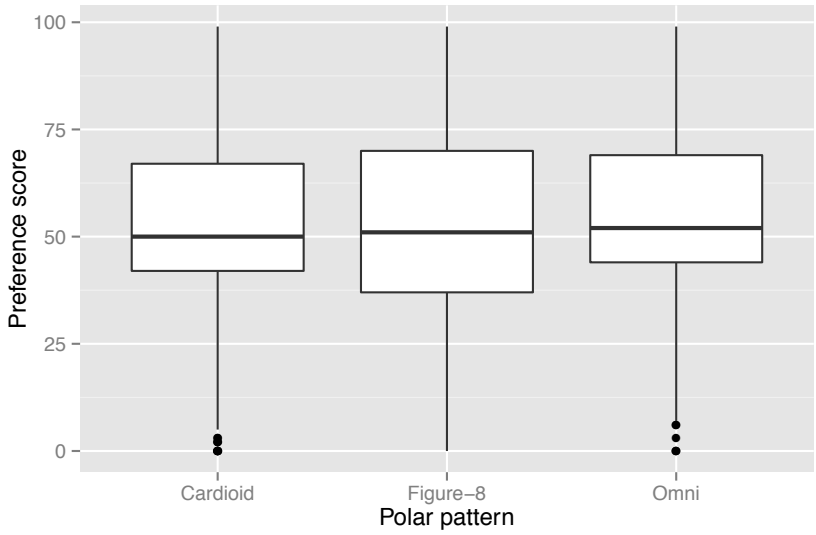


Figure 33: Polar pattern preference ratings, pooled across all subjects

3.2.4.4 Individual Preferences

When the preferences of individual subjects were tested, differences were revealed in only two of the 29 cases: subject 2, $H(2) = 15.6$, $p = .012$; and subject 28, $H(2) = 3.97$, $p = .037$. (p-values were corrected with Holm-Bonferroni) For subject 2, comparisons of mean ranks showed that Figure-8 (bi-directional) had a significantly lower preference rank than Cardioid (difference = 19.3). For subject 28, Figure-8 had a lower rank than Omnidirectional (difference = 17.7). In both cases, the critical difference ($\alpha = 0.05$ corrected for the number of tests) was 11.8. Raw preference scores for subjects 2 and 28 are shown in Figure 34.

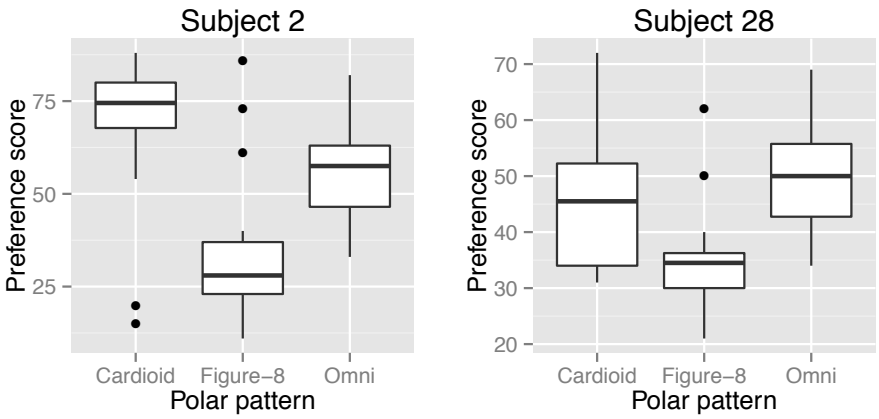


Figure 34: Preference scores for subjects 2 and 28

For the majority of subjects (27 out of 29), the rankings given to the three microphone polar patterns were not significantly different (Figure 35).

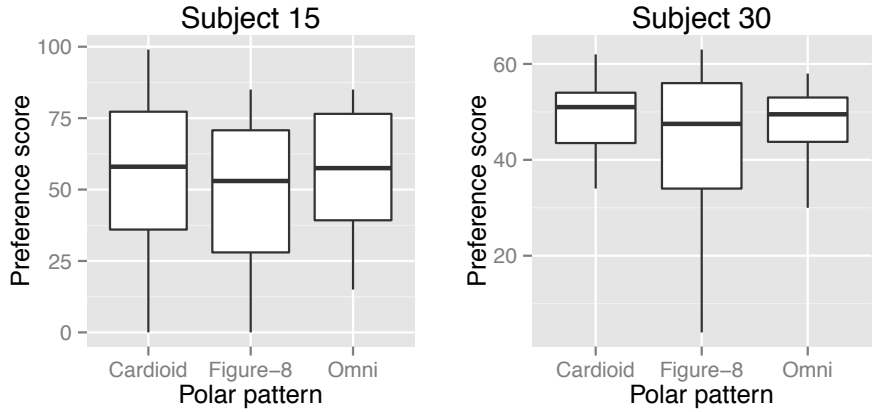


Figure 35: Preference scores for subjects 15 and 30. These subjects were typical in exhibiting no significant preference for any polar pattern.

3.2.4.5 Main Effect of Instrument

A Kruskal-Wallis test was conducted to investigate for a main effect of instrument on preference. A main effect of instrument on preference rank was found, $H(3) = 9.15$, $p = 0.027$. A multiple comparisons test following Siegel and Castellan [148] showed a significant difference only between the Harp and Cello sources, with Harp having a slightly higher mean

rank than Cello (difference = 90.2; critical difference = 80.4; α = 0.05, adjusted for the number of tests).

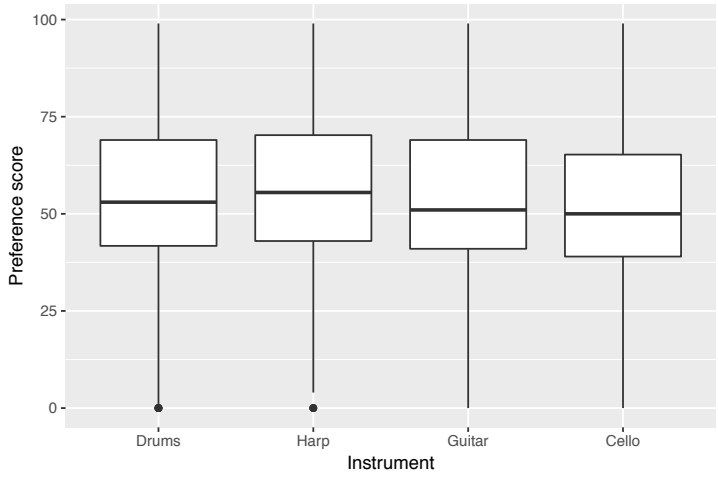


Figure 36: Main effect of instrument on pooled preference rankings

3.2.4.6 Instrument or Room Specific Effects

Since the testing stimuli were derived from four different sound sources, recorded in three different acoustic spaces, it would valuable to know whether certain polar patterns were preferred for specific instrument – room combinations. Four additional Kruskal-Wallis tests were performed on polar pattern and preference, with each instrument considered separately, and all subjects pooled together. In all four cases, no significant effects were found.

3.2.4.7 In-Trial Optional Listener Comments

Although not required to do so, many subjects took the time to fill in comments for at least some of the trials, giving a brief explanation, often two or three key words, as to why they preferred a specific polar pattern. Several subjects filled in very detailed comments for each trial, sometimes also including why a specific polar pattern was least preferred. These comments were searched by the primary author for terms or synonyms of terms common to subjective spatial audio evaluation. Terms with similar or identical meanings, such as "envelopment" and "immersion," were pooled together, referencing lexicon studies by Le Baggousse et al. [120] and Zacharov and Pederson [108]. The pooled terms were then sorted

by polar pattern and counted. For all three polar patterns, “most enveloping/presence”, “wider sound source”, and “most clarity” were the three most common reasons given for preference. Similarly, all three polar patterns had the same most common reasons for being least preferred in a given trial: “least enveloping”, “thin sound”, and “confusing imagery”.

3.2.5 Discussion

For a clear majority of test subjects, no significant preference for any one height channel microphone polar pattern was shown. This result was true for several different musical instruments recorded in acoustic spaces ranging from a large reverberant scoring stage, to a resonant isolation booth, typical of pop drums, piano, or vocal recording.

Of the 29 subjects, 22 left general comments about the test. Ten of those subjects commented on the subtlety or difficulty of the test. Below are several sample comments:

Subject 001: “I found it very hard to hear any differences with the cello, harp and guitar.”

Subject 004: “In general, the differences were, for me, very subtle. In some cases, I did not even perceive a difference.”

Subject 027: “I found the cello recordings virtually indistinguishable.”

As mentioned in section 3.2.3.4, each stimulus was mixed with the goal of creating a balance that conveyed a significant level of immersion and envelopment, but did not overtly emphasize the height channels at the cost of overall mix cohesion. The majority of the subjects who participated in this experiment had limited previous experience listening to three-dimensional audio. This lack of familiarity with the playback medium, when combined with stimuli whose microphone signal balances prioritize mix cohesion over obviousness of height information, likely lead to differences in reproduction conditions that were either too subtle, or too foreign to most listeners to make strong preference judgements. This view

offers some explanation as to why there were so many subject responses with a value of 50. It seems probable that had the differences in sonic qualities between the three microphone polar patterns been stronger and more obvious, subjects would have felt more compelled to move the preference sliders a correspondingly larger degree. It is difficult, then, in many cases to ascertain whether subjects truly did not perceive any difference between stimuli, or whether the differences they did perceive were reported in such a way that was too subtle to be detected by the statistical tests. Given the fact that many of the subjects were able to articulate why they had chosen a particular polar pattern as most preferred, the latter seems more likely.

When looking at the raw data for each subject, there were numerous instances where subjects were inconsistent with their preferences even with averaged responses over multiple trials. This was true for all four stimuli. This is also supported by post-test comments. For example:

Subject 010: "The subtle differences in the 3 mixes for each track had me questioning myself, especially in the middle of the test."

Subject 003: "I don't think I was consistent."

Here again, we may be seeing the influence of the lack of familiarity with 3D audio, as well as lack of significant professional audio production experience for a majority of test subjects. An inconsistency in preference does not necessarily indicate of lack of preference, but rather a lack of experience making critical judgements about multichannel audio stimuli. For example, one subject's post-test comment read: "Very good mixes in general, might be too good to say which one is bad." It is possible that if the same test were performed with a group of better trained listeners, more definitive results could be obtained.

Results were somewhat more concrete in terms of listener appraisal of the stimuli. It is interesting to note that regardless of which polar pattern was selected as most preferred, the same three aspects of the sound scene appeared to dominate the decision-making process: "envelopment/presence/immersion", "sound source width", and "clarity". This is also supported by a strong tendency for subjects to define their least preferred stimuli as "confusing", i.e. unclear, or "least enveloping". We can also surmise that a fullness of sound was generally preferred by listeners in this study. These general trends in listener comments, when combined with the observations of listeners reported in Section 3.1, indicate that an optimal microphone technique for three-dimensional acoustic music recording should be designed to prioritize strong levels of listener envelopment, a dimensional broadening of the sound source, and clarity of the sound scene, while avoiding any confusing or unnatural spatial imagery, including upward vertical smearing of sonic images. This is in agreement with previous recommendations from Theile and Wittek [75], Hamasaki and Van Baelen [10], and Lee [78].

The aesthetic methodology used to mix the stimuli for this study was based on an assumption that more homogeneously balanced height channels would give the test and subsequent results a greater ecological validity, as this research is concerned primarily with the creation of microphone techniques that are practical for implementation in real-world recording conditions. It is possible, however, that a more exaggerated approach to balancing the height channels would have made differences between polar patterns more obvious, which could have led to more significance in the listener data. Another factor to consider is the scaling bias that was likely created by having the preference sliders default to 50 for each trial: this method will not be repeated going forward. There is also the question of how relevant preference data even is for this type of investigation: preference is not necessarily an indication of performance, as it is based primarily on personal taste. It may be more valuable

to compare the effects of height channel polar pattern on perceived difference in auditory attributes. Using a pool of more highly trained listeners, such a study may yield results that are more universally applicable outside the experimental conditions of this study. Another approach for an alternative study could be to instruct the subject to set their own balance for the height channels during a pre-test training session, so as to optimize the listening experience during the primary listening test for immersion and ability to discern between stimuli. This more personalized method might remove some of the inherit bias in the test created at the stimulus-mixing stage.

The process of creating the three-dimensional test stimuli was valuable and educational, yielding numerous ideas for microphone arrays that could be explored in future recordings. The potentially positive result of this research is that recording engineers currently exploring three-dimensional music recording should not necessarily feel bound by the example of past microphone techniques that specify that certain polar patterns be used for certain applications. Rather, microphone choice and placement should be based on capturing the specific type of sonic information most relevant to a given reproduction channel in order to create a unified, coherent sound scene.

4 A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound

Abstract

Based on results from previous research, as well as a new series of experimental recordings, a technique for three-dimensional orchestral music recording is introduced. This technique has been optimized for 22.2 Multichannel Sound, a playback format ideal for orchestral music reproduction. A novel component of the recording technique is the use of dedicated microphones for the bottom channels, which vertically extend and anchor the sonic image of the orchestra. Within the context of highly dynamic orchestral music, an ABX listening test confirmed that subjects could successfully differentiate between playback conditions with and without bottom channels.

4.1 Introduction

4.1.1 22.2 Multichannel Sound

In recent years, much work has been done to introduce and standardize various three-dimensional audio formats for cinema, home theatre, and broadcast [77], [149], [73] , [6], [5], [11]. Japan Broadcasting Corp. (NHK) has developed and introduced Super Hi-Vision, "an ultra-high definition video system with 4000 scanning lines and a viewing angle of 100°. [11]" Super Hi-Vision includes a complementary immersive audio format: 22.2 Multichannel Sound (22.2) [8], standardized by SMPTE [61] and the ITU [5]. Utilizing ten playback channels at ear level, nine above the listener (top layer), and three at floor level (bottom layer) (Figure 37), 22.2 has been shown to significantly increase the impression of "presence" over a wide listening area, as compared with 5.1 surround sound [9]. NHK has produced numerous special programs featuring audio recorded and mixed for 22.2, and plans to be broadcasting Super Hi-Vision to consumers in time for the 2020 Tokyo Olympics [64].

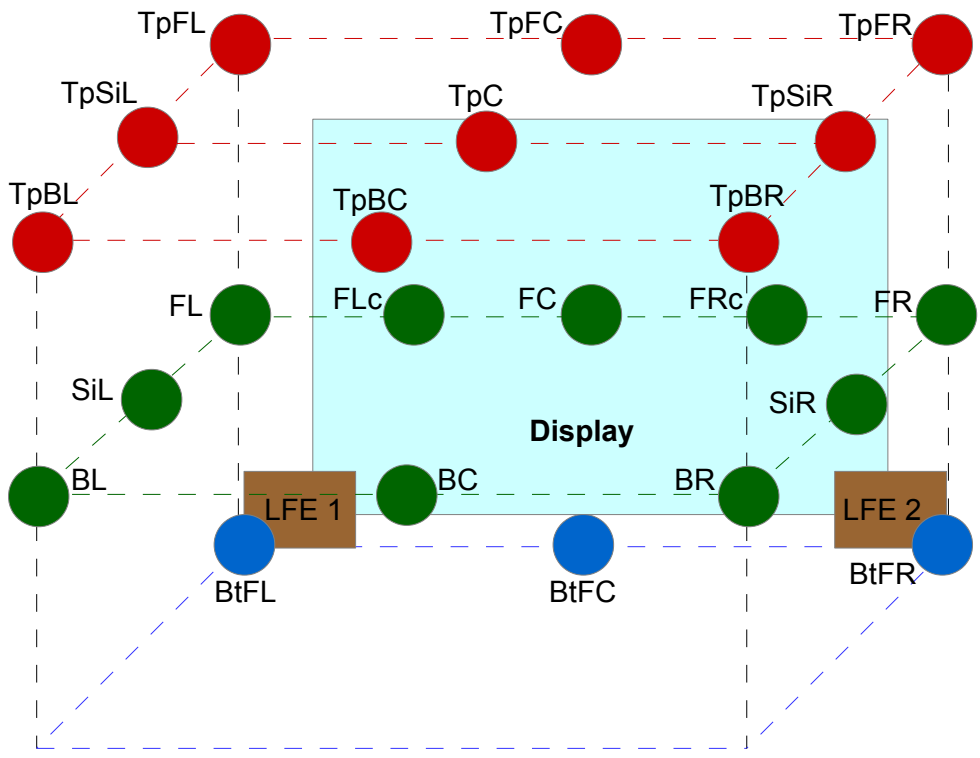


Figure 37: 22.2 Multichannel Sound layout. 9 Top layer channels, 10 Middle layer channels, 3 Bottom layer channels, 2 LFE.

4.1.2 3D Audio and Classical Music Recording

Listeners of recorded classical music have become accustomed to an idealized, realistic recreation of a live performance in an acoustic space [46], [45]. In multichannel audio, this aesthetic typically involves a "concert" perspective, i.e., instruments are reproduced in front of the listener, while ambience surrounds from the sides, behind, and above. Hinata et al. state, "From years of experience in mixing 5.1 surround sound and 22.2 multichannel sound, it is known that close sounds heard from the sides and back create a psychological feeling of pressure, which results in a small spatial impression. [62]" Many acoustic music recording techniques have been developed for both stereo and 5.1 surround sound, typically designed to capture a sound scene with an ideal balance of direct and diffuse sound [31], [1]. These techniques, however, fail to capture the fully immersive experience of listening to a live performance in a real acoustic environment. Three-dimensional audio brings the listener closer to a "natural" listening experience, also improving the depth, presence, envelopment, and intensity of music recordings [2], [3], [4]. Several authors have introduced three-dimensional music recording techniques or concepts primarily aimed at classical ensemble capture [72], [75], [87], [71]. These techniques tend to be designed and optimized for smaller-scale three-dimensional audio formats. Specific to 22.2, most publications have discussed sound capture methods for special events [77], [63] and live sports broadcast [62], but not acoustic music.

4.1.3 Spatial Impression in Multichannel Music Reproduction

In concert hall acoustics, spatial impression is typically divided into two broad categories: Apparent Source Width (ASW), and Listener Envelopment (LEV). For multichannel music production it is envelopment, and in the case of classical music, environmental envelopment, that is the more important of the two spatial attributes. Berg and Rumsey have undertaken extensive research into perceived spatial quality of reproduced sound, finding that "an

enveloping sound gave rise to the most positive descriptors and that the perception of different aspects of the room was most important for the feeling of presence. [112]" "Presence" is defined as "The experience of being in the same acoustical environment as the sound source, e.g. to be in the same room. [112]" When examining correlation between various subjective spatial attributes, Berg and Rumsey found that "preference" was most strongly correlated with "envelopment", while "naturalness" was most strongly correlated with "presence" [112]. In multichannel audio, creating a strong sense of envelopment is key to achieving strong levels of listener enjoyment and immersion.

Hanyu and Kimura have shown that LEV "increases if there is adequate spatial balance in the direction of arriving reflections. [124]" In David Griesinger's model of spatial impression, "background spatial impression" (BSI) is closely tied to envelopment. Griesinger claims that in order to achieve high levels of spaciousness, large fluctuations in the interaural intensity difference (IID) and interaural time difference (ITD) at the two ears during background sound are required [43]. Griesinger suggests that maximum spaciousness will occur when the reverberant component of a recording is fully decorrelated, and recommends that component should be "reproduced by an array of decorrelated loudspeakers around the listener. [43]" Hiyama et al. showed that for loudspeakers placed at even intervals around the listener, "at least six loudspeakers are needed to reproduce the spatial impression of (a) diffuse sound field. [150]"

4.1.4 22.2 Multichannel Sound for Orchestral Music Recording

Most current three-dimensional audio formats retain the traditional 60° frontal sound reproduction angle associated with stereo and 5.1 surround sound [77], [5]. 22.2 Multichannel Sound, however, has a frontal sound reproduction angle of 120°. This wider reproduction angle is ideal for reproducing the sonic image of an ensemble as large as a symphony orchestra. Hamasaki et al. have shown that the use of five frontal speakers, as opposed to

three, is essential for increasing the impression of presence [9]. As seen in Figure 37, the even spatial distribution of loudspeakers at both ear level and above in 22.2 is ideal for the reproduction of early and late lateral reflections, as well as a fully decorrelated reverberant sound field. This ensures maximum listener envelopment across the audible frequency spectrum, as well as contributing to a dimensional broadening of the orchestral image. In two separate studies by Hamasaki and his co-authors [2] [10], when excerpts of orchestral music were used as stimuli, 22.2 was shown to be rated superior to 5.1 for a number of perceptual attributes related to spatial impression, as well as localization accuracy. Similar results were seen in a study by Shim et al. [55] for a wide range of multichannel audio stimuli. Sporer et al. [151] investigated localization of sound objects for several different multichannel audio reproduction formats, and found that 22.2 provided the best localization accuracy, as compared with 10 and 5-channel reproduction formats.

4.1.5 Bottom Channels in 22.2 Multichannel Sound

Within the current literature, only Hamasaki and his co-authors have specifically addressed recording orchestral music for 22.2 (see Section 5.1.1). However, the use of and potential benefits of the bottom channels are not discussed [9], [10]. Typically located below the FL, FC, and FR loudspeakers, at floor level (Figure 37), the bottom channels were originally intended to reproduce special effects germane to on-screen action. An ideal presentation of orchestral sound would make use of these channels to extend the ensemble image to the floor, recreating the conductor's perspective.

From numerous experimental recordings of classical and pop/rock music at McGill University, it has been observed that the bottom channels add a great deal of perceptual "weight" to the instruments or ensembles being reproduced, providing a lower vertical extension that anchors the sonic image. This "anchoring" effect is highly useful, as any correlated or semi-correlated sonic information present in the height channels has the

tendency to cause instrument images to shift upward, which may not be desirable [143], [152]. Martin and King [153] have shown the importance of vertically extending sound images using the bottom channels in re-mixing one-dimensional content for three-dimensional playback environments. Listeners have been shown to prefer higher levels of vertical immersive content in a three-dimensional playback environment [142]; increasing the downward vertical extent of the direct sound image by use of the lower channels allows the recording engineer to maximize the level of immersive ambience.

Using dedicated microphones for the bottom channels adds the advantage of capturing early floor reflections, as well as additional low frequency content due to the complex radiation patterns of orchestral instruments [26]. Loudspeakers at or near floor-level are capable of more efficient low-frequency reproduction to the listener, as they do not suffer from low frequency spectral notches caused by interference between direct sound and floor reflected sound that would be present in sound reproduced from speakers at ear level or above [21]. Roffler and Butler [20] found that tonal stimuli have intrinsic spatial characteristics: different tone bursts reproduced by a single loudspeaker will be located by listeners as being higher or lower in space, depending on their frequency. Cabrera and Tilley observed that, “Having low frequencies originate from lower sources is in harmony with the pervasive pitch-height metaphor. [21]” The lower channels allow the recording/mixing engineer to concentrate low frequency power below and in front of the listener, which is in keeping with an apparent natural human aesthetic.

4.2 Design of Microphone Technique

Based on the above considerations, previous research presented in Chapter 3 [143], [152], established one and two-dimensional recordings techniques [46], [45], [62], [31], as well as a number of experimental three-dimensional music recordings, a new technique was designed to record orchestral music, optimized for a 22.2 reproduction environment. The technique is

designed to incorporate new considerations for creating convincing three-dimensional sound images, but retains compatibility with stereo recording techniques.

4.2.1 Orchestral Sound Capture

The primary component of the orchestral sound image is reproduced by the five front speakers (FL, FLc, FC, FRc, FR) and the three bottom speakers (BtFC, BtFL, BtFR). The main frontal sound capture is based on the classic "Decca Tree" model of three spaced omnidirectional microphones assigned to FLc, FC, FRc, with two additional omnidirectional "outrigger" microphones placed at the lateral three-quarter points of the orchestra assigned to FL and FR (Figures 38, 39). The FLc, FC and FRc microphones are fitted with acoustic pressure equalizers: diffraction attachments that increase microphone directivity at high frequencies, as well as give a natural boost in the 1kHz–5Khz range of the frequency spectrum [81]. The use of omnidirectional microphones is critical for capturing the complete low frequency spectrum of the orchestra, as they do not suffer from proximity effect. These "front" microphones should be placed somewhat closer to the orchestra than is typical for a stereo-only recording, so as to capture less ambient sound. This is aided by the increased directivity introduced by using acoustic pressure equalizers. When also being used as the main system for a stereo mix, the "front" microphones can be combined with several ambience microphones as necessary.

Three directional microphones are placed adjacent to the FL, FC and FR microphones, ideally within a meter of the floor, angled downward at an angle of approximately –45º (Figure 38, 39). These microphones are routed to the bottom channels, and are meant to give the orchestral image a downward vertical extension, as well as capture early floor reflections and low frequency content from instruments with frequency-dependent directivity.

4.2.2 Ambient Sound Capture

Microphones routed to the remaining fourteen playback channels are all directional, mostly cardioid, placed in such a way as to prioritize ambient sound capture and decorrelation between channels (Figure 40). Cardioids are chosen for their high degree of rear sound rejection, as well as being less susceptible to low frequency loss than hypercardioid or bi-directional microphones. Some amount of low frequency roll-off due to proximity effect is desirable, especially in the height channels, as it reinforces the above mentioned "pitch-height metaphor" [21]. To optimize listener envelopment, microphone signals that prioritize early and late reflection capture should be decorrelated across the audible frequency spectrum. This can be achieved through distant spacing between microphones. Hamasaki et al. [46] found that a distance of at least 2m between microphones was necessary to ensure decorrelation above 100Hz, while Griesinger [154] has suggested that spacing microphones at a distance greater than the critical distance of the recording venue will ensure decorrelation at low frequencies.

Ideally, microphone capsule direction should roughly mirror playback speaker direction. For example, microphones routed to the SiL and SiR speakers would face the side walls, primarily capturing lateral reflected sound. The exception would be the TpFL, TpFC and TpFR microphones, which may need to be oriented further away from the orchestra to minimize direct sound capture, depending on venue acoustics. Microphone positions need not be dictated by the layout and relative distances between loudspeakers of a given reproduction format, e.g., techniques such as the "Polyhymnia Pentagon", where five omnidirectional microphones are positioned based on a standard 5.1 loudspeaker layout [44]. Rather, microphone placement should be based on capturing an ideal reverberant sound field from the performance venue, and should be optimized, ideally, while monitoring in a 3D sound reproduction environment.

4.3 Implementation of Design

The proposed recording technique was implemented during recording sessions for the 90-piece, National Youth Orchestra of Canada (NYOC). The recordings took place over three days in McGill University's Music Multimedia Room, a large scoring stage measuring 24.4m x 18.3m x 17m, with an RT60 of approximately 2.5s (Figure 38). Monitoring of the recordings took place in the adjacent Studio 22, an acoustically treated listening room that fulfills ITU-R BS.1116 [105] requirements. 22 Musikelectronic Geithain GmbH *M-25* speakers are arranged for 22.2 Multichannel Sound reproduction, as per [61]. Monitoring in a 22.2 playback environment was essential in order to understand the complex sonic relationships between the different points of ambient sound reproduction. Hearing how all 22 playback channels resolved to form a single audio scene was critical for optimal microphone placement and adjustment.

Height channel microphones were hung from various catwalks above the studio floor in such a way that positional optimization could take place during recording breaks. Microphone choice and placement were as seen in Table 6 and Figures 38, 39, 40 and 41. Most of the recording venue's floor space was occupied by the orchestra, with only 3.9m of free space from the conductor to the back wall of the studio (Figure 38). As such, certain ambience microphones were spaced extremely widely to gain greater distance from the "frontal" microphones, thereby ensuring an appropriate amount of depth in the audio scene. To avoid strong rear wall reflections in the Back Centre channel, a laterally oriented bi-directional microphone was used.

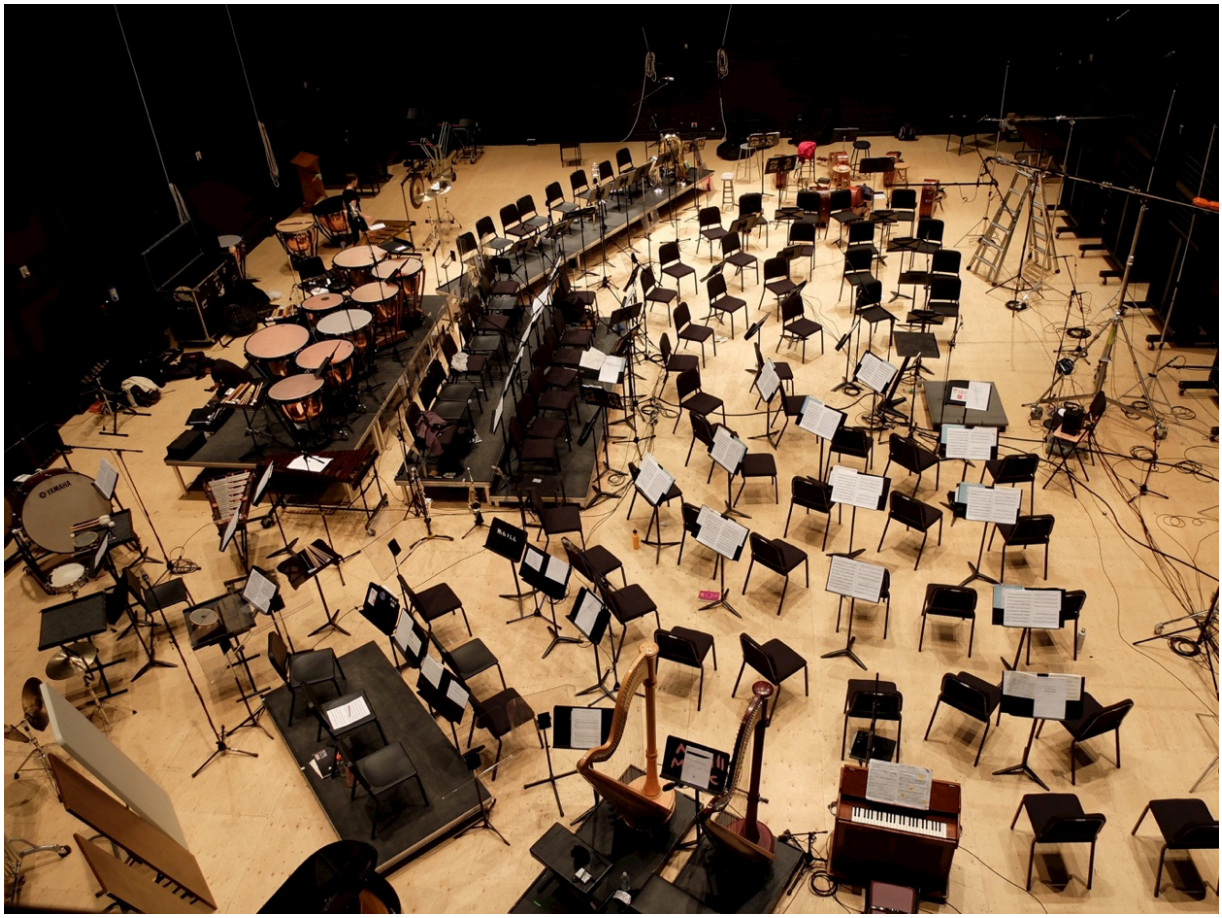


Figure 38: National Youth Orchestra in Music Multimedia Room. Decca tree is positioned above the conductor's podium.

Figure 39: Frontal sound capture microphones, as seen from viola section. Red = height layer, green = main layer, blue = bottom layer.

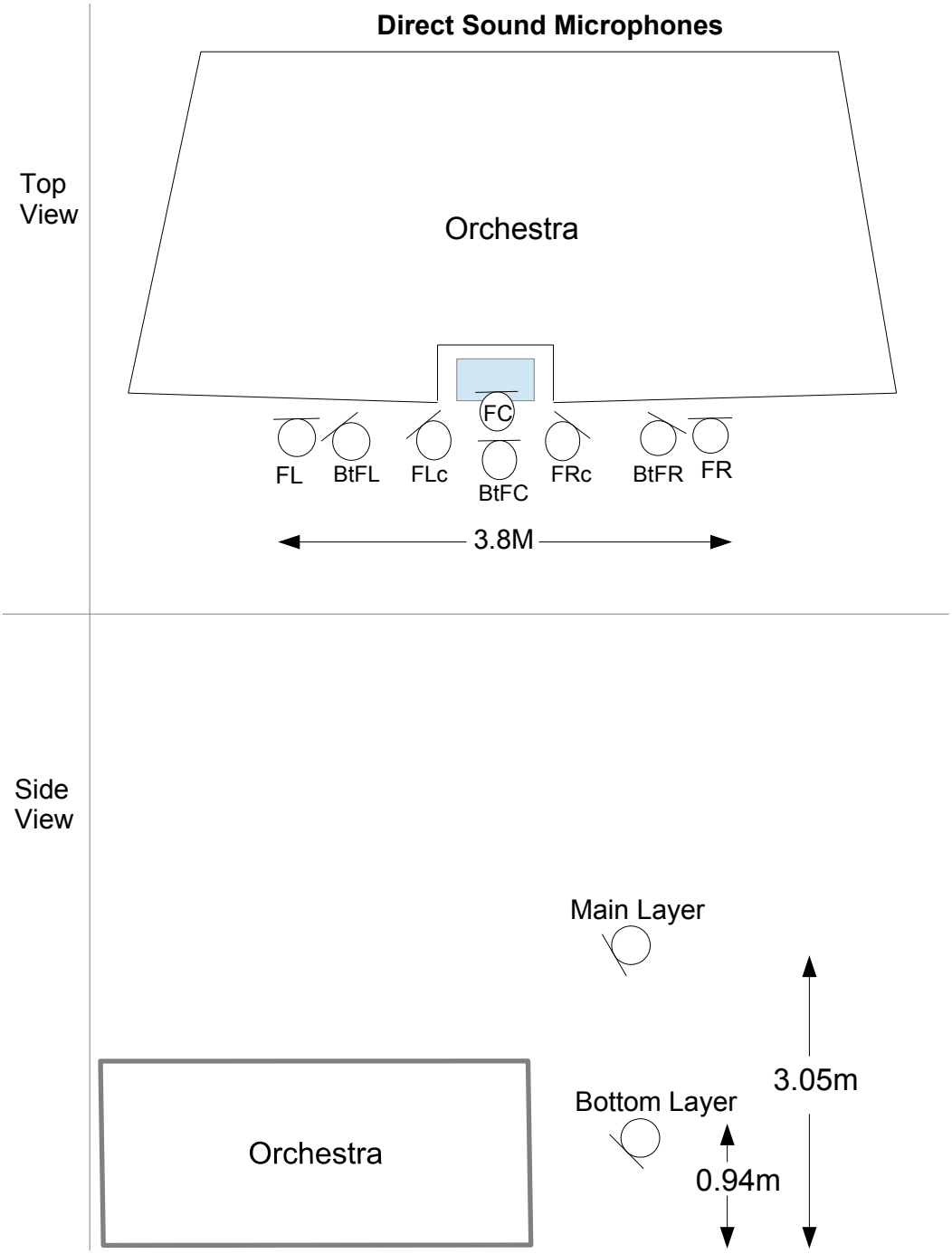


Figure 40: Orchestral sound capture microphones

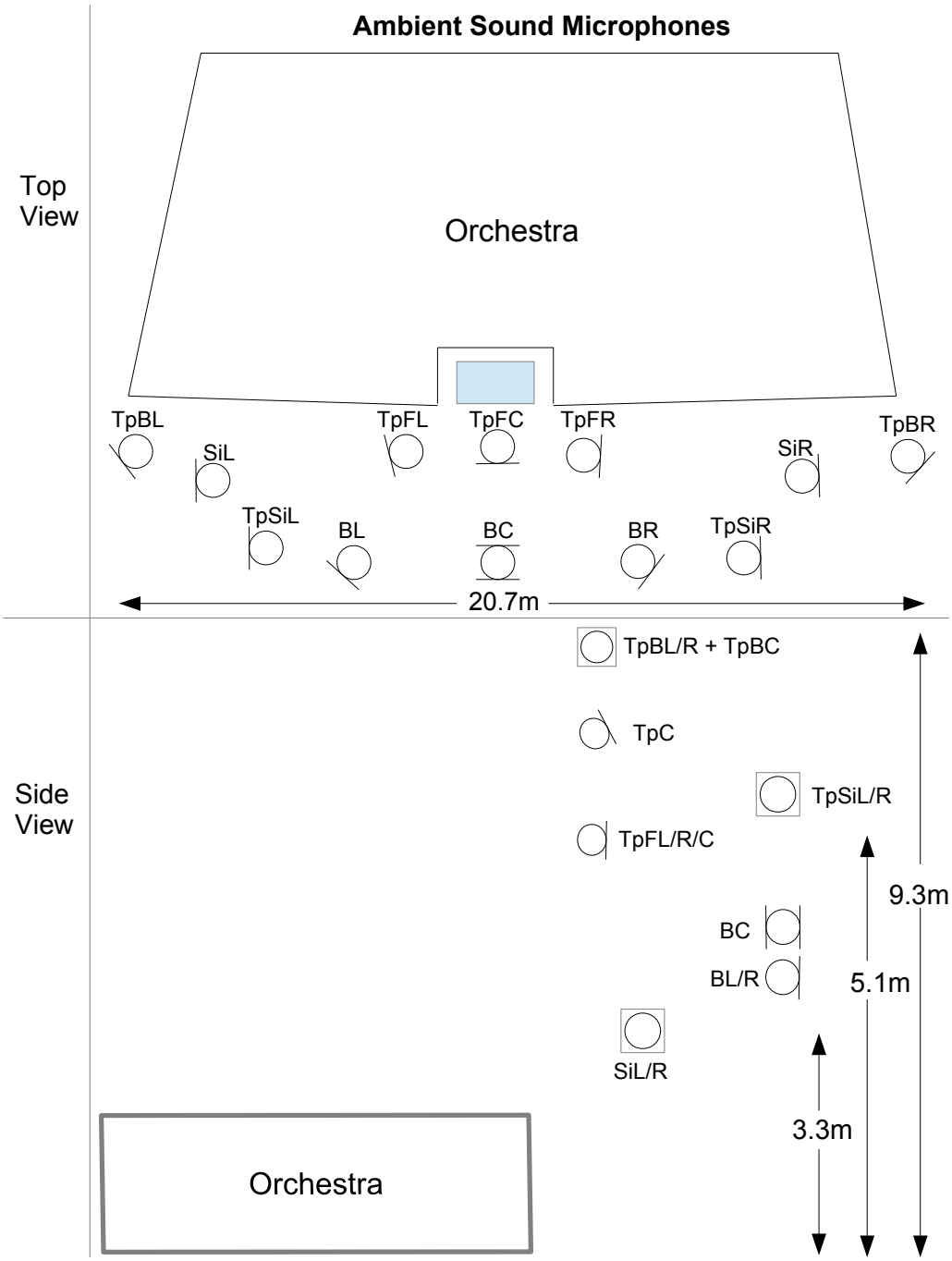


Figure 41: Ambient sound capture microphones

Table 6: Microphones Used for Test Recording

Channel	Microphone	Polar Pattern
FL	Schoeps MK 2S	Omnidirectional
FR	Schoeps MK 2S	Omnidirectional
FC	Schoeps MK 2H w/Acoustic Pressure Equalizer	Omnidirectional
BL	Schoeps MK 21	Wide Cardioid
BR	Schoeps MK 21	Wide Cardioid
FLc	Schoeps MK 2H w/ Acoustic Pressure Equalizer	Omnidirectional
FRc	Schoeps MK 2H w/ Acoustic Pressure Equalizer	Omnidirectional
BC	Neumann KM 120	Bi-directional
SiL	DPA 4011	Cardioid
SiR	DPA 4011	Cardioid
TpFL	Schoeps MK 4	Cardioid
TpFR	Schoeps MK 4	Cardioid
TpFC	Schoeps MK 4	Cardioid
TpC	DPA 4011	Cardioid
TpBL	Schoeps MK 4	Cardioid
TpBR	Schoeps MK 4	Cardioid
TpSiL	Schoeps MK 4	Cardioid
TpSiR	Schoeps MK 4	Cardioid
TpBC	Schoeps MK 4	Cardioid
BtFC	DPA 4011	Cardioid
BtFL	DPA 4011	Cardioid
BtFR	DPA 4011	Cardioid

4.4 Evaluation of Recording

A balanced mix was created of "Mars, The Bringer of War" from Gustav Holst's *The Planets*, using only the above described 22-microphone "main system". Numerous informal listening sessions have taken place to form a preliminary evaluation of the recording technique. Because this is the first 22.2 orchestral recording made at McGill University, there are no "reference recordings" with which to compare it in a formal listening test.

The “Mars mix” has been heard by fulltime and visiting faculty of the Graduate Program in Sound Recording at McGill University, as well numerous visiting researchers and recording engineers. In general, comments have been positive. Many have noted the large, coherent, and realistic orchestral image, natural depth of field, excellent instrument and sectional image clarity, and enveloping reverberation. The overall impression seems to be a naturalistic listening experience. The same recording has been heard as a part of informal listening sessions in four studios in Japan designed or equipped for 22.2 Multichannel Sound reproduction: 1) NHK Science & Technology Research Laboratories, 2) A dubbing studio at NHK’s Shibuya production center [64], 3) “Studio B” at Tokyo University of the Arts’ Senju Campus, 4) Yamaha Corp. in Hamamatsu. The recording was well received; listener comments and observations were consistent with those collected at McGill. It was also observed that the sound of the mix was generally consistent across all playback venues, which themselves varied in terms of size, acoustical treatment, and speaker radius. Further informal evaluations at the BBC, Rochester Institute of Technology, and TC Electronics have also yielded positive listener impressions. An excerpt of the recording was used as part of recent research by Zacharov et al. investigating a new method for multichannel sound evaluation [155].

4.5 Evaluation of Bottom Channels

4.5.1 Listening Test

A double-blind ABX listening test was designed to determine if within the context of highly dynamic orchestral music recordings, subjects could successfully differentiate between playback conditions with and without the three bottom channels. Although a seemingly simple task, this was felt to be a good first question to answer before moving on to a more complex perceptual evaluation of the contribution of the bottom channels toward orchestral music reproduction. All testing took place in Studio 22 (Figure 29).

Test stimuli consisted of three 25s excerpts from "Mars, The Bringer of War": 1) a relatively soft passage, 2) a passage ranging from mezzo-forte to forte, and 3) the very loud ending of the piece. Mixes of each stimulus were prepared with and without bottom channel content. A Neumann KU 100 dummy head was set at Studio 22's listening position. Playback of each stimulus was recorded, then analyzed using an integrated loudness measurement. Bottom and non-bottom channel mixes, per musical excerpt, were then level matched to within 0.2dB of each other.

4.5.1.1 Administration of Listening Test

Test subjects were seated at the listening position in Studio 22. Prior to performing the listening test, each subject took part in a brief training session during which they were given time to familiarize themselves with the three musical excerpts, the Pro Tools session being used as the testing interface, and the "with bottom channels" and "without bottom channels" conditions. For each trial, subjects were presented with one of the three musical excerpts on a repeating loop, and asked to compare mixes labelled "X" "A" and "B" by selecting between three VCA groups in Pro Tools. Subjects were instructed to identify the mix that was "different from X". (During preliminary tests, subjects found this to be a more logical task then identifying the mix that was "the same as X".) Subjects recorded their answers on an online form that also included a short demographic survey and comments section to be completed after the listening test. Subjects were also asked to rate the difficulty of the task. Each participant saw each excerpt three times, for a total of nine trials. Mix stimuli assignment to VCA groups (X, A and B) was randomized, as was the order of excerpt presentation. Playback of stimuli was time-aligned for seamless switching.

4.5.2 Results

14 subjects performed the listening test. All had at least two years of experience or training as recording engineers, and all reported having normal hearing. Each subject completed nine

trials for a total of 126 trials. The participant's response was marked "correct" if they successfully discriminated between the A and B stimuli (with X as reference) and "not correct" if they did not. Throughout the entire analysis, the overall success rate was examined as well as the success rates for each musical excerpt (soft, medium, and loud). The percentage of correct responses is shown in Figure 42. An overall success rate of 69% was achieved; a binomial test (Table 7) shows that this result is highly significant. When looking at the three musical excerpts individually, significant discrimination rates were achieved for both the medium (81%) and soft (67%) excerpts. However, the 60% discrimination rate for the loud excerpt was not significantly above chance. Participants rated the task difficult overall, giving it a mean rating of 3.9 (S.E. 3.8-4.1) on a scale from 1 (Easy) to 5 (Hard).

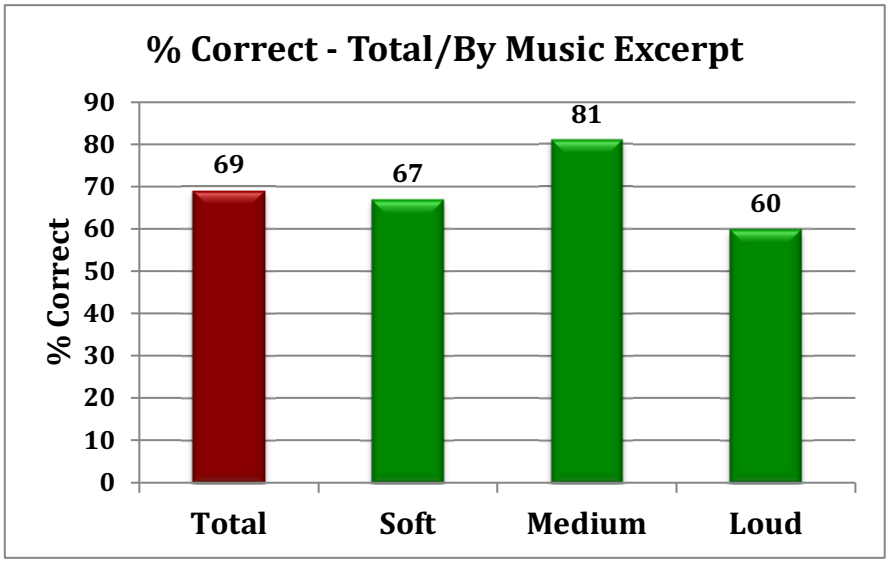


Figure 42: Percentage of correct responses

Table 7: Binomial Test

Data Group	Probability	95% Conf. Interval	*p*-value
Total	.6905	0.60-0.77	<0.001
Soft	.6667	0.50-0.80	0.044
Medium	.8095	0.66-0.91	<0.001
Loud	.5952	0.43-0.74	0.28

4.6 Discussion and Future Work

4.6.1 Informal Evaluations

Based on preliminary observations, it can be said that the proposed recording technique captures a broad, vertically anchored orchestral image with a natural depth of field and clear image localization, as well as highly enveloping ambience.

4.6.2 Bottom Channel Evaluation

In comments written in post-test surveys, as well as those made verbally to the examiners afterwards, most subjects commented on the subtlety and difficulty in detecting when the bottom channels were in use. This is also reflected in the analysis of the perceived difficulty rating. For orchestral music this is not surprising, especially considering how the bottom channels were mixed in comparison with the main front channels (typically 7dB lower in output). As such, a 69% probability of success ($p<0.001$) is considered a valid result in demonstrating the ability of listeners to discriminate between playback conditions with and without bottom channels.

Most subjects also commented on what they were listening for when attempting to discriminate between playback conditions. Many felt they could discern more low frequency information when the lower channels were active, particularly in the mezzo forte/forte excerpt. Not surprisingly, it was often observed that the image of the orchestra extended further towards the floor when the bottom speakers were active. Several subjects commented that the bottom channels contributed to a broadening of the orchestral image, a similar impression to what is described in concert hall acoustics as Apparent Source Width. This is quite interesting, as ASW is typically associated with lateral reflections [129].

4.6.3 Source Material for Spatial Audio Evaluation

Analysis showed that the "loud ending" musical excerpt had the lowest percentage of correct differentiation of playback conditions, and that the discrimination rate was not above chance. It is likely that for this loud passage the difficultly experienced by listeners was due to the dynamic envelope of the music. This passage was made up of brief fortissimo tutti orchestra chords separated by moments of silence. The large variance in the overall dynamic envelope makes it very difficult for the participants to find an appropriate place to "switch" between A, B and X. Rumsey has discussed the importance of choosing appropriate source material as stimulus within the context of listening tests designed to evaluate sound quality, noting that the choice of source material "can easily dictate the results of an experiment, and should be chosen to reveal or highlight the attributes in question. [109]" Similarly, ITU-R BS.1116-1 states that "Only critical material is to be used in order to reveal differences among systems under test. Critical material is that which stresses the systems under test. [105]" For these types of listening tests that seek to reveal subtle differences in sound quality, it is advisable to use source material that is relatively static in both dynamic envelope as well as spectrum for the entire length of the excerpt, thus making any differences between stimuli more apparent to the listener.

4.6.4 Future Work

A major hindrance to any serious subjective evaluation of the technique and its subsequent recordings is a lack of "reference" 22.2 optimized orchestral material with which to compare. The study covered in Chapter 5 will address this issue. The proposed technique will be setup and optimized to record several days of orchestral rehearsals. As a comparison, several other three-dimensional orchestral music recording techniques will be setup for simultaneous recording. This should yield several different recordings that can be used for extensive subjective comparison between large-scale three-dimensional recording techniques, as well as

further validation of the new technique proposed herein. A more comprehensive evaluation of the effectiveness of bottom channels in music reproduction is also required. It would be valuable to extend this investigation beyond orchestral music, and include recordings of other genres of music, such as chamber music, jazz, and pop/rock.

4.7 Conclusions

A three-dimensional orchestral music recording technique, optimized for 22.2 Multichannel Sound reproduction has been developed. The technique prioritizes the capture of a natural orchestral sound image with realistic horizontal and vertical extent, as well as a highly diffuse, enveloping reverberant sound field. Using the proposed technique, a recording was made of the National Youth Orchestra of Canada. Preliminary evaluations of the recording have been positive. A subsequent listening test showed that subjects can successfully differentiate between playback conditions with and without the use of the bottom channels in an orchestral music mix. More test recordings using the proposed technique are required, as well as further, more formal subjective evaluation of its effectiveness.

5 Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio

Abstract

A study was conducted to evaluate a recently developed microphone technique for three-dimensional orchestral music capture, optimized for 22.2 Multichannel Sound (22.2). The proposed technique was evaluated against a current 22.2 production standard for three-dimensional orchestral music capture, as well as a coincident, higher order ambisonics capture system: the Eigenmike™. Analysis of the results showed no significant difference in listener evaluation between the proposed technique and the current production standard in terms of the subjective attributes “clarity”, “scene depth”, “naturalness”, “environmental envelopment”, and “quality of orchestral image”.

5.1 Introduction

5.1.1 Recording Acoustic Music for 3D Playback

Chapter 4 introduced a new method for three-dimensional orchestral music recording, optimized for Japan Broadcasting Corp. (NHK)'s 22.2 Multichannel Sound (22.2) [156], [8]. The technique is designed to take advantage of several aspects of the 22.2 reproduction environment that make it uniquely suited to orchestral music reproduction. Featuring 10 playback channels at ear level, nine above the listener, and three below, 22.2 is one of the most advanced and robust of the currently standardized 3D audio formats [5]. Five frontal speakers at ear level with a reproduction angle of 120º, combined with three bottom (ground level) channels, allows for the creation of a large, coherent, stable orchestral image that gives the listener the impression of an idealized conductor's perspective. An even spatial distribution of surrounding loudspeakers allows for realistic reproduction of early and late reflections, and a reverberant field that is highly decorrelated at all frequencies. These factors are key to achieving strong levels of listener envelopment [124], [43], [150] as well as a dimensional broadening of the sound source image [123]. Chapter 4 [156] concluded that while the proposed recording technique performed well in informal listening tests that took place at five different 22.2 reproduction facilities, a more formal subjective evaluation was required.

Several authors have proposed microphone techniques for three-dimensional classical music recording, optimized for 9.1 or similar formats [72], [75], [71], [152]. However, few of these new techniques have been examined through formal subjective listening tests. Ryaboy investigated perceptual differences between two recording techniques: Double MS+Z and Twins Square [74] (see also: Sections 2.4.2, 2.4.3). Results of a double-blind listening test were reported as showing significant differences between the two techniques regarding "localization" (horizontal and vertical) and "perceived room size".

Hamasaki et al., introduced a technique for three-dimensional orchestral music recording as part of an investigation comparing 19.1 (22.2 without bottom channels) with reduced playback conditions: 17.1 (FL and FR removed), 10.1 (mid-layer only) and 5.1 [9]. Hamasaki and Van Baelen describe an updated version of the same technique in [10]. A study in [10] showed listeners rated Hamasaki's three-dimensional orchestral recordings significantly higher than stereo and 5.1 mixes of the same material for many subjective attributes, including "deep", "elevation", "spaciousness", "envelopment", and "good sound". The technique described by Hamasaki and his co-authors (or variations on said technique) has been used by NHK recording engineers for numerous orchestral music recordings, and is as such considered a current production standard optimized for 22.2. No known study has subjectively compared multiple three-dimensional capture methods optimized for 22.2.

5.1.2 Recording Array Comparisons for 5.1 Surround

Within the realm of 5.1 surround sound, there is far more literature exploring subjective comparisons and evaluations of recording techniques. Kassier et al. [146], and Hietala [157] examined differences between spaced (e.g. Fukada Tree) and semi-spaced (e.g. OCT surround) techniques. Within the context of an informal comparison, listeners consistently preferred Fukada Tree paired with Hamasaki Square [146]. Camerer and Sold [158], Kim et al., [147], Kamekawa et al., [36], Paquier et al., [159], A. Sitek and B. Kostek [160], and Peters et. al. [161] all undertook investigations that included evaluating perceptual differences or preferences between spaced, semi-spaced, and coincident surround recording techniques. These publications often investigated different aspects of multichannel sound, and as such, depending on the research question, certain spaced or semi-spaced recording techniques tended to perform better than others. However, a consistent trend found within these publications is that regardless of the subjective or preference attribute(s) being investigated, coincident techniques tend to be rated on the negative end of the spectrum. This

was true for 1st order ambisonics techniques [158], [147], [159], [160], [161], higher order ambisonics (HOA) [159], and Double MS [36].

5.1.3 Motivation

The primary aim of this study is to evaluate the effectiveness of the three-dimensional orchestral music recording technique proposed in Chapter 4, as compared with a current production standard for 22.2-optimized orchestral music capture in terms of salient spatial sound attributes. A secondary aim is to compare the performance of these two spaced techniques with a coincident, HOA-based capture system.

5.2 Recording Techniques Under Investigation

A detailed explanation of **Technique 1**'s design rational can be found in Chapter 4. Primarily direct orchestral sound is captured by a modified "Decca Tree" of five omnidirectional microphones, the middle three of which are outfitted with acoustic pressure equalizers [81]. Three directional microphones placed 1m above the stage floor provide signal for the bottom channels, vertically extending and anchoring the orchestral image. Widely spaced directional microphones capture decorrelated, spatially diffuse ambience, and are assigned to the remaining main layer and height channels (Figures 38–41). The technique is designed to retain the traditional "concert perspective" that is typical of most multichannel classical music recordings. Microphone orientation typically mirrors assigned playback channel orientation: for example, the TpFL microphone would have a horizontal orientation of around 60º, and a vertical orientation of approximately 45º.

Technique 2 was designed by Hamasaki and his co-authors, as described in [9] and [10]. The technique is a logical extension of Hamasaki's earlier publications on multichannel music recording, particularly "Reproducing Spatial Impression With Multichannel Audio", co-authored with Hiyama [47]. Direct sound from the orchestra is captured by an array of 5

hypercardioid microphones spaced at equal intervals across the sound stage. In [9], ambient sound is captured with an array of laterally oriented bi-directional microphones, an extension of the well-known "Hamasaki Square" [47] (see also: Section 2.2.7.2). The placement and spacing of the bi-directional microphones ensures minimal capture of direct and rear wall sound reflections, and that the ambient sound field is decorrelated across the audible frequency spectrum. Several of these ambience microphones are assigned to the front channels, to be mixed in if the recording engineer feels the orchestral sound is too "dry". In [10] this approach is updated using vertically oriented hypercardioid microphones as height channels. This version of the technique is representative of current 3D orchestral music recordings being made by NHK recording engineers, and thus can be considered a de-facto production standard for 22.2. Neither [9] nor [10] specify microphones for the bottom channels. For this study, three Sanken CUB-01 miniature boundary microphones have been added to the technique, each placed as far down-stage as possible (Figures 43, 44). These microphones were chosen for their minimal visual impact, an important factor in broadcast sound recording, as well as to contrast with the bottom channel microphones used in Technique 1.

As seen in the introduction, several studies comparing multichannel recording techniques have included coincident microphone systems. When considering the complexity, cost, and time associated with setting up either Techniques 1 or 2, the potential advantages to using a single-point, ambisonics-based capture system become obvious. As such, for this study, the Eigenmike (em32) was chosen as a **3rd recording technique**. The em32 from Mh acoustics is a spherical microphone array where each of the 32 capsules is calibrated for magnitude and phase response. The accompanying software *Eigenstudio* converts the microphone signals into 3rd order ambisonics b-format audio files. For this study, 16 channels were recorded following the ACN channel order convention with N3D normalization [162].

5.3 Setup and Optimization of Recording Techniques

The three techniques under investigation were installed in Pollack Hall, a medium sized concert hall with a seating capacity of 590. The hall measures 36m long by 18m wide by 12m high. Reverb times for the empty hall are shown in Table 8. The side and stage walls are equipped with acoustic curtains designed to decrease RT60. For this study, all acoustic curtains were "out" (removed) except for the stage curtains which were set to "¾ out" to control onstage reflections.

Table 8: RT 60 for Pollack Hall

f (Hz)	**63**	**125**	**250**	**500**	**1k**	**2k**	**4k**
RT60	2.3s	2.0s	1.7s	1.8s	1.8s	1.7s	1.4s

The microphones for all three techniques were installed the day before a week of orchestral rehearsals, with the goal of having all three techniques fully optimized before recording the final dress rehearsal. All microphones were routed to RME Micstacy preamps and A/D converters. Two streams of optical MADI output from the Micstacys were routed via fibre optic lines to Studio 22 (Figure 29), a multichannel audio mixing room in an adjacent building. Studio 22 is equipped with 28 full-range, two-way loudspeakers (Musikelectronic Geithain GmbH *M-25*) and a stereo sub-woofer, arranged for reproduction of both 22.2 Multichannel Sound, and Auro 3D 9.1. The studio fulfills ITU-R BS.1116 requirements [105].

5.3.1 Placement and Optimization: Techniques 1 and 2

For Techniques 1 and 2, microphone choice and placement was based on [156], [9] and [10], as well as extensive experience recording orchestral music (Table 9, Figures 43 and 44). A current NHK production engineer provided valuable insight as to the optimization of Technique 2. Placement of the front five microphones for Technique 2 was based on

available hanging points, and the increased "reach" of hypercardioid microphones as compared with omnidirectional microphones. Like the "Hamasaki Square", Technique 2 included 3 frontal ambience (FrAmb) microphones to be mixed in with the direct orchestral sound as necessary. Microphones for both techniques were either hung or placed on telescopic stands in the hall, depending on their desired height and location. Adjustments were made based on monitoring the orchestra's rehearsals.

Table 9: Microphones used per technique. For a detailed explanation of channel naming, see Figure 37 and Table 1

Channel	Technique 1	Polar Pattern (1)	Technique 2	Polar Pattern (2)
FL	Schoeps MK 2S	Omnidirectional	Neumann KM 185	Hypercardioid
FLc	Schoeps MK 2H	Omnidirectional	Neumann KM 185	Hypercardioid
FC	Schoeps MK 2H	Omnidirectional	Senn. MKH 8050	Supercardioid
FRc	Schoeps MK 2H	Omnidirectional	Neumann KM 185	Hypercardioid
FR	Schoeps MK 2S	Omnidirectional	Neumann KM 185	Hypercardioid
BL	Schoeps MK 21	Wide Cardioid	Schoeps MK 8	Bi-directional
BC	Neumann KM 120	Bi-directional	Neumann KM 120	Bi-directional
BR	Schoeps MK 21	Wide Cardioid	Schoeps MKH 8	Bi-directional
SiL	Neumann KM 184	Cardioid	Senn. MKH 30	Bi-directional
SiR	Neumann KM 184	Cardioid	Senn. MKH 30	Bi-directional
TpFL	Schoeps MK 4	Cardioid	Neumann KM 185	Hypercardioid
TpFC	Schoeps MK 4	Cardioid	Neumann KM 185	Hypercardioid
TpFR	Schoeps MK 4	Cardioid	Neumann KM 185	Hypercardioid
TpC	Schoeps MK 41	Supercardioid	Schoeps MK 41	Supercardioid
TpBL	Schoeps MK 4	Cardioid	Neumann KM 185	Hypercardioid
TpBC	Schoeps MK 4	Cardioid	Senn. MKH 8050	Supercardioid
TpBR	Schoeps MK 4	Cardioid	Neumann KM 185	Hypercardioid
TpSiL	Schoeps MK 4	Cardioid	Senn. MKH 50	Supercardioid
TpSiR	Schoeps MK 4	Cardioid	Senn. MKH 50	Supercardioid
BtFL	DPA 4011	Cardioid	Sanken CUB-01	Cardioid
BtFC	DPA 4011	Cardioid	Sanken CUB-01	Cardioid
BtFR	DPA 4011	Cardioid	Sanken CUB-01	Cardioid
FrAmb L	N/A	N/A	Senn. MKH 800	Bi-directional
FrAmb C	N/A	N/A	Senn. MKH 800	Bi-directional
FrAmb R	N/A	N/A	Senn. MKH 800	Bi-directional

5.3.2 Placement and Optimization: Technique 3

Professional recording engineers tend to place microphones based on previous experience, known best practices, and most importantly, what they are hearing inside the recording venue and the monitoring environment. Recording with the Eigenmike, as such, presents a unique set of challenges. There is little published information detailing placement and optimization strategies for music recording using spherical HOA microphones, especially where the desired sound scene utilizes the traditional "ensemble in front, ambience surrounding" perspective. Daniels discusses several experimental recordings done with spherical HOA microphones, mixed for two-dimensional playback [163]. For a large ensemble recording where the goal was to keep the ensemble imaged in front of the listener, Daniels placed a 20 capsule HOA sphere near several other (unidentified) 5.1 microphone arrays. Barrett [164] and Power [165] both used the Eigenmike for music recording as part of their respective studies, but provided no methodology for placement and/or optimization.

The 32-channel output from the Eigenmike is recorded to a computer running *Eigenstudio* software via firewire output from an mh acoustics EMIB termination box. There is no effective way to monitor a 22.2 rendering of these signals in real time. For this study, the beampattern of an omnidirectional microphone was output from the *Eigenstudio* recording software and routed to Studio 22 for monitoring. Though not ideal, this gave the recording team some degree of information regarding distance and balance of instrumental groups for microphone placement optimization. The result was the Eigenmike being placed in the centre of Technique 1's "Decca Tree" (Figures 43 and 44).

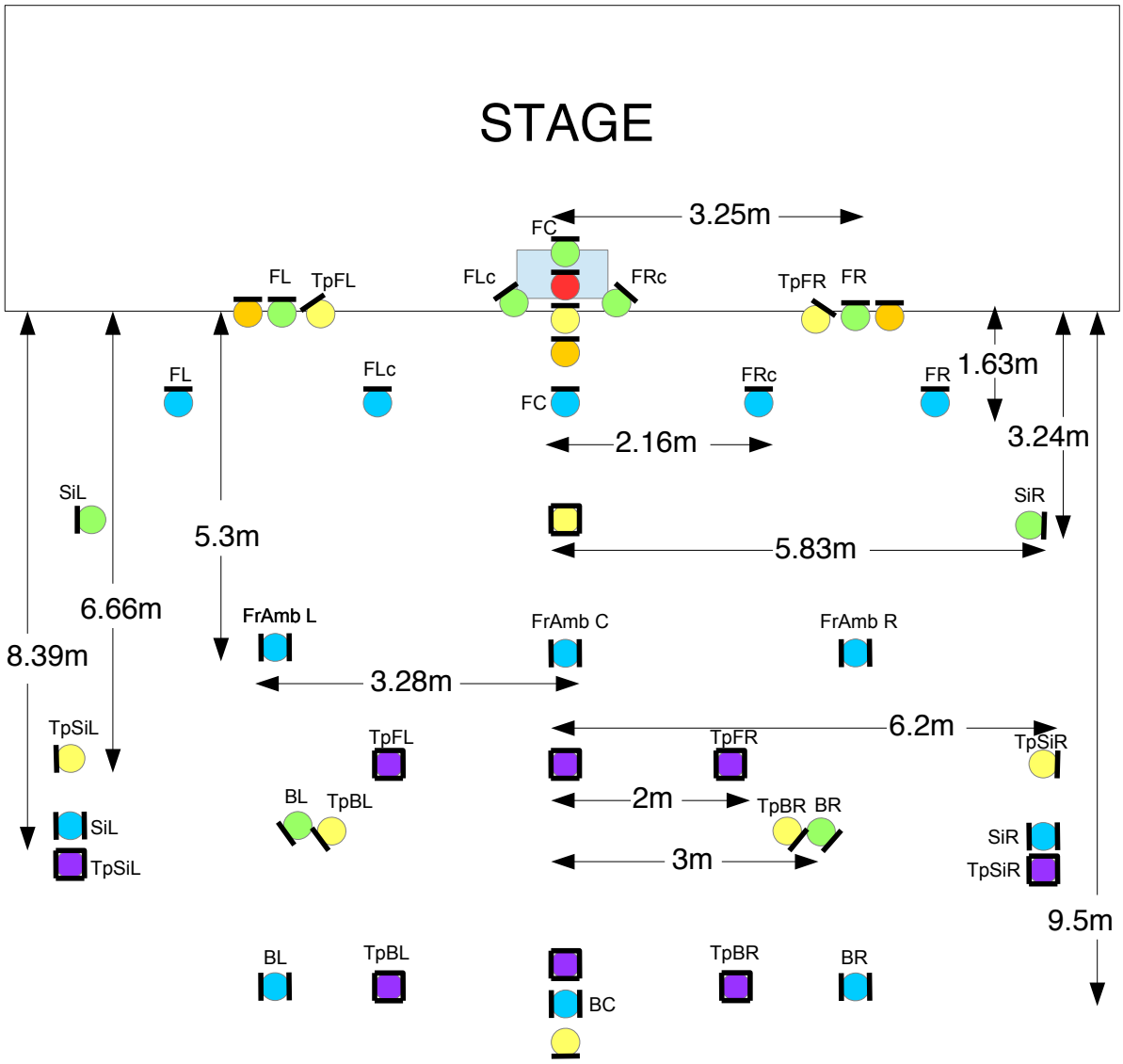


Figure 43: Microphone placement, overhead view. Height is referenced to stage floor.

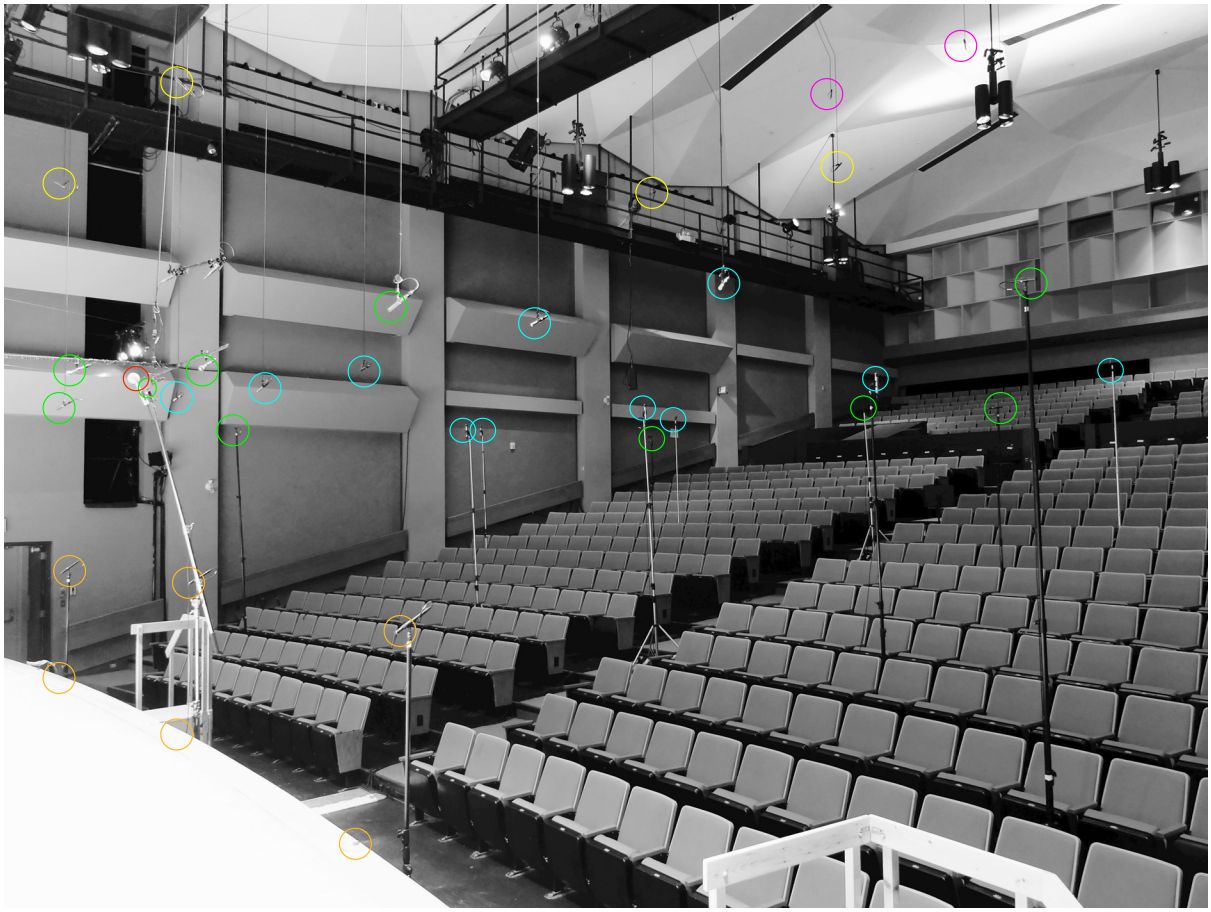


Figure 44: Orchestral capture microphones, as seen from on stage. Colours correspond to Figure 43. Not all microphones can be seen.

5.4 Experimental Design

5.4.1 Creation of Stimuli

All three techniques simultaneously captured the final orchestral dress rehearsal. Techniques 1 and 2 were recorded to Pro Tools 10 at 96kHz/24bit resolution. Spot microphones for the woodwinds, bass and tympani were also recorded. Technique 3 was recorded to a separate laptop computer, whose audio interface was locked to the RME Micstacys' master clock. A single, 30-second musical excerpt was chosen as stimuli. The musical passage contains dense orchestration representative of the piece it was derived from (Tchaikovsky's 5th Symphony), and has a fairly even dynamic envelope.

The techniques under investigation were balanced by a team of three recording engineers with extensive professional experience recording and mixing orchestral music. It was observed that Technique 2 did not contain enough low frequency content for a satisfying mix, largely due to the low frequency roll-off typical of highly directional microphones. In accordance with Hamasaki and Hiyama [47], the FL and FR omnidirectional channels from Technique 1 were added to Technique 2's mix, low-passed at 200Hz. Once ideal balances were achieved, 24-channel mixes of the musical excerpt were made for each technique.

To create an optimal 22.2 mix of the Eigenmike recording, a custom-made decoder for the speaker positions in Studio 22 was built. Using Heller's Ambisonic Decoder Toolbox [166] the decoder matrix for a dual band All-Round decoder [167] was calculated, which allowed for adjustment of balance between high and low frequencies with phase matched filters per [168]. The crossover frequency (400Hz) and the gain for the balance (+1dB HF) were chosen to perceptually match the mixes from Techniques 1 and 2.

The three resultant stimuli were level matched by ear. These results were then confirmed by objective means. A Neumann KU-100 Dummy Head microphone was placed in the listening position at ear level, and used to record the playback of each stimulus. Integrated loudness measures (LUFS 9) were then performed for each recording. All stimuli were found to be within 0.5dB of each other.

5.4.2 Design and Implementation of Listening Test

A double-blind listening test was designed to identify possible salient perceptual differences between recordings made using the three techniques. The test was implemented using Cycling 74's Max/MSP software. Twenty-three subjects performed the test: all were either current students or faculty within the Graduate Program in Sound Recording at McGill University. All subjects reported having normal hearing.

Subjects were seated in Studio 22's listening position, were explained the testing conditions, and given time to familiarize themselves with the testing interface and stimuli (Figure 45). Definitions of the perceptual attributes being investigated were provided both verbally and in written form (Table 10). Based on previous research into spatial audio evaluation [111] [155], "clarity", "scene depth", "naturalness", "environmental envelopment" and "sound source envelopment" were chosen. "Quality of orchestral image", a new term, was included based on the sonic imaging goals of Technique 1.

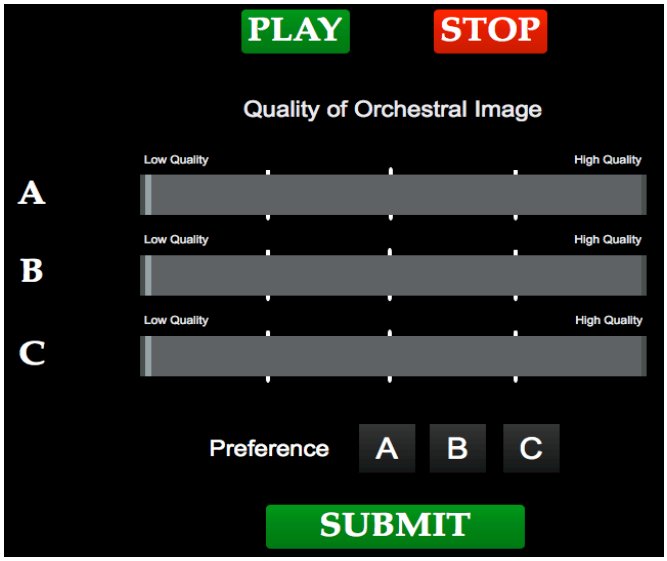


Figure 45: Testing GUI

For each trial, subjects were asked to evaluate mixes labelled "A", "B", and "C" for a given attribute, using a set of continuous sliders (0-100). Anchor words were provided at the extremes of each slider. Since absolute anchors were not given at intervals along the scales, these measurements are relative and not absolute. To reduce scaling bias, subjects were instructed to always rate the mix they felt was the "most" or "best" of a given attribute as 100%, then using that as a reference, rate the other two accordingly. More than one mix could be rated 100%. After completing ratings for a given attribute, the subject was asked to choose the mix they preferred, regardless of the perceptual attribute being investigated. Subjects could switch between playback of "A", "B", and "C" or stop the audio at any point. Playback

of stimuli was time-aligned and continuously looped. The test was administered in blocks of three trials per attribute, for a total of 18 trials. This was done to allow subjects to focus on one perceptual attribute at a time. For each trial, stimulus assignments to "A", "B", and "C" were randomized. The order of attribute trial blocks was also randomized. Subjects were instructed to set a comfortable listening level before completing the first trial, and then leave the level unchanged for the remainder of the test. At the test's midway point was an enforced rest period of 1 minute. Subjects took an average of 25 minutes to complete the test. Upon completion, subjects were instructed to fill out a short demographic survey.

Table 10: Sound attribute names and definitions

Attribute Name	Definition
Sound Source Envelopment	The sense of being enveloped by a group of sound sources. [109]
Environmental Envelopment	The sense of being enveloped by reverberant or environmental sound. [109]
Clarity	"The clearer the sound, the more details you can perceive in it." [118]
Naturalness	A sound is natural if it gives you a realistic impression, as opposed to sounding artificial. [118]
Quality of Orchestral Image	A "high quality" orchestral image is defined as being a cohesive, anchored sound image, with well-defined horizontal and vertical extent.
Scene Depth	The overall impression of the depth of the sound image. Takes into consideration both overall depth of scene, and the relative depth of the individual sound sources. [155]

5.5 Results

5.5.1 Attributes

To remove possible scale biases, all scores were normalized as z-scores within each attribute for each participant. The mean ratings, for each attribute, for each technique are shown in Figure 46. For all attributes, Techniques 1 and 2 were rated quite high and similar, whereas Technique 3 was rated quite low. The results of a one-way repeated measures ANOVA on each attribute can be seen in Table 11, and show that the differences seen in the ratings for

each attribute is significant. Post hoc Bonferroni corrected pair-wise t-tests show that significant differences exist between Technique 3 and both other techniques for all attributes. Post hoc tests also show that a significant difference between Techniques 1 and 2 exists only for the attribute "sound source envelopment".

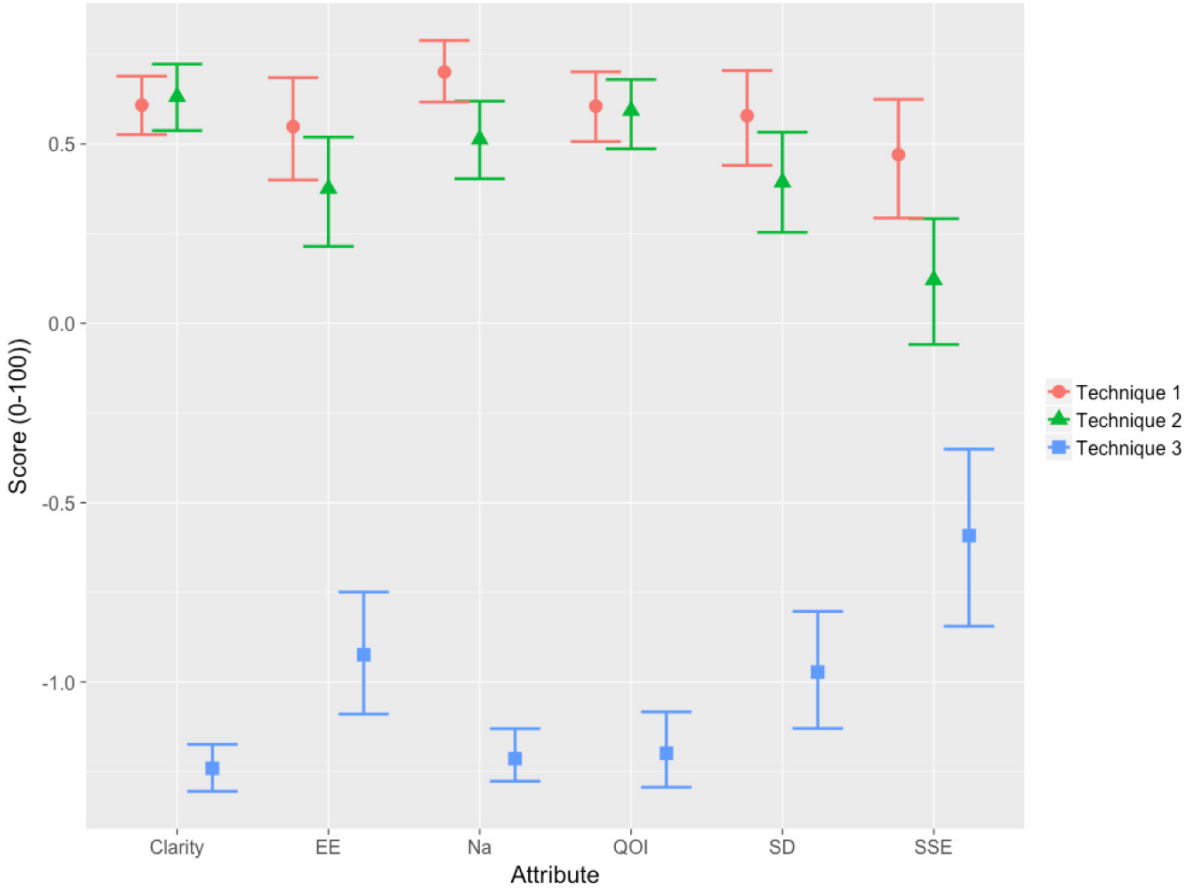


Figure 46: Average rating for each attribute. Colour represents the three different recording techniques. EE = environmental envelopment, Na = naturalness, QOI = quality of orchestral image, SD = scene depth, SSE = sound source envelopment.

Table 11: ANOVA and Post Hoc on Attribute Ratings

Attribute	Tech1 Mean (SD)	Tech2 Mean (SD)	Tech3 Mean (SD)	F (df)	p	p Tech1 vs Tech 2
Clarity	.609 (.362)	.632 (.339)	-1.24 (.276)	302 (2, 44)	<.001	
EE	.549 (.604)	.376 (.649)	-.924 (.777)	31.4 (2, 44)	<.001	
QOI	.606 (.422)	.593 (.389)	-1.20 (.435)	301 (2, 44)	<.001	
Na	.701 (.367)	.513 (.467)	-1.21 (.299)	316 (2, 44)	<.001	
SD	.579 (.577)	.393 (.619)	-.972 (.730)	35.2 (2, 44)	<.001	
SSE	.470 (.714)	.122 (.707)	-.591 (1.05)	8.45 (2, 44)	<.001	<.01

5.5.2 Preference

Preference for each technique was measured by counting the number of times a given technique was chosen as the most preferred (Table 12).

Table 12: Contingency Table for Preference

	Technique 1	Technique 2	Technique 3	Total
Count	243	167	3	413
% of total	59%	40%	0.7%	

A Chi-Square test shows that difference in preference is significant, $x^2(2) = 218.6$, $p < .001$. Technique 3's large deviation from the expected random frequency (~33.3%) is likely the cause of this significant difference. Given that Technique 3 received so few preference counts, it can be dropped from the analysis, and a binomial test on the counts for Techniques 1 and 2 can be performed. In this case, Technique 1 was significantly preferred over Technique 2, $p < .001$ with a confidence interval of 0.54-0.64.

With this test design, attributes and preference were rated by the participants simultaneously. It is therefore important to know if the attribute being rated for a given trial influenced preference. In this case, the attribute being rated did not have a significant effect on the preference ratings $x^2(2) = 13.18$, $p = 0.21$.

5.5.3 Correlation of Attributes

Having the same three stimuli (techniques) rated along several different attributes allows for investigating rating correlation between said attributes. The results (Table 13) show that there is a high positive correlation between all pairs of attributes. The correlation coefficients are significant to at least the $p = 0.05$ level.

Table 13: Pearson correlation matrix between attributes

	Clarity	EE	QOI	Na	SD	SSE
Clarity	1.00	-	-	-	-	-
EE	0.64	1.00	-	-	-	-
QOI	0.88	0.60	1.00	-	-	-
Na	0.87	0.60	0.84	1.00	-	-
SD	0.70	0.48	0.72	0.65	1.00	-
SSE	0.43	0.33	0.37	0.33	0.48	1.00

The relationship between attribute ratings for each stimulus and preference ratings is visualized in Figure 47. It shows that when a given technique is preferred, it also receives much higher ratings along all attributes. The magnitude of this difference appears to be similar between all attributes.

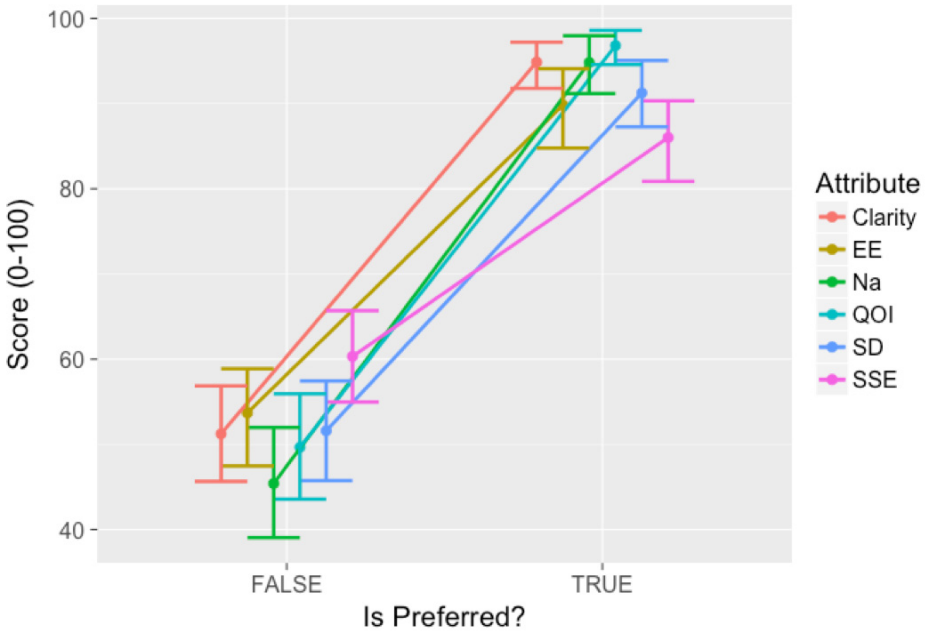


Figure 47: Average rating for each attribute according to preference. Colour represents the attribute.

5.6 Discussion

5.6.1 Overall Performance of Recording Techniques

Figure 46 shows a clear similarity of ratings between Techniques 1 and 2 for all subjective attributes under investigation: "clarity", "scene depth", "naturalness", "environmental envelopment", "sound source envelopment", and "quality of orchestral image". Given this, and its consistently high mean scores across all attributes, the three-dimensional recording technique proposed in Chapter 4 should be considered a well-performing, valid production technique for three-dimensional orchestral music recording. Additionally, concepts from both Technique 1 and 2 could also easily be combined to form any number of hybrid techniques. For example, broadcast recordings involving picture would benefit from the bottom channel microphone design from Technique 2, which is more visually transparent.

Also very clear are the consistently low scores across all perceptual attributes for Technique 3. This matches well with the trend observed in previous research comparing two-

dimensional recording techniques (Section 5.1.2). For example, [158] observed a lack of "depth" and "adequate spatial impression" for the Soundfield MKV, a 1st order ambisonics recording system. These observations are echoed in the current study, with the Eigenmike performing poorly for "scene depth", "environmental envelopment" and "sound source envelopment". Spherical HOA microphones, although a convenient alternative to large spaced microphone arrays, may not yet be suited to professional music recording, especially given the monitoring difficulties discussed in Section 5.3.1.

The design rational behind Techniques 1 and 2 share similar goals in terms of spatial impression, clarity, and stability of ensemble imaging; both techniques are based on separate capture components for direct and ambient sound, and both techniques use large spacing between microphones, particularly those intended to capture decorrelated ambience. The major design differences between the two techniques are primarily related to microphone type and placement. Even there, however, there are enough similarities that the two techniques were able to share certain microphone channels between them. Techniques 1 and 2 both provide a great deal of flexibility at the mixing stage: either technique could yield any number of possible versions of a particular sound scene. As such, despite the technical differences between the two techniques, both could be used to create aesthetically similar mixes, which was likely the case in this study, since all three techniques were mixed by the same team of engineers. This may explain some of the similarity in ratings between techniques 1 and 2. Technique 3, by comparison, offers far less flexibility during the mix stage in terms of changes to the reproduced sound scene. Had Technique 3 not been included in this study, there may have been a greater variation in ratings between Techniques 1 and 2.

5.6.2 Naturalness and Sound Source Envelopment

"Naturalness" appears frequently as a subjective attribute in multichannel audio evaluation, and has been shown to correlate strongly with the impression of "presence" [111] and overall

preference of sound quality [169]. Frequently observed by both subjects and researchers were unpleasant and unnatural "out of phase" sonic artefacts present in Technique 3. This may explain why amongst all attributes, Technique 3's mean rating was lowest for "naturalness". In this study, a lack of perceived "naturalness" may also be an issue of perspective bias. Techniques 1 and 2 deliver a "cinematic" perspective for reproduced orchestral music, with the orchestra appearing entirely in front of the listener – a perspective most listeners have grown accustomed to. Technique 3, however, presents a much wider orchestral image, with direct sound sources covering almost 180° of frontal sound, likely due to the spherical nature of the Eigenmike. It is possible that the more "wrap-around" direct sound perspective delivered by Technique 3 is also perceived as being "unnatural".

Rumsey has written, "Envelopment, on the other hand, must be subdivided into environmental envelopment and source-related envelopment, the former being similar to LEV in concert halls and the latter to envelopment by one or more dry or direct foreground sound sources." [109] It was assumed that Technique 3's wider orchestral image would be rated highly for Sound Source Envelopment. However, although that attribute represented Technique 3's highest rated mean, it still scored well below Techniques 1 and 2. Clearly, listeners did not find a wider "wrap around" orchestral image to be more enveloping. In this study, the listeners' impression of "sound source envelopment" may be closer to Griesinger's concept of Continual Spatial Impression [43], a fusion of continuous direct sound and reflected energy that results in a sense of envelopment connected to the sound source. Technique 1 appears to best represent this type of spatial impression.

5.6.3 Cultural Bias in Preference

Technique 1 was created by a graduate of McGill University's Graduate Program in Sound Recording. That this technique was significantly preferred by current students and faculty within that same program could point to a strong bias within the results. Arrangements have

been made to perform the same listening test at Tokyo University of the Arts to investigate possible cultural trends within 3D microphone technique preference.

*****

5.7 Addendum: Additional Testing and Confirmation of Results

In the interest of further validating the results reported in Section 5.5, the experiment described in Section 5.4 was repeated at Tokyo University of the Arts (GEIDAI), with a pool of subjects drawn from students and faculty in the Department of Musical Creativity and the Environment.

5.7.1 Listening Room

The listening test took place in Studio B at Tokyo University of the Art's Senju Campus. The room is acoustically treated, with a measured T60 of 340ms, and a background noise level that does not exceed NC15. The room's volume is 208.1m3, substantially larger then McGill's Studio 22 (116.67m3). Studio B is equipped with 22 KS Digital C5 loudspeakers and two KS Digital ADM B2 subwoofers, arranged for 22.2 reproduction, as per ITU standards outlined in BS.2051-0 [5].

5.7.2 Subject Pool

12 subjects performed the listening test, all either current students or faculty at GEIDAI. All subjects reported having normal hearing. Most subjects had more than 10 years of musical training, had completed or were currently enrolled in a technical ear training program, and had previous experience listening to 3D audio. As with the McGill listener pool (Group 1), subjects took an average of 25 minutes to complete the test.

5.7.3 Listening Test

The test was administered in the same way as described in Section 5.4. Written and verbal instructions and attribute term definitions were provided in both English and Japanese. Subjective attribute names and definitions were translated from English to Japanese by a native Japanese speaker who is also fluent in English, and who has lived in North America for several years. These translations were confirmed by several native and non-native Japanese speakers who are fluent in both languages.

5.7.4 Results and Discussion

An examination of the data from the GEIDAI subjects (Group 2) found variation in how listeners used the rating scale, e.g., some subjects contracted their responses into a smaller range of the scale than others. To equalise these differences, all scores (groups 1 and 2) were normalized as z-scores within each attribute for each participant.

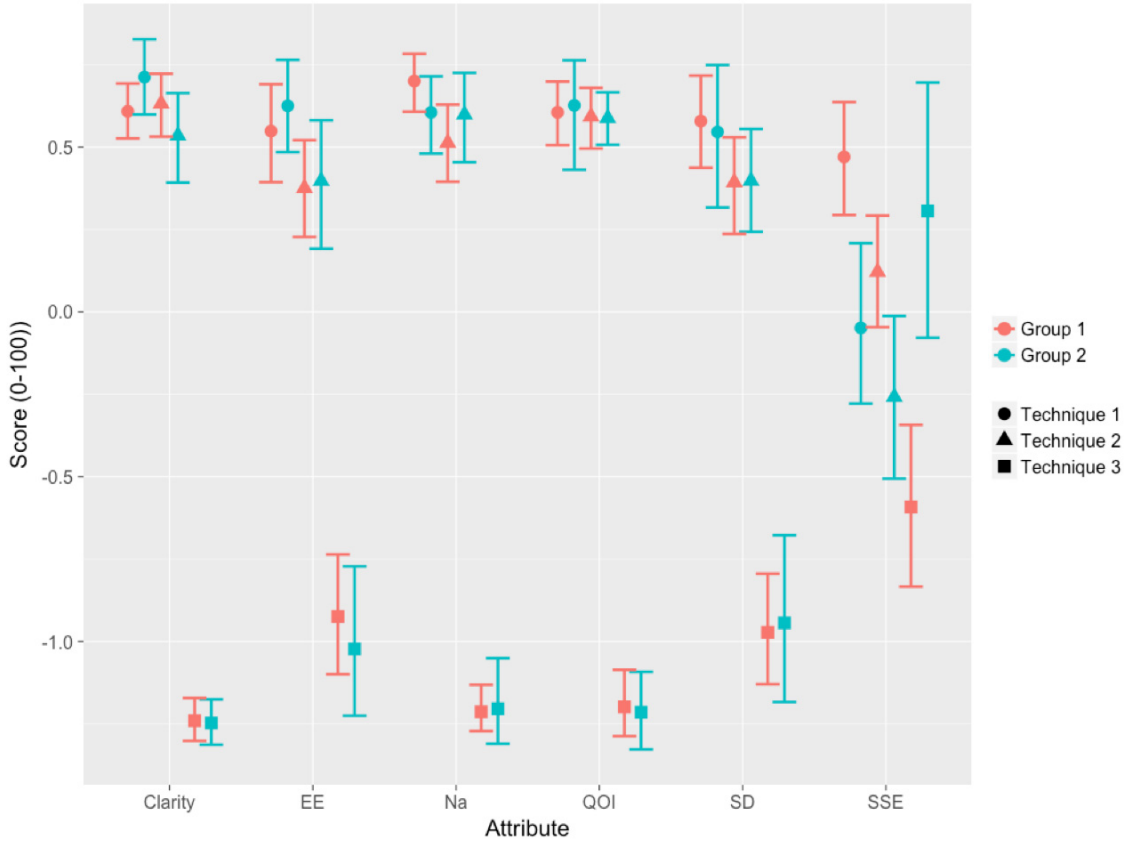


Figure 48: Attribute ratings by recording techniques, by listener group.

Figure 48 shows that attribute ratings are very similar between the McGill and GEIDAI participants. A series of pairwise t-tests (Bonferroni correction) found no significant differences between the two listener groups for all attributes (p= 1 for all tests) except for "sound source envelopment". For "sound source envelopment" differences do not exist between listener groups for Techniques 1 and 2 (p = 0.06 and p = 0.52 respectively) but do exists between groups for Technique 3 ($p < 0.001$).

Results shown in Figure 48 confirm the findings of the original study: the three-dimensional orchestral music capture technique proposed in Chapter 4 is a valid production technique for immersive content creation. It is interesting to note the significant difference in "sound source envelopment" ratings for Technique 3 between the two participant groups. With the McGill group, it is theorized in section 5.6.2 that subjects understood "sound source envelopment" to mean a sense of envelopment associated with the sound of the orchestra that is not necessary directly related to its perceived physical extent. Though provided with the same attribute definitions and verbal instructions, it seems likely that the GEIDAI listeners understood "sound source envelopment" as being more directly connected to the horizontal extent of the orchestral image, i.e. the direct sound of the recording. Perhaps Japanese listeners can more easily associate the concept of "envelopment" with direct sound sources than their North Americans counterparts. Studio B at GEIDAI has a larger speaker radius than McGill's Studio 22, and this may have contributed to a change in mix imaging that led to an even greater sense of being surrounded by direct sound components when listening to the HOA example. Further analysis of possible cultural or social differences between attribute ratings and preference data will be performed and discussed as part of a future study.

6 Listener Discrimination Between Common Speaker-Based 3D Audio Reproduction Formats

Abstract

A study was undertaken to determine whether listeners could discriminate between four currently standardized three-dimensional audio formats within the context of reproduction of acoustic music. Results of a double-blind listening test showed that listeners could discriminate between NHK 22.2 Multichannel Sound (22.2) and several other, lower channel count 3D reproduction formats with a high degree of success, regardless of the musical stimulus. Listeners were also able to discriminate between three relatively similar 3D audio formats: ATSC 11.1, KBS 10.2, and Auro 9.1, though with significantly less success than when the 22.2 format was involved. This suggests each of these formats deliver a perceptually different listening experience, with 22.2 being particularly different from the other formats under investigation. A small library of high-quality 3D audio recordings was created to facilitate the study.

6.1 Introduction

Over the last decade, numerous three-dimensional audio playback formats have been introduced and standardized for cinema, broadcast, and home theatre environments [5], [7], [77], [6], [69]. These formats differ in terms of number of speakers, speaker positions in the horizontal and vertical planes, and workflow concept: channel-based, object-based, or some hybrid of the two. Each system possesses inherent pros and cons in terms of reproduction of acoustic music, i.e., music performed using primarily acoustic instruments, such as classical, folk, or jazz. Recent research from the Graduate Program in Sound Recording at McGill University [143], [152], [156], [94], [95], Lee and Gribben [82], [170], Hamasaki and Van Baelen [10], Kim et al. [171], [172], and others has explored developing and evaluating music production and reproduction techniques for 3D audio. Much of the work done at McGill University has used NHK 22.2 Multichannel Sound (22.2) as the primary audio format. In recent years, work has been done by researchers at NHK's Science and Technology Research Laboratories (STRL) to introduce simplified, consumer-oriented playback systems for 22.2 [11], [66], [173], including loudspeaker-based systems with a reduced number of channels, and binaural headphone-based reproduction. However, the complexity and cost of the format still presents a daunting challenge to both content creators and end users.

The primary goal of this study is to investigate whether the same listening experience delivered by 22.2 for the recording and reproduction of acoustic music can be achieved with smaller scale, currently standardized 3D audio formats. Or, put another way: can listeners discriminate between acoustic music recordings created for 22.2 and remixes of the same content for reduced-channel reproduction formats? This knowledge is important for 3D audio content creators and distributers to better understand how their work may translate from one format to another, and whether there is adequate justification for continuing to pursue higher

channel-count reproduction systems, such as 22.2, in terms of delivering a perceptually unique experience to consumers. High-quality testing material that covers a range of common 3D audio formats, including 22.2, is currently in short supply. Such material is critical to examine the subtle perceptual differences that may or may not exist between different 3D audio formats. The secondary aim of this study is the creation of a library of such testing material, utilizing recordings made by sound engineers well versed in contemporary 3D audio production techniques. This library will be made available to other researchers.

6.2 Previous Research

Hamasaki et al. showed that 22.2 is easily discernible from, and generally more favourably rated than 5.1 surround sound and 2.0 stereo, especially for sound field reproduction [2]. In another study, Hamasaki et al. found that over a wide listening area, listeners rated "presence" significantly higher for 22.2 as compared with an almost identical playback condition that did not use the FL and FR channels [9] (see Figures 36 and 50 for 22.2 speaker layout). "Presence" is defined as "the sense of being inside an enclosed space" [109]. In that study, a 22.2 recording of Tchaikovsky's Symphony No. 6 was used as stimulus. An excerpt was remixed for other playback conditions under investigation by a professional mixing engineer. The mixing engineer "created the best balance, the best spatial impression and the best sound stage with each sound reproduction system at the center of the listening area […] Each sound stimulus was also carefully adjusted to provide the same loudness and same impression of reverberation. [9]"

Several authors have compared 22.2 with smaller-scale 3D reproduction formats [68], [55], [155], but results have been somewhat conflicting. Kim et al. compared 22.2 with 11.2 (Samsung), 10.2 (Samsung) and 10.2 (USC). Comparing the systems in terms of "Overall Quality", listeners perceived "little difference" between Samsung 10.2 and 22.2 [68]. Three types of test materials were used: two segments from film soundtracks, and one music

excerpt, which was comprised of a clip from a 5.1 mix of The Eagle's "Hotel California" with additional reverb added to fill out the 22.2 playback condition. 22.2 mixes were downmixed following passive coefficient schemes to create content for the other speaker layouts. Kim et al. discuss the possible influence of program material on the audibility of perceptual differences between playback conditions, concluding that the results "should be verified with wider variety of program material, and also in different applications. [68]"

Shim et al. compared 22.2 with 10.2 (KBS, same as in [68]) and 5.1, in terms of several perceptual attributes [55]. Stimuli were derived from film soundtracks, and did not prominently feature acoustic music. Stimuli were not downmixed, but remixed for each reproduction format. Listeners rated 22.2 "significantly better" than 10.2 for the perceptual attributes "naturalness", "listener envelopment" and "overall". The authors concluded "the more loudspeakers are used, the better attributes are reproduced as commonly known. [55]"

Zacharov et al. assessed several "next generation" audio systems within the context of evaluating the Multiple Stimulus Ideal Profile Method. As with the previous two studies, most stimuli were not music specific, in this case using excerpts primarily drawn from radio dramas [155]. The sole music stimulus used in [155] was an excerpt of the ending of Holst's "Mars: The Bringer of War", from the same recording described and evaluated in Chapter 4 [156]. This excerpt, originally recorded and mixed for 22.2, was then downmixed for the other systems under investigation (11.1 and 5.1). Downmixing to 5.1 utilized a passive coefficient scheme from [67]; a similar scheme of the authors' own design was used for downmixing to 11.1. Results of a listening test showed the "ideal profile" being statistically similar to both 22.2 and 11.1, suggesting little difference between the two formats.

6.2.1 Critical Testing Material

Zacharov et al. conclude that a "set of critical test material is also vitally important in the development of effective methods for assessing advanced sound systems. [155]" It is well known that when comparing different reproduction systems, critical material that stresses the systems under test is necessary [105]. It is questionable whether the music stimulus used in [68] can be defined as "critical material". Music recordings that are of poor quality, or those that are excessively dynamic in nature, make the task of discriminating between or rating subtle differences extremely difficult, especially within the context of "ABX" or "Triad" style listening tests. Chapter 4 [156] showed that the ability of listeners to discriminate between 22.2 mixes with and without the lower channels present was at least somewhat dependent on the dynamic and spectral envelopes of the audio excerpts used as stimuli. The orchestral music stimulus used by Zacharov et al. [155] is an excerpt with an extreme dynamic range, and may have limited the ability of subjects to compare the different formats under test. Rumsey has noted that the choice of source material for listening tests designed to evaluate sound quality "can easily dictate the results of an experiment, and should be chosen to reveal or highlight the attributes in question. [109]" A recent study by Francombe et al. comparing various spatial audio reproduction formats found listeners preferred both 9.1 and 5.1 over 22.2 for a wide range of musical material [174]. The authors, however, acknowledge that this result may in part be due to their inexperience producing content for more complex multichannel formats [174]. This all points to a strong need for studies examining perceptual differences between 3D audio formats where the stimuli are derived from high-quality recordings made by professional sound engineers who are experienced in creating content for the different formats under investigation.

6.3 Creation of Stimuli

One of the primary aims of this study was to create high-quality 3D music recordings that could be used in a wide range of future listening tests. The film, television, and radio drama content used as stimuli in [68], [55], [155] were produced primarily from either existing stereo or 5.1 multitrack recordings or content that had been made using methods typical of current object-based [175] approaches. Both approaches to multichannel content creation tend to focus on sound sources as mono point-source signals. In contrast, stimuli in this study were produced using methods that aim to capture and reproduce highly realistic sound images and sound scenes by combining complex microphone arrays, and optimizing production techniques specifically for 3D playback environments.

6.3.1 Stimuli Recording and Production

Three excerpts were chosen from recent 22.2 recordings to be used as testing stimuli, each representing a different genre of acoustic music. These recordings and their associated production techniques have been evaluated through formal listening tests [156], [95], [176], and numerous informal evaluations at 22.2 studios in North America, Europe, and Japan. Leading 3D audio experts at McGill University, Tokyo University of the Arts, Rochester Institute of Technology, and the BBC have deemed these recordings “critical testing material.” To create stimuli for the other 3D reproduction formats under investigation, one option would be to apply a downmixing algorithm to the original 22.2 mixes, such as those described by Sugimoto [66] and Ando [177]. This methodology aims to avoid the introduction of the “human” variable: the aesthetic or technical bias present in any remixing engineer. However, automated downmixing may introduce unwanted spatial or timbral artefacts, especially if correlated or semi-correlated microphone signals are combined in the process [177]. For this study, the original 22.2 mixes were manually remixed for each reproduction format under test, a methodology similar to [10], [9], [55], and [174]. This

methodology has its own pros and cons: a skilled mixer should be able to create stimuli that retain a high degree of spatial and timbral fidelity for each reproduction format, but the question of how similar the mixes are between formats becomes highly subjective. The remixing methodology described in Section 6.3.2 aims to achieve a high degree of consistency between each format, using multiple mixing engineers to remove or average out some of the human bias.

6.3.1.1 Musical Excerpt 1: "Orchestra"

"Orchestra" is a 25 second excerpt from Holst's "Mars: The Bringer of War" from *The Planets*. The piece was performed by a 90-piece symphony orchestra in a large scoring stage, captured with a 22 microphone "main system". An extensive explanation of the recording methodology can be found in Chapter 4. This specific excerpt was chosen because of its dense orchestration and relatively stable dynamic envelope.

6.3.1.2 Musical Excerpt 2: "Bass"

"Bass" is a 20 second excerpt from a recording of a solo bass. The piece, a jazz/new-music free improvisation, was recorded in the same scoring stage as "orchestra", but features a much longer reverb time of around 5 seconds, due to the acoustical treatment in the studio having been removed prior to the recording. The recording methodology combined a multi-microphone direct sound capture system based on ideas described by Martin et al. in [94] and [95] with an array of widely spaced ambience microphones similar to [156] (Figure 49). The goal was to capture both the complex spectrum, and horizontal and vertical extent of the instrument, thereby creating a sonic image that is very similar to what one would hear standing in front of the bass. The performer's use of pizzicato playing leaves a great deal of space in the music, affording the listener a very strong impression of the room's late reflected energy, contributing to a strong sense of "envelopment" and "presence".

Figure 49: Solo bass in scoring stage, with direct sound and ambience arrays.

6.3.1.3 Musical Excerpt 3: "Jazz"

"Jazz" extends the sound capture concept of "bass" to a jazz trio (tenor sax, bass, drums), recorded in a medium sized concert hall (Figure 50). The hall measures 36m long by 18m wide by 12m high, with an average T60 of 1.8 s. As with the solo bass recording, the goal was to create a sound scene that realistically captures the size and spectrum of each instrument, accurate on-stage sound source positioning, and the performance space's acoustic signature. The 40 second excerpt used as stimulus maintains a stable dynamic envelope throughout.

Figure 50: Jazz trio in concert hall with direct sound microphone arrays

6.3.2 Stimuli Remixing

Having been recorded and mixed for 22.2, each musical excerpt was then remixed for each format under investigation: 11.1 (ATSC 3.0) [69], 10.2 (KBS/ITU) [5], and Auro 9.1 [7] (Figures 51 – 54). 11.1 and 10.2 were chosen as they are both currently standardized broadcast formats. Auro 9.1 is a popular format for 3D film mixing and music recording. All mixes were created by a professional recording engineer with over three years' experience recording and mixing music for three-dimensional reproduction, and more than 10 years' experience recording and mixing multichannel audio for music, film, and live performance. The quality and similarity of these mixes was then confirmed by two other professional recording engineers; adjustments to mixes were made as necessary until all parties were satisfied. To simplify mixing and testing conditions, the LFE channels were not used.

Remixing was performed using Merging Technologies' Pyramix 10 audio workstation, which allows for multiple mix sessions to be open simultaneously. As such, near-instantaneous comparisons between formats were possible. For each format, a mix was created that was as close to the original 22.2 reference as possible in terms of instrument or ensemble size, positioning on the sound stage, balance, timbre, clarity, depth of field, and quality and balance of early and late reflections. For the "orchestra" and "jazz" examples, there was an unavoidable narrowing of the sound stage for the formats other than 22.2, owing to a lack of loudspeakers at ±60°. Panning of phantom images of direct sound sources between front and side or front and rear speakers was avoided, as this led to sound images that were unstable and often suffered from some degree of comb filtering. This was not an issue for the "bass" example, as it contains only a single, centre-panned instrument. Another unavoidable difference between 22.2 and the other formats is the loss of the lower channels, which results in a slight vertical narrowing of the ensemble sound image. However, microphone signals originally panned to the bottom channels could be panned to main layer speakers, which aided in maintaining timbral similarity between mixes. Components of direct sound multi-microphone arrays that were deemed redundant or timbrally destructive when remixing to reduced channel formats were muted. For 11.1, 10.2 and 9.1, it was possible to keep the balance, EQ, and panning of signals within the L, C, and R channels identical for all remixes.

The primary function of the height and surrounding channels in these recordings was reproduction of ambience. When possible, microphone signals containing primarily ambient information were kept at the same spatial position, and same level relative to frontal direct sound sources. This contributed to a consistency of timbre and quality of reverberance between mixes. For example, 22.2, 11.1 and 10.2 all share main-layer loudspeaker positions at ±90° and ±135°. Thus, between those formats, ambience signal to channel allocation and

balance remained similar for the main reproduction layer. In situations where playback channels are reduced between 22.2 and other formats, each ambience track was auditioned for each remaining playback channel, to determine which combination yielded the spatial impression most similar to the 22.2 reference. In general, panning ambience microphone signals between loudspeakers was found to have a detrimental effect on the timbre and balance of early and late reflections, and was avoided. Balance, panning, and minimal equalization changes were the only adjustments needed during the remixing process.

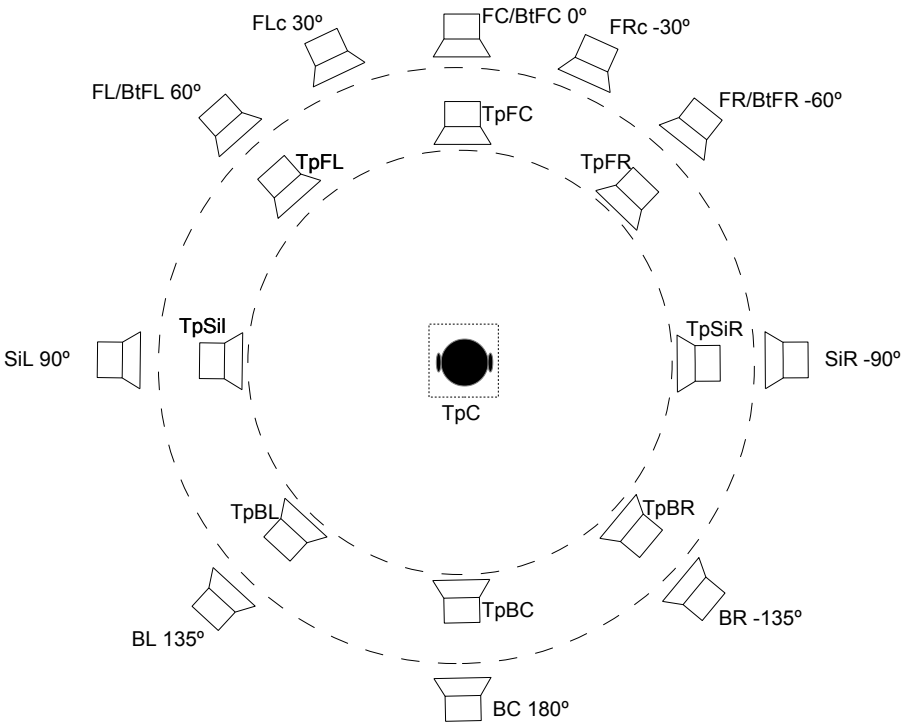


Figure 51: 22.2 speaker layout, viewed from above.

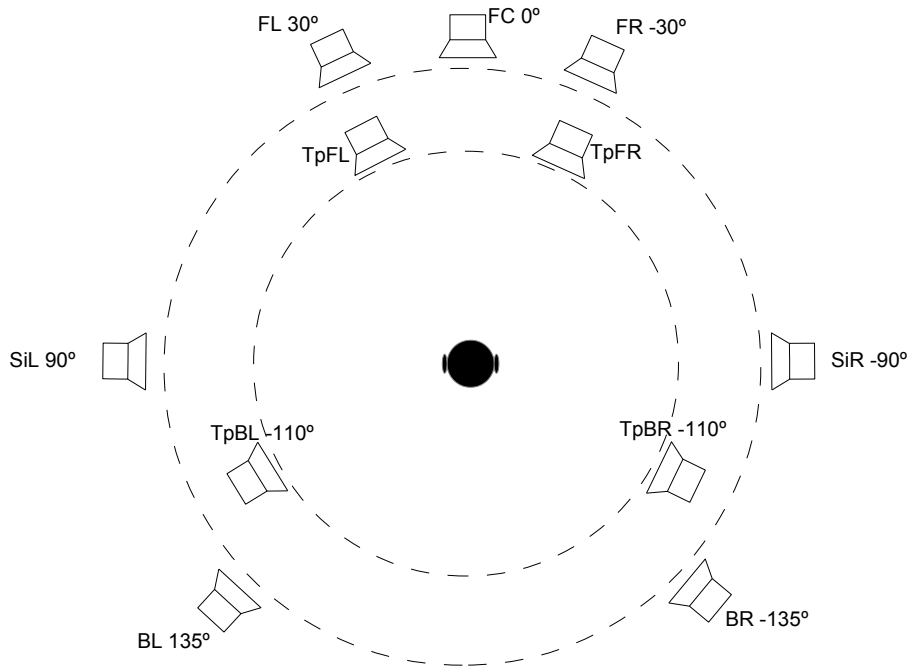


Figure 52: 11.1 speaker layout, viewed from above

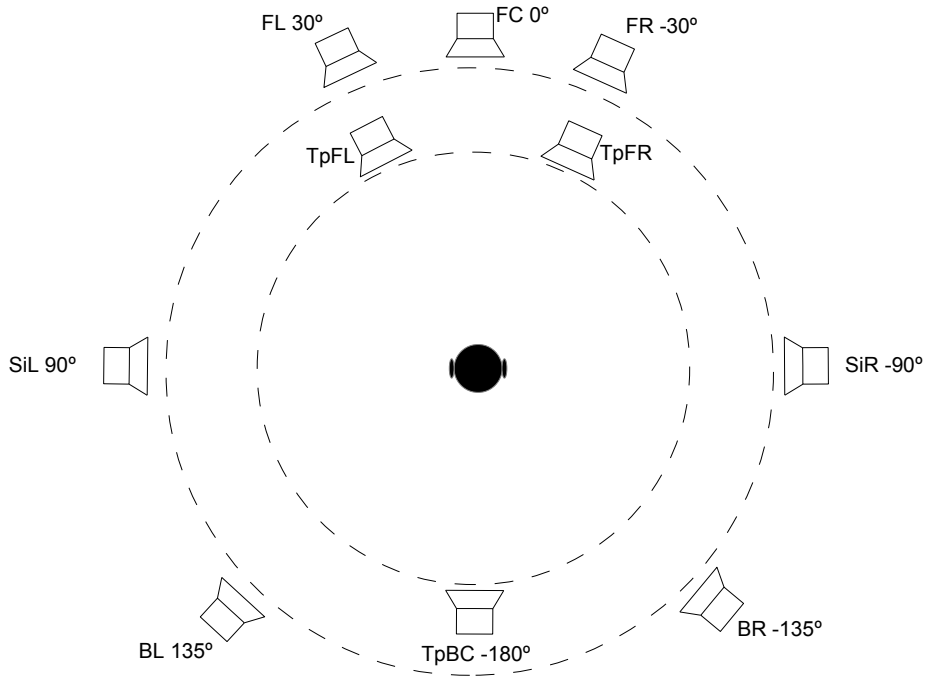


Figure 53: 10.2 speaker layout, viewed from above

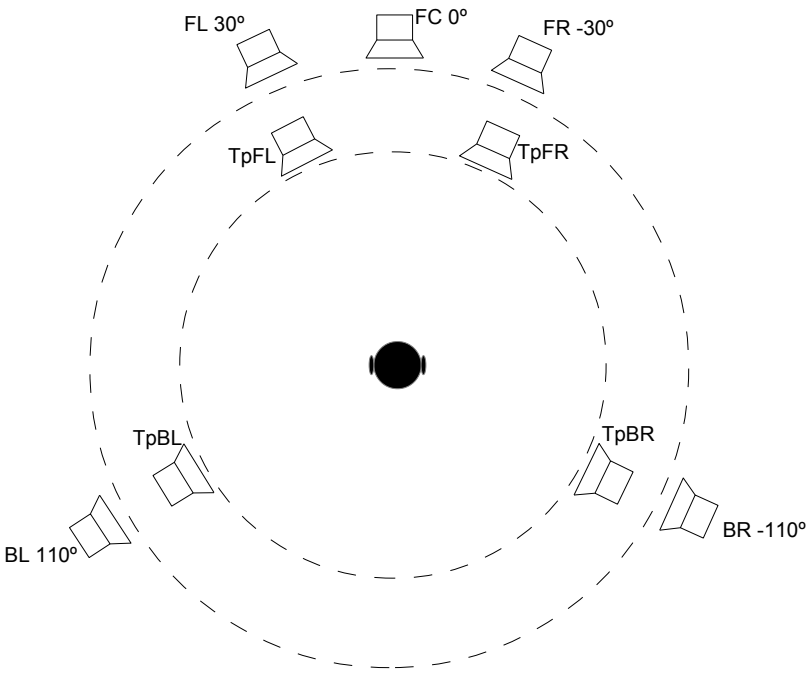


Figure 54: 9.1 speaker layout, viewed from above

6.3.3 Level Matching

For each musical excerpt, all stimuli were loudness matched by ear. These results were then confirmed by objective means. A Neumann KU-100 Dummy Head microphone was placed in the listening position at ear level, and used to record the playback of each stimulus. Recordings were made to Pro Tools 12 using an RME Fireface UFX. Integrated loudness measures were taken for each recording using the HOFA 4U software loudness meter (EBU +9 scale [178]). All stimuli, for each musical excerpt, were within 0.3 dB of each other.

6.4 Listening Environment

All stimulus mixing and listening tests took place in McGill University's Studio 22. The studio is equipped with 28 full-range, two-way loudspeakers (ME Geithain *M-25*) powered by Flying Mole class D amplifiers, and an Eclipse TD725SWMK2 stereo sub-woofer. The loudspeakers are arranged for reproduction of both 22.2 Multichannel Sound and Auro 3D

9.1. From these speaker positions all reproduction systems under investigation (Figures 47–50) could be achieved. The room's dimension ratios and reverb time fulfil ITU-R BS.1116 requirements [105], [140]. Continuous background noise does not exceed NR15 [140]. Studio 22's combined room and reproduction response shows a level deviation no greater than ±3 dB for the range of 20 Hz to 18 kHz. All height channels are positioned at an angle of elevation of 35º. The bottom channels for 22.2 share the same azimuths as the FL, FC, and FR loudspeakers, at an angle of elevation of –20º. The listening position is set at a point equidistant to all loudspeakers in the main layer.

6.5 Listening Test

A double-blind listening test was conducted to determine whether subjects could discriminate between 22.2 and other common 3D audio formats. A secondary aim of the test was to determine whether successful discrimination is possible between 11.1, 10.2, and 9.1, which are all similar in terms of number of playback channels and layout. Pairwise comparisons were made with a simple triad test, implemented using Cycling 74's Max/MSP. Subjects were seated in Studio 22's listening position and informed they would be comparing musical excerpts mixed for four different 3D audio formats. For each trial, one of the three musical excerpts was played on a continuous loop. Subjects were instructed to switch between stimuli labelled "A", B" and "C" at their leisure, and determine which two were the same (Figure 55). Playback of the different stimuli was time aligned to ensure seamless switching. Stimulus assignment to letters "A", "B", and "C", as well as the order of musical excerpts and format pairings were randomized within the testing program. Each of the three musical excerpts was mixed for four different formats, yielding 12 different stimuli. This results in a total of 18 possible pair-wise comparisons. Each pairwise comparison was evaluated twice, for a total of 36 trials. The test took an average of 25 minutes to complete, which is in line with ITU recommendations for the similar "double-blind triple stimulus with hidden

reference" methodology [105]. There was an enforced rest period of 1 minute after the 18th trial. After test completion, subjects filled out a short demographic survey that included the optional question "Please comment on what aspects of the recorded sound you were able to use to discriminate between examples."

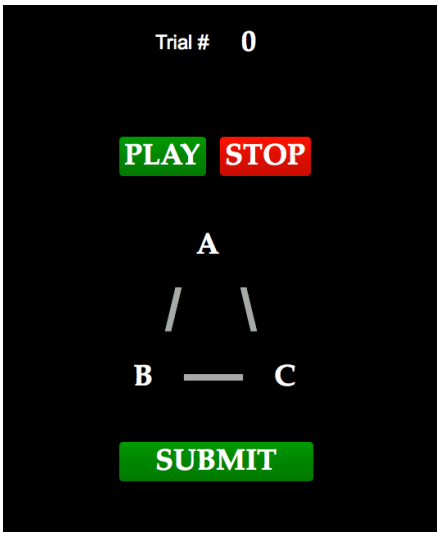


Figure 55: Testing GUI. Subjects indicated their selection by clicking on the line connecting the two mixes they believed to be the same.

6.5.1 Participants

Twenty subjects took part in the listening test. The subject pool was drawn from current students, recent graduates, and faculty of the Graduate Program in Sound Recording at McGill University. All reported having normal hearing, had previous experience performing triad or pairwise comparison-style listening tests and ear training activities, and had previous experience hearing three-dimensional music recordings. The decision to restrict the subject pool to more experienced listeners was based on a desire to obtain data with greater statistical power. Schoeffler and Herre [179] showed that experienced listeners tend to be more consistent in their ratings of spatial audio stimuli than naïve listeners. Olive [180] found that for evaluation of different loudspeaker types, experienced listeners were more discriminating then untrained listeners.

6.6 Results

6.6.1 Effects of Participant Demographics

A series of Chi-Squared tests showed there was no significant effect of age, years of musical experience, or years of production experience on subjects' discrimination ability, which appeared to be normally distributed with no outliers.

6.6.2 Playback Format Comparison

Figure 56 and the results of six binomial tests (Table 14) show that discrimination between all pairs of playback formats was significantly above chance (0.33). In all cases except for the 10.2 vs. 11.1 comparison, and particularly with all comparisons involving the 22.2 format, the discrimination rate is well above chance, reaching a maximum of 90% success for the 10.2 vs. 22.2 comparison.

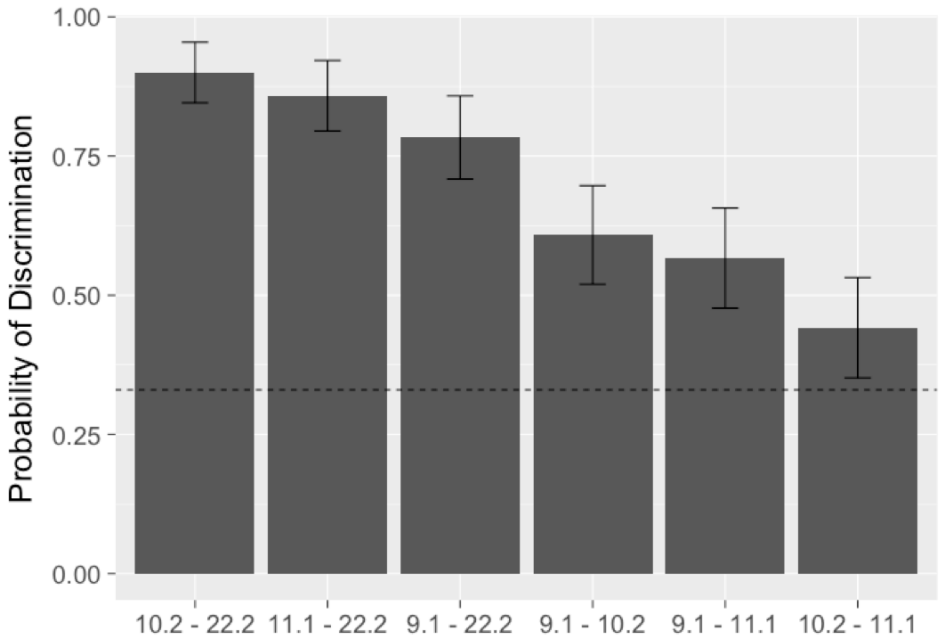


Figure 56: Probability of discrimination for each pair of playback formats. Dotted horizontal line indicates probability of chance (33%)

Table 14: Binomial test on format discrimination (chance probability = 0.33)

Formats	Discrimination	95% Confidence Interval	p
10.1-22.2	0.90	0.83-0.95	<0.001
11.1-22.2	0.86	0.78-0.92	<0.001
9.1-22.2	0.78	0.70-0.85	<0.001
9.1-10.1	0.61	0.52-0.70	<0.001
9.1-11.1	0.57	0.47-0.66	<0.001
10.1-11.1	0.44	0.35-0.54	0.015

A mixed-effects logistic regression model was built using the pairwise comparison to predict a correct response. The subject number was input as a random effect (intercept) to compensate for any overall differences in discrimination ability between participants. This model was a statistically better fit to the data when including the pairwise comparison as a predictor than when not including it $X^2(5) = 29.29$, $p < .001$.; there are statistically different discrimination rates between the different sets of comparisons. These differences were explored using multiple comparisons with Tukey contrasts (Table 15).

Table 15 shows there were no significant differences between pairs of comparisons when the 22.2 format was involved in both comparisons. When the 22.2 format was involved in only one of the pairs being compared, significant differences existed for all comparisons. When the 22.2 format was not involved in either comparison, there were no significant differences between pairs of comparisons. This essentially creates two groups of comparisons: those with the 22.2 format and those without. There are no significant differences within these two groups, but there are significant differences between the two groups. Ultimately, this suggests that the 9.1, 10.2, and 11.1 formats are all similar when compared to the 22.2 format.

Table 15: Mixed-effects Logistic Regression Model with Tukey Contrasts. *Pairwise format comparison* predicting *correct response*, *subject number* as a random effect.

Pair A	Pair B	*Z* Value	Adjusted *p*
10.1-11.1	10.1-22.2	6.847	**<.001**
10.1-11.1	11.1-22.2	6.371	**<.001**
10.1-11.1	9.1-10.1	2.581	.099
10.1-11.1	9.1-11.1	1.938	.373
10.1-11.1	9.1-22.2	5.286	**<.001**
10.1-22.2	11.1-22.2	-0.988	.920
10.1-22.2	9.1-10.1	-4.929	**<.001**
10.1-22.2	9.1-11.1	-5.433	**<.001**
10.1-22.2	9.1-22.2	-2.428	.143
11.1-22.2	9.1-10.1	-4.240	**<.001**
11.1-22.2	9.1-11.1	-4.799	**<.001**
11.1-22.2	9.1-22.2	-1.509	.653
9.1-10.1	9.1-11.1	-0.658	.986
9.1-10.1	9.1-22.2	2.922	**.039**
9.1-11.1	9.1-22.2	3.538	**.005**

6.6.3 Effect of Program Material on Discrimination

The probability of discrimination rates between the different musical excerpts were found to be similar (Jazz = 0.74, Orchestra = 0.68, Bass = 0.66). A mixed-effects logistic regression model was built using the musical excerpt to predict a correct response. The subject number was input as a random effect (intercept). This model was not a statistically better fit to the data when including the musical excerpt as predictor than when not including it $X^2(2) = 4.31$, $p = .116$; there were no differences in discrimination rates between the different music excerpts. A model was built to examine any possible interaction effects between the music excerpt and the pair of formats between compared, however, the model failed to converge and is therefore not reliable.

6.6.4 Perceptual attributes collected from subjects

Immediately after completing the test, subjects were asked to complete a short demographic survey. Of the 20 participants, 16 gave answers to the optional question "Please comment on what aspects of the recorded sound you were able to use to discriminate between examples." These comments were then searched by the primary author for terms or synonyms of terms common to subjective spatial audio evaluation. For example, "*I tried to focus on small changes in spatial impression - positioning and density of reverb*", would be simplified as "spatial impression". Terms with similar or identical meanings, such as "ambience" and "reverb" were pooled together, referencing lists of attributes from previous work [109], [107], [118], and a thesaurus. The three most common attributes reported were "timbre" (70% of participants), "spatial position of direct sounds" (81%), and "spatial impression" (94%).

6.7 Discussion and Conclusions

6.7.1 Listener Discrimination

Figure 56 and Table 14 show subjects could discriminate between 22.2 and other common 3D audio reproduction formats with a very high degree of success. This result was the same for all subjects, regardless of age group, musical training, or music production experience. This suggest that within the context of three-dimensional reproduction of acoustic music, there are clear perceptual differences between 22.2 and 11.1, 10.2, and 9.1. These differences appear to be equally appreciable across multiple recordings using different recording and mixing techniques. Discrimination was also possible between 9.1, 10.2, and 11.1, formats that are similar in terms of number of channels and position of speakers. It is not surprising that the lowest mean discrimination rate was found for the 10.2 vs. 11.1 comparison, as these formats are almost identical (Figures 52 and 53). Perhaps most germane to the goals of this study are the results shown in Table 15 and summarized in Section 6.6.2. The key idea here is

that significant differences between pairs of format comparisons only exist when the 22.2 format is involved in one of the comparisons. This suggests a clear and marked difference between the 22.2 format and all other formats under investigation, which are all perceptually more similar to each other for reproduction of acoustic music. This also suggests that the experience of hearing high-quality acoustic music recordings created for 22.2 cannot be duplicated using reproduction formats with a reduced channel count. These findings support results from previous research by Hamasaki et al. [9] and Shim et al. [55], while addressing Kim et al.'s conclusion that further investigation of the perceptual differences between various channel-based 3D audio formats was necessary [68].

6.7.2 Perceptual differences between formats

As seen in Section 6.6.4, "spatial impression", "spatial position of direct sounds", and "timbre" were the most frequently used subjective attributes by subjects when describing what aspects of the recorded sound were useful for discriminating between reproduction formats. References to sound source positioning are likely primarily related to 22.2 versus other format comparisons, as the position of direct sound sources remained identical for the 11.1, 10.2, and 9.1 mixes. The switch from a frontal sound image that spans 120º to 60º would be obvious to most listeners, particularly for the "jazz" and "orchestra" excerpts that involved direct sound from non-central locations.

Comments related to "spatial impression" were dominated by subjective attributes related to what Griesinger would define as "background spatial impression" [43], and were primarily references to late reflected sound energy, such as "spatial distribution of ambience", "reverberance", "envelopment". Results from [155] appear to show a clear difference between 22.2 and 11.1 for the subjective attribute "engulfment", a term related to "spatial impression". Results reported in [2] and [9] show strong perceptual differences between 22.2 and 5.1 for the attributes "envelopment", "reverberant", "depth" and "presence", all of which

are related to early and late reflected sound energy. [55] shows a consistent significant difference between 22.2 and 10.2 for "listener envelopment", regardless of the type of listening material or subject seating position (centre or off-centre). This, combined with the current study's subject comments, suggests that aspects of spatial impression and particularly late reflected sound energy represent the most salient differences between channel-based 3D audio reproduction systems, with respect to reproduction of acoustic music. Many three-dimensional recordings of acoustic music retain a "concert" perspective, wherein the surround and height channels are reproducing primarily ambient sound information. Therefore, it is likely the spatial impression that is most affected by changes to the number of and spatial positions of said channels.

6.7.3 Future Work

Olive describes the advantages of using experienced listeners for this type study: "Training and experience in controlled tests lead to significant gains in performance so that fewer listeners are required to achieve the same statistical power. [180]" The relative ease with which experienced listeners were able to discriminate between 22.2 and the other formats under investigation, regardless of years of music production experience, suggests that this result could be generalized to a less discriminating, more general population. To confirm this, however, it is necessary to run the same experiment again with non-experienced listeners.

Like [10], [9], [55], and [174], for this study, musical stimuli were derived from mixes made specifically for each format under investigation. It may prove valuable to repeat the experiment using downmixed material as stimuli, to see if results remain similar.

Zacharov et al. highlight the need for critical testing material for experiments involving 3D audio formats [155]. The recordings and mixes created for this study will help meet this need. Future recordings are planned to increase the musical range of this stimulus

set to include pop/rock material. For access to the stimuli or max patch from this study, for research or evaluation purposes, please contact Will Howie (wghowie@gmail.com).

7 CONCLUSIONS

The importance of 3D audio within both the research community and the commercial marketplace has been rapidly increasing over the last several years. Though the rate of technological change is staggering, with 3D audio becoming more accessible to consumers every year, more work needs to be done to ensure that audio capture methods remain up to date. Rumsey wrote in 2002:

"High technical quality or fidelity, it can be argued, may be taken for granted at this point in the history of audio engineering. Although not all audio devices exhibit the highest technical quality, the technical quality of the best sound reproduction available to the consumer exhibits very low levels of distortion, a wide frequency range, a flat frequency response, and low noise, with specifications that match or exceed the limits of human perception. Although improvements may still be made in these domains, the technical quality curve is becoming asymptotic to the ideal, and product development is in a region of diminishing returns. Spatial quality and character, on the other hand, have some way to go before the curve could be said to be asymptotic to some ideal. [109]"

Although Rumsey was referring primarily to two-dimensional multichannel audio, the sentiment remains true today. There is a danger that the technical quality of 3D audio

playback environments may soon or already has surpassed the technical quality of sound recordings and mixes created for said systems. As with the early days of stereo or 5.1 content production, we must now explore and define the technical and aesthetic considerations for music recording for three-dimensional audio environments. As such, there is a strong need for research that addresses music capture methods for 3D audio systems of all kinds, but especially those systems that exhibit the greatest potential for immersive listening experiences. Hamasaki et al. showed that 22.2 produces a superior impression of presence and realism when compared to other two and three-dimensional playback formats [9]. The authors concluded that "additional subjective evaluation experiments are required to assess the further capability of the 22.2 multichannel audio system [9]." As seen in Section 2.4, several techniques and concepts have been introduced for 3D music recording, but few of these have been subjected to stringent evaluations. Such experiments are vital to understand if capture techniques are fully exploiting the capabilities of the audio systems they have been designed for. There exists already a number of well-known high-quality "reference" recordings for stereo and 5.1 surround sound that can be used as stimuli for perceptual tests. The creation of high-quality three-dimensional audio recordings that can be deemed "critical testing material," and which stress the limits of the playback environment they are intended for, is important not only for the advancement of the art of audio practitioners, but also for experimental research examining perceptual aspects of three-dimensional sound reproduction. This thesis has attempted to address these gaps in research.

7.1 General Conclusions

Chapter 3.1 describes the design and implementation of a fourteen-channel microphone array for three-dimensional acoustic music capture. The array is based on previous research in 3D recording and concert hall acoustics, particularly studies focused on listener envelopment. The recording technique was evaluated through informal listening sessions

where several important observations were made: 1) The addition of height channel information to the reproduction of acoustic music greatly increases the impressions of envelopment and presence, as reported in previous research. 2) Late reflected energy in side channels (±90º) appears to be particularly important for achieving strong levels of envelopment, which is in line with previous research in concert hall acoustics. 3) Omnidirectional microphones for height channels are prone to capturing too much direct sound information, which can destabilize the frontal sound image. 4) An above-the-head centre channel contributes to a greater homogeneity within the ambient sound field.

Chapter 3.2 explores the relationship between height information capture and microphone polar patterns through a series of experimental recordings of different ensembles in different acoustic spaces. A special two-channel microphone array was designed to capture all possible height channel polar patterns simultaneously. Results of a double-blind listening test showed that listeners showed no significant preference between three different polar patterns for left and right (±90º) height channels: omnidirectional, cardioid, and bi-directional. This suggests that within the context of a commercially balanced 3D music mix, differences to the overall sound scene created by changing the polar pattern of height channel microphones may be far subtler or far less meaningful to most listeners than originally thought, particularly for listeners with little previous experience hearing 3D audio content. Practitioners should consider themselves free to capture the kind of height information that best compliments their desired sound scene aesthetic, and not be bound by specific microphone types, as was often the case in stereo or 5.1 recording techniques.

Chapter 4 details the design, implementation, and preliminary evaluation of a three-dimensional orchestral music capture technique, optimized for 22.2 Multichannel Sound. The technique is based on the experimental recordings and listening results from Chapter 3, as well as previous research in one, two, and three-dimensional music recording, and spatial

impression in multichannel sound reproduction. The technique is designed to prioritize the capture of a natural orchestral sound image with realistic horizontal and vertical extent, stable direct sound source localization, and a highly diffuse reflected sound field. Through a series of informal listening sessions at five different 22.2 reproduction facilities, it was observed that the technique achieved its sound capture goals, and that the sound scene impression remained constant regardless of the type of room, loudspeakers, or loudspeaker radius and azimuths. A subsequent listening test showed that within the context of dynamic orchestral music, subjects could successfully differentiate between playback conditions with and without the bottom channels. For the conditions with bottom layer active, many subjects observed a vertical extension of the orchestral sound image, as well as an increase in low frequency content, which validates components of the technique's design rationale. Results also highlight the need to use music excerpts with stable dynamic envelopes for "ABX" or "triad" style comparative listening tests, as had been suggested in previous research.

Chapter 5 describes a subjective listening test designed to compare three different orchestral music capture techniques optimized for 22.2, and to confirm that the technique proposed in Chapter 4 is valid for broadcast and commercial recording. Subjects rated the proposed technique as well or better than a current production standard used by Japan Broadcasting Corp., as well as a spherical higher order ambisonic capture system, for the subjective attributes "clarity", "scene depth", "environmental envelopment", "sound source envelopment", "naturalness", and "quality of orchestral image". The proposed technique was also significantly preferred over the other two techniques. All subjective attributes used in the listening test were found to be correlated with each other, with no one attribute being a significant predictor of overall listener preference. While the two spaced recording techniques were rated highly across all attributes, the Eigenmike was rated very poorly for all attributes, particularly "naturalness". This, when combined with findings from previous research

comparing 2D recording techniques, suggests that ambisonics-based recording techniques are still not viable for use in commercial multichannel music recording. The experiment was performed again at Tokyo University of the Arts, with similar results.

Chapter 6 addresses the question of whether the perceptual listening experience delivered by high quality 22.2 acoustic music recordings can be achieved using 3D reproduction formats with a reduced channel count. Results of a double-blind listening test show that subjects can discriminate between 22.2 and three other common 3D audio formats with a high degree of accuracy. This result is the same across multiple recordings of acoustic music (including orchestral music), created using different recording and mixing techniques. Results also show that significant differences between pairs of format comparisons only exist when the 22.2 format is involved in one of the comparisons. This suggests that clear perceptual differences exist between 22.2 and other common 3D audio formats for the reproduction of acoustic music. It also suggests that when compared to 22.2, the other 3D formats under investigation are all perceptually similar. This strengthens one of the central arguments of this thesis: that 22.2 has the potential to reproduce orchestral music in a way that is perceptually unique among currently standardized 3D audio formats.

7.2 Further Discussion

7.2.1 Adaptation of Recording Techniques

Although the novel orchestral music recording technique described in this thesis was optimized for 22.2 Multichannel Sound, the design concepts behind it are not limited to that format, or orchestral music capture. Figure 11 shows how the technique can easily be adapted to other reproduction formats by simply subtracting redundant microphone channels. The basic approach of using omnidirectional microphones with acoustic pressure equalizers for primarily direct sound capture, and widely spaced directional microphones for decorrelated

ambience capture can be applied to any instrument or ensemble. This approach has already been used in subsequent test recordings, such as the solo piano example shown in Figure 11, with consistently good results. Individual components of the recording technique can also be applied outside of 3D audio. While recording the stimuli for the study described in Chapter 5, a separate production team captured the orchestral rehearsals and concerts for national broadcast on CBC Radio 2. This team had a split of Technique 1's "Decca Tree" (see: Section 5.3), as well as several ambience microphones. It was observed that Technique 1's front height channels combined very well with the Decca Tree to give a complete stereo image. Chapter 4 shows how 3D recording techniques designed with the concert hall in mind can be implemented in a way that still achieves a rich depth of field in spaces where physical depth is difficult to achieve through microphone placement, such as a recording studio or scoring stage. The extensive documentation in Chapters 3-6 describing how and why each recording technique was designed and setup can serve as a guide to recording engineers and researchers working within the new paradigm of 3D audio. Practitioners should feel free to mix and match ideas from these techniques to achieve a spatial sound scene that best matches their desired aesthetic outcome in a given situation.

7.2.2 Realistic Sound Reproduction: Approaching an Infinite Transducer

In some ways, the research in this thesis can be seen as a revisitation of the work of Snow, Fletcher, and Steinberg at Bell Labs in the 1930s. Working with a "screen" analogy, these researchers conceived of a system designed to capture orchestral sound in a concert hall that would then be reproduced simultaneously in another hall, perhaps many thousands of kilometres away [181]. In its truest form, this stereophonic sound reproduction system would consist of an infinite number of tiny microphones hung in a screen, placed in front of the orchestra to capture direct sound. These microphones would be connected to a corresponding number of tiny loudspeakers hung in a similar fashion at the reproduction venue (Figure 57).

"Then the sound projected at the audience will be a faithful copy of the original sound and an observer will hear the sound in true auditory perspective. [181]" A simplified version of the technique using three microphones and three loudspeakers was tested in 1933, and was found to produce good results [182], [183].

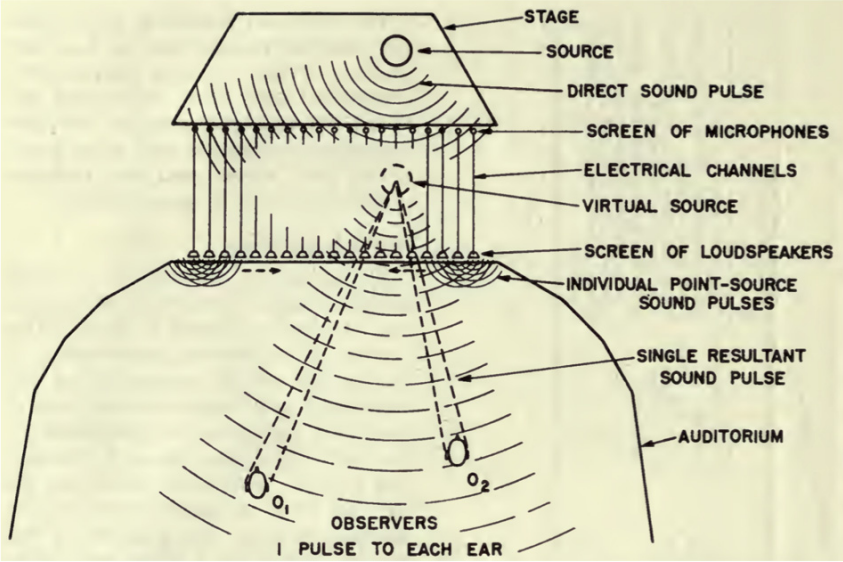


Figure 57: Snow's "Ideal stereophonic system", reproduced with permission from [181]

One of the primary goals in the development of the recording techniques discussed in Chapters 3–6 was achieving a strong sense of realism within acoustic music recordings. This was accomplished by designing microphone capture systems that exploit the number and location of speaker positions within the 22.2 audio format to create instrument or ensemble images with realistic localization, relative physical size, timbre, and tone colour, as well as a faithful reproduction of the recording venue's acoustic signature. 22.2's large number of loudspeakers for frontal sound field reproduction, combined with the inclusion of a bottom layer of loudspeakers, offers the ability to reproduce strong, focused sound images that have realistic horizontal *and* vertical extent, as seen in Chapters 4 and 6, [2], [9], [95], and [153].

Results from Chapter 6, [2], [9], and [55] show that 22.2 is perceptually unique or superior among common 2D and 3D channel-based audio formats for subjective attributes

such as “spatial impression,” “envelopment,” and “presence”. Oode et al. [184] investigated different loudspeaker configurations to determine how the number of surround and height channels in an audio reproduction system affects the sensation of “listener envelopment” (LEV). Oode et al. found that for configurations with loudspeakers only at ear-level, the sensation of LEV became saturated with 12 loudspeakers, and did not increase as more loudspeakers were added. However, for loudspeaker configurations that included a height layer, the sensation of LEV continued to increase as the number of loudspeakers increased [184] (Figure 58). This all indicates a likely correlation between number and position of loudspeakers in a 3D audio reproduction system, and the ability of said system to deliver a realistic impression of a given sound scene. In that respect, among current channel-based 3D audio systems, 22.2 may be the best suited to this task. This begs the question: is there is an optimal or ideal number and arrangement of points of sound reproduction before we reach a condition of diminishing returns or perceptual saturation? Or, taking a cue from Snow, Fletcher and Steinberg, are we simply moving along a path that will not end until the realization of an infinite transducer?

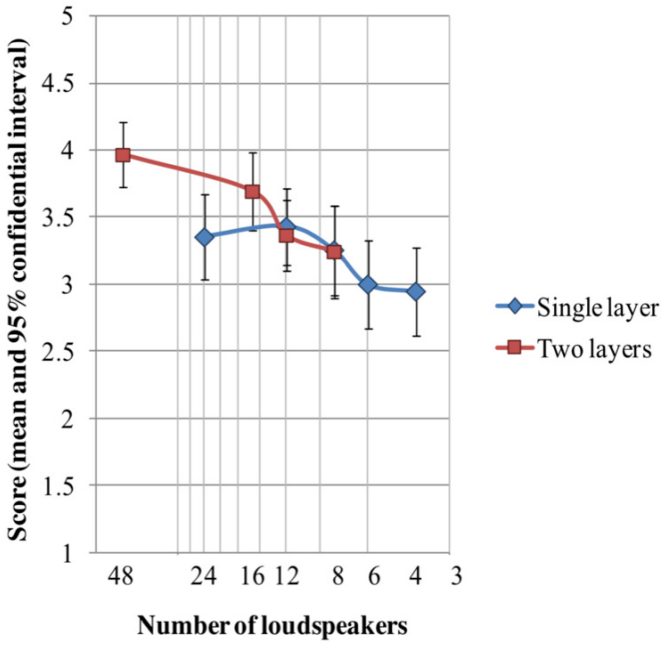


Figure 58: Results for sensation of LEV, reproduced with permission from [184]

7.2.3 Considerations for ITU-R BS.2159-7

Currently in its seventh addition, ITU-R BS.2159-7 [77] summarized numerous standards and recommendations from the ITU regarding three-dimensional audio (advanced sound systems), as well as important related research and standards. There are several sections of this document that, in light of the research presented in this thesis, could now be updated. References to various sections of BS.2157-9 will be printed in *italics* to avoid confusion with similarly named sections of this manuscript.

BS.2157-7 Section 4.1 contains a list of facilities capable of 22.2 sound reproduction that is now out of date, missing facilities at Tokyo University of the Arts, Rochester Institute of Technology, University of Huddersfield, TC Electronics, and others.

Section 6.2.1 "Principles of three-dimensional sound mixing" appears to be based on a limited number of English-language publications from the NHK on this topic, and presents a somewhat narrow perspective on recording and mixing for 22.2. The "conventional applications" of each speaker layer are defined as:

"**Top Layer**: Reverberation and ambience. Sound localized above, such as loudspeakers hung in gymnasiums […] Unusual sound, such as meaningless sound.

Middle Layer: Basic sound field formation. Envelopment reproduction.

Bottom Layer: Sound of water such as the sea, rivers, and drops of water. Sounds of the ground in scenes with a bird's-eye view [77]"

Chapters 4 and 6 of this thesis, as well as [95] and [153] show the importance of both the top and especially bottom loudspeaker layers in a 22.2 system for creating musical instrument and ensemble sonic images that have realistic, well-defined horizontal and vertical extent, and remain spatially anchored over a wide listening area. These new recording and

mixing concepts suggest a need to revise and update the above list of "conventional applications". Additionally, the technical descriptions of three-dimensional music recording found in Chapters 4–6 could be summarized and added to *Section 6.2.3*, which presents examples of audio production for 22.2 that are now somewhat out of date.

Section 7.4 describes a number of recent studies investigating the performance quality of various multichannel sound systems. A summary of the study detailed in Chapter 6 would be a valuable addition to *Section 7.4*, as it is the only currently published research that specifically addresses listener discrimination between various 3D audio formats. Results from Chapter 6 that show strong perceptual differences between 22.2 and other common 3D audio formats for the reproduction of acoustic music would be valuable in updating *Section 7.2*, which discusses a study whose results are now somewhat questionable, based on the quality of the musical stimulus used for its listening test.

7.3 Future Work

The most important contribution this thesis makes to future research is likely the creation of high quality stimuli for multiple 3D audio formats, as detailed in Chapter 6. Chapter 4's study on listener perception of lower channels in a 3D audio environment remains the only one of its kind. Having now created a wider variety of high quality material optimized for 22.2, more testing can and should be done to investigate the perceptual effects and importance of bottom, ground-level channels in 3D audio reproduction environments, especially for "vertical imaging", as discussed in Chapter 4 and by Martin et al. in [94], [95], and [153].

The study described in Chapter 5 has been repeated at Tokyo University of the Arts (Section 5.7) as well as Rochester Institute of Technology, with the aim of validating previous results and investigating possible inter-cultural differenced in perception of three-dimensional audio. Although significant cultural differences were not observed, analysis of

the data did reveal some significant trends with regard to the impact of musical and audio training on subject consistency. A new set of listening tests has been performed at McGill to further investigate listener consistency within the context of three-dimensional audio reproduction, an area of research that has seen little published scholarship as of yet. This work follows logically from previous work by Olive [180] and Bech [185] exploring subject performance in stereo sound reproduction listening tests.

The results of Chapter 6 should be further validated in several ways. The same listening test should be performed again using naïve listeners, to confirm the high success rate of discrimination between 22.2 and other 3D audio formats observed is not limited to "expert" or "trained" listeners. This would help generalize the results to a larger population. Chapter 6 also raises the question of downmixing vs. remixing methodologies for stimulus creation. It would be relatively simple and highly valuable to perform the same listening test again, but using downmixed content for the 11.1, 10.2 and 9.1 playback conditions, again, in hopes of making the results of the original study more universally applicable. The stimulus set created for Chapter 6 consists entirely of "acoustic music" material: classical, new music, and jazz. Work has already begun on a large-scale three-dimensional pop/rock recording, based on the multi-microphone direct sound capture techniques mentioned in Chapter 6. This recording will help expand the musical breadth of the 3D audio stimulus set. There are also plans for a 22.2 recording of new electroacoustic compositions that would feature a more "inside the ensemble" 360º perspective, which should prove valuable for covering a wider range of reproduction aesthetics within future perceptual tests. As mentioned in Chapter 6, this material will be made available to researchers at academic institutions outside of McGill.

For the sake of brevity and publication space limitations, Section 6.3.1 provides only a summary of how each of the three 22.2 stimuli recordings were made. The techniques used to create the "bass" and "jazz" recordings are very new, having emerged from recent experimental recording sessions. Upon completion of the 22.2-optimized "pop" recording mentioned above, a paper is planned to introduce the rationale behind and implementation of these direct sound capture techniques, which are designed to capture realistic sonic image size and instrument spectrum using an advanced multi-microphone placement scheme. Although designed for channel-based 3D audio formats, this approach would also work very well in an object-based workflow: each instrument could be represented by a "multitrack" or "multichannel" object – a relatively new concept in the field. It would also be valuable to objectively measure whether these multi-microphone techniques are in fact capturing a greater, more representative amount of an instrument's spectrum then traditional mono and stereo microphone techniques.

8 BIBLIOGRAPHY

[1] F. Rumsey, Spatial Audio, Burlington: Focal Press, 2013.

[2] K. Hamasaki et al., "Effectiveness for height information for reproducing presence and reality in multichannel audio system," in *AES Convetion 120*, Paris, 2006.

[3] T. Kamekawa et al., "Evaluation of spatial impression comparing 2ch stereo, 5ch surround, and 7ch surround with height for 3D imagery," in *AES Convention 130*, London, 2011.

[4] S. Kim et al., "Subjective Evaluation of Multichannel Sound with Surround-Height Channels," in *AES Convention 135*, New York, 2013.

[5] "Advanced sound system for programme production," *ITU-R BS.2051-0,* 2014.

[6] "Dolby Atmos for the Home Theatre," Dolby Laboratories, Inc., San Francisco, USA, 2015.

[7] B. Van Daele and W. Van Baelen, "Productions in Auro 3D: Professional workflow and costs," Auro Technologies, 2012.

[8] K. Hamasaki and K. Hiyama, "Development of a 22.2 Multichannel Sound System," *Broadcast Technology,* vol. 25, pp. 9-13, Winter 2006.

[9] K. Hamasaki et al., "Advanced multichannel audio systems with Superior Impressions of Presence and Reality," in *AES Convention 116*, Berlin, 2004.

[10] K. Hamasaki and W. Van Baelen, "Natural Sound Recording of an Orchestra with Three-Dimensional Sound," in *AES Convention 138*, Warsaw, 2015.

[11] K. Hamasaki, "The 22.2 Multichannel Sounds And Its Reproduction At Home And Personal Environment," in *AES 43rd International Conference*, Pohang, 2011.

[12] J. Blauert, Spatial Hearing: The Psychoacoustics of Human Sound Localization, Cambridge: MIT Press, 1997.

[13] Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera and H. Kato, Principals and Applications of Spatial Hearing, Zao: World Scientific Publishing Co. Pte. Ltd., 2009.

[14] P. Damaske and B. Wagener, "Subjective investigations of sound fields," *Acustica,* vol. 19, no. 4, pp. 198-213, 1967.

[15] S. Roffler and R. Butler, "Factors That Influence the Localization of Sound in the Vertical Plane," *J. Acoust. Soc. Am.,* vol. 43, no. 6, pp. 1255-1259, 1968.

[16] R. Butler and R. Humanski, "Localization of Sound in the vertical plane with and without high frequency spectral cues," *Percept. Psychophys.,* vol. 51, no. 2, pp. 182-

186, 1992.

[17] V. R. Algazi, C. Avendano and R. O. Duda, “Elevation localization and head-related transfer function analysis at low frequencies,” *J. Acoust. Soc. Am.,* vol. 109, no. 3, pp. 1110-1122, 2001.

[18] H. Lee, “Sound Source and Loudspeaker Base Angle Dependency of Phantom Image Elevation Effect,” *J. Audio. Eng. Soc.,* vol. 65, no. 9, pp. 733-748, 2017.

[19] C. Pratt, “The Spatial Character of High and Low Tones,” *Journal of Experimental Psychology,* vol. 13, no. 3, pp. 278-285, 1930.

[20] S. Roffler and R. Butler, “Localization of tonal stimuli in the vertical plane,” *J. Acoust. Soc. Am.,* vol. 43, pp. 1260-1266, 1968.

[21] D. Cabrera and S. Tilley, “Vertical localization and image size effects in loudspeaker reproduction,” in *AES 24th International Conference*, Banff, 2003.

[22] H. Lee, “Perceptual Band Allocation (PBA) for the Rendering of Vertical Image Spread with a Vertical 2D Loudspeaker Array,” *J. Audio Eng. Soc.,* vol. 64, no. 12, pp. 1003-1013, 2016.

[23] H. Lee, “Phantom Image Elevation Explained,” in *AES Convention 141*, Los Angeles, 2016.

[24] J. Hebrank and D. Wright, “Spectral cues used in the localization of sound sources on the median plane,” *J. Acoustic. Soc. Am. ,* vol. 56, no. 6, pp. 1829-1834, 1974.

[25] R. Wallis and H. Lee, “Directional Bands Revisited,” in *AES Convention 138*, Warsaw,

2015.

[26] C. Hugonnet and P. Walder, Stereophonic Sound Recording: Theory and Practice, New York: Wiley, 1998.

[27] G. Martin et al., "Sound Source Localization in a Five-Channel Surround Sound System," in *AES Convention 107*, New York, 1999.

[28] "Multichannel stereophonic sound system with and without accompanying picture," *Recommendation ITU-R BS.775-3,* August 2012.

[29] J. Corey and W. Woszczyk, "Localization of lateral phantom images in a 5-channel system with and without simulated early reflections," in *AES Convetion 113*, Los Angeles, 2002.

[30] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality," *J. Acoust. Soc. Am. ,* vol. 74, no. 6, pp. 1726-1733, 1983.

[31] M. Dickreiter, Tonmeister Technologies, New York: Temner Enterprises Inc., 1989.

[32] H. Irimajiri et al., "The Highly Preferred Sound Levels of Ambience Microphone Arrays to Front Microphone Arrays with the Fixed Levels for Surround Recording," AES Surround Study Group, 2006-2007.

[33] M. Williams, The Stereophonic Zoom, Gloucestershire: Rycote Microphone Windshields Ltd and Human Computer Interface, 2002.

[34] M. Gray, "The Decca Sound: Secrets Of The Engineers," Polymath Perspective, 13

June 2012. [Online]. Available: http://www.polymathperspective.com/?p=2484. [Accessed 6 October 2017].

[35] "M 50 – The Historic Omni Directional," [Online]. Available: https://www.neumann.com/?lang=en&id=hist_microphones&cid=m50_publications. [Accessed 6 October 2017].

[36] T. Kamekawa et al., "Corresponding Relationships between Physical Factors and Psychological Impressions for Microphone Arrays for Orchestral Recording," in *AES Convention 123*, New York, 2007.

[37] H. Wittek and G. Theile, "The recording angle – based on localisation curves," in *AES Convention 112*, Munich, 2002.

[38] H. Wittek, "APP "Image Assistant 3" Beta," 28 August 2015. [Online]. Available: https://www.hauptmikrofon.de/stereo-3d/image-assistant/ima-3-app. [Accessed 14 February 2018].

[39] H. Lee et al., "An Interactive and Intelligent Tool for Microphone Array Design," in *AES Convention 143*, New York, 2017.

[40] H. Lee, "Perceptually Motivated Amplitude Panning (PMAP) for Accurate Phantom Image Localisation," in *AES Convention 142*, Berlin, 2017.

[41] H. Lee and F. Rumsey, "Level and Time Panning of Phantom Images for Musical Sources," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 978-988, 2013.

[42] M. Williams, "Unified theory of microphone systems for stereophonic sound

recording," in *AES Convetion 82*, London, 1987.

[43] D. Griesinger, "Spatial Impression and Envelopment in Small Rooms," in *AES Convention 103*, New York, 1997.

[44] T. Kamawaka, "An Explanation of Various Surround Microphone Techniques," Sanken, [Online]. Available: http://www.sanken-mic.com/en/qanda/index.cfm/18.56. [Accessed 26 March 2017].

[45] A. Fukada, "A challenge in multichannel music recording," in *AES 19th International Conference*, Schloss Elmau, 2001.

[46] K. Hamasaki et al., "Approach and Mixing Technique for Natural Sound Recording of Multichannel Audio," in *AES 19th International Conference*, Schloss Elmau, 2001.

[47] K. Hamasaki and K. Hiyama, "Reproducing Spatial Impression With Multichannel Audio," in *AES 24th International Conference*, Banff, 2003.

[48] H. Wittek, "Double M/S - a Surround recording technique put to test," Schoeps Mikrofone, 2010.

[49] M. Gerzon, "Practical Periphony: The Reproduction of Full-Sphere Sound," in *AES Convetion 65*, London, 1980.

[50] M. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio Eng. Soc.,* vol. 33, no. 11, pp. 859-871, 1985.

[51] R. Brice, "Ambisonics," Psatial Audio, 2014.

[52] M. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," Mathematical Institute, University of Oxford, 1975.

[53] Soundfield, "ST450 MKII Portable Microphone System: User Guide, Version 1," Soundfield, London.

[54] F. Hollerweger, "An Introduction to Higher Order Ambisonic," Plone Foundation, 2008.

[55] H. Shim et al., "Perceptual Evaluation of Spatial Audio Quality," in *AES Convention 129*, San Francisco, 2010.

[56] M. Boone et al., "Spatial Sound-Field Reproduction by Wave-field Synthesis," *J. Audio Eng. Soc.,* vol. 43, no. 12, pp. 1003-1012, 1995.

[57] S. Kim, M. Ikeda and B. Martens, "Reproducing Virtually Elevated Sound via a Conventional Home-Theater Audio System," *J. Audio Eng. Soc.,* vol. 62, no. 5, pp. 337-344, 2014.

[58] J. Kotches et al., "DVD Benchmark - Part 6 - DVD-Audio," Home Theatre and Audio Review, March 2005. [Online]. Available: http://hometheaterhifi.com/volume_8_4/dvd-benchmark-part-6-dvd-audio-11-2001.html. [Accessed 2 May 2017].

[59] M. Bishop, Interviewee, *email correspondence.* [Interview]. 12 October 2017.

[60] "10.2 Surround Sound," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/10.2_surround_sound. [Accessed 2 May 2017].

[61] "Ultra High Definition Television Audio Characteristics and Audio Channel Mapping for Program Production," in *SMPTE ST 2036-2-2008*, White Plains, The Society of Motion Picture and Television Engineers, 2008.

[62] T. Hinata et al., "Live Production of 22.2 Multichannel Sound for Sports Programs," in *AES 40th International Conference*, Tokyo, 2010.

[63] K. Irie and T. Miura, "The production of 'The Last Launch of the Space Shuttle" by Super Hi-Vision TV," in *Broadcast Engineering Conference, NAB show 2012*, Los Vegas, 2012.

[64] I. Sawaya et al., "Dubbing Studio for 22.2 Multichannel Sound System in NHK Broadcasting Center," in *AES Convention 138*, Warsaw, 2015.

[65] I. Sawaya et al., "Discrimination of Changing Loudspeaker Positions of 22.2 Multichannel Sound System Based on Spatial Impressions," in *AES Convention 134*, Rome, 2013.

[66] T. Sugimoto, "Downmixing Method for 22.2 Multichannel Sound Signal in 8K Super Hi-Vision Broadcasting," *J. Audio Eng. Soc.,* vol. 63, no. 7/8, pp. 590-599, 2015.

[67] T. Komori et al., "Subjective loudness of 22.2 multichannel programs," in *AES Convention 138*, Warsaw, 2015.

[68] S. Kim et al., "New 10.2-channel Vertical Surround System (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers," in *AES Convention 129*, San Francisco, 2010.

[69] "ATSC Standard: A/342 Part 1, Audio Common Elements," Advanced Television Systems Committee, Washington, DC, 24 January 2017.

[70] "Experience," Auro Technologies, [Online]. Available: http://www.auro-3d.com/consumer/experience/. [Accessed 2 May 2017].

[71] D. Bowles, "A Microphone Array for Recording in Surround-Sound with Height Channels," in *AES Convention 139*, New York, 2015.

[72] P. Geluso, "Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays," in *AES Convention 132*, Budapest, 2012.

[73] "Productions in Auro-3D," Auro Technologies, 2012.

[74] A. Ryaboy, "Exploring 3D: A subjective evaluation of surround microphone arrays catered for Auro-3D reproduction," in *AES Convention 139*, New York, 2015.

[75] G. Theile and H. Wittek, "Principals in Surround Recording with Height (v2.01)," in *AES Convention 130*, London, 2011.

[76] R. King et al., "A Survey of Suggested Techniques for Height Channel Capture in Multi-channel Recording," in *AES Convention 140*, Paris, 2016.

[77] "Multichannel sound technology in home and broadcasting applications," ITU-R BS.2159-7, Geneva, 2015.

[78] H. Lee, "The Relationship between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking," in *AES Convention 131*, New York, 2011.

[79] R. Wallis and H. Lee, "Vertical Stereophonic Localization in the Presence of Interchannel Crosstalk: The Analysis of Frequency-Dependent Localization Thresholds," *J. Audio Eng. Soc.,* vol. 64, no. 10, pp. 762-770, 2016.

[80] R. Wallis and H. Lee, "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources," *Applied Sciences,* vol. 7, no. 278, 2017.

[81] W. Woszczyk, "Acoustic Pressure Equalizers," *Pro Audio Forum,* pp. 1-24, 1990.

[82] H. Lee and C. Gribben, "Effect of Vertical Microphone Array Spacing for a 3D Microphone Array," *J. Audio Eng. Soc.,* vol. 62, no. 12, pp. 870-884, 2014.

[83] "Use Case: Morten Lindberg, 2L – Norway," Merging Technologies, [Online]. Available: http://www.merging.com/news/use-cases/morten-linderg-2l-norway. [Accessed 12 July 2015].

[84] Merging Technologies, "An Exceptionally Successful 2016 For Unamas And Mick Sawaguchi," [Online]. Available: http://www.merging.com/news/news-stories/an-exceptionally-successful-2016-for-unamas-and-mick-sawaguchi. [Accessed 25 1 2018].

[85] M. Sawaguchi, Interviewee, *Personal Correspondance with the Author.* [Interview]. 25 02 2016.

[86] M. Williams, "Microphone Array Design for localisation with elevation cues," in *AES Convention 132*, Budapest, 2012.

[87] M. Williams, "The Psychoacoustic Testing of a 3D Multiformat Microphone Array Design, and the Basic Isosceles Triangle Structure of the Array and the Loudspeaker

Reproduction Configuration," in *AES Convention 134*, Rome, 2013.

[88] M. Williams, "Microphone Array Design applied to Complete Hemispherical Sound Reproduction – from Integral 3D to Comfort 3D," in *AES Convention 140*, Paris, 2016.

[89] M. Williams, "Downward Compatibility Configurations when using a univalent 12 Channel 3D Microphone Array Design as a Master Recording Array," in *AES Convention 137*, Los Angeles, 2014.

[90] H. Wittek, Interviewee, *Correspondence between H. Wittek and W. Howie.* [Interview]. 12 July 2016.

[91] K. Ono et al., "Portable spherical microphone for Super Hi-Vision 22.2 multichannel audio," in *AES Convention 135*, New York, 2013.

[92] R. J. Ellis-Geiger, "Music Production for Dolby Atmos and Auro 3D," in *AES Convention 141*, Los Angeles, 2016.

[93] C. Baume and A. Churnside, "Upping the Auntie: A Broadcaster's Take on Ambisonics," in *AES Convention 128*, London, 2010.

[94] B. Martin et al., "Microphone Arrays for Vertical Imaging and Three-Dimensional Capture of Acoustic Instruments," in *AES Conference on Sound Field Control*, Guilford, 2016.

[95] B. Martin et al., "Subjective Graphical Representation of Microphone Arrays for Vertical Imaging and Three-Dimensional Capture of Acoustic Instruments, Part I," in *AES Convention 141*, Los Angeles, 2016.

[96] E. Bates et al., "Comparing Ambisonic Microphones – Part 2," in *AES Convention 142*, Berlin, 2017.

[97] M. Ikeda et al., "New Recording Application for Software Defined Media," in *AES Convention 141*, Los Angeles, 2016.

[98] I. Choi et al., "Objective Measurement of Perceived Auditory Quality in Multichannel Audio Compression Coding Systems," *J. Audio Eng. Soc.,* vol. 56, no. 1/2, pp. 3-17, 2008.

[99] S. Choisel and F. Wickelmaier, "Relating auditory attributes of multichannel sound to preference and to physical parameters," in *AES Convention 120*, Paris, 2006.

[100] R. Conetta et al., "Spatial Audio Quality Perception (Part 2): A Linear Regression Model," *J. Audio Eng. Soc.,* vol. 62, no. 12, pp. 847-860, 2014.

[101] S. Kim et al., "Predicting Listener Preferences for Surround Microphone Technique through Binaural Signal Analysis of Loudspeaker-Reproduced Piano Performances," in *AES Convention 121*, San Francisco, 2006.

[102] S. Kim and B. Marten, "Deriving Physical Predictors for Auditory Attribute Ratings Made in Response to Multichannel Music Reproductions," in *AES Convention 123*, New York, 2007.

[103] J-H. Seo et al., "Perceptual Objective Quality Evaluation Method for High-Quality Multichannel Audio Codecs," *J. Audio Eng. Soc.,* vol. 61, no. 7/8, pp. 535-545, 2013.

[104] G. Sunish et al., "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *J. Audio*

Eng. Soc., vol. 58, no. 12, pp. 1013-1031, 2010.

[105] "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," in *ITU-R Recommendation BS.1116-1*, Geneva, International Telecom Union, 1997, pp. 1-26.

[106] S. Bech, "Methods for subjective evaluation of spatial characterstics of sound," in *AES 16th Interntional Conference*, Rovaniemi, 1999.

[107] S. Bech and N. Zacharov, Perceptual Audio Evaluation - Theory, Method and Application, Southern Gate, Chichester: John Wiley & Sons Ltd, 2006.

[108] N. Zacharov and T. H. Pederson, "Spatial sound attributes - development of a common lexicon," in *AES Convention 139*, New York, 2015.

[109] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.,* vol. 50, no. 9, pp. 651-666, 2002.

[110] F. Rumsey, "Subjective Assessment of the Spatial Attributes of Reproduced Sound," in *AES 15th International Conference*, Copenhagen, 1998.

[111] J. Berg and F. Rumsey, "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *AES 19th International Conference*, Schloss Elmau, 2001.

[112] J. Berg and F. Rumsey, "Systematic Evaluation of Perceived Spatial Quality," in *AES 24th International Conference*, Banff, 2003.

[113] J. Berg and F. Rumsey, "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique," *J. Audio Eng. Soc.,* vol. 54, no. 5, pp. 365-379, 2006.

[114] J. Berg and F. Rumsey, "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques," in *AES Convention 112*, Munich, 2002.

[115] N. Zacharov and K. Koivuniemi, "Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping," in *AES Convention 111*, New York, 2001.

[116] N. Zacharov and K. Koivuniemi, "Unravelling the perception of spatial sound reproduction: Techniques and experimental design," in *AES 19th International Conference*, Schloss Elmau, 2001.

[117] K. Koivuniemi and N. Zacharov, "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," in *AES Convention 111*, New York, 2001.

[118] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *J. Acoust. Soc. Am.,* vol. 121, no. 1, pp. 388-400, 2007.

[119] T. Kamekawa and A. Marui, "Developing Common Attributes to Evaluate Spatial Impression of Surround Sound Recordings," in *AES 40th International Conference*, Tokyo, 2010.

[120] S. Le Bagousse et al., "Categorization of Sound Attributes for Audio Quality Assessment—A Lexical Study," *J. Audio Eng. Soc.,* vol. 62, no. 11, pp. 736-747, 2014.

[121] C. Guastavino and B. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.,* vol. 116, no. 2, pp. 1105-1115, 2004.

[122] F. Rumsey et al., "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *J. Acoust. Soc. Am.,* vol. 118, no. 2, pp. 968-976, 2005.

[123] T. D. Rossing, "Acoustics in Halls for Speech and Music," in *Springer Handbook of Acoustics*, New York, Springer, 2007, pp. 302-315.

[124] T. Hanyu and S. Kimura, "A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections," *Applied Acoustics,* vol. 62, pp. 155-184, 2001.

[125] A. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, Cambridge: MIT Press, 1990.

[126] T. Letowski, "Sound Quality Assessment: Concepts and Criteria," in *AES Convention 87*, New York, 1989.

[127] W. C. Sabine, "Architectural Acoustics," *Proceedings of the American Academy of Arts and Sciences,* vol. 42, no. 2, pp. 51-84, 1906.

[128] J. Eargle, Handbook for Recording Engineers, New York: Van Nostrand Reinhold Company Inc., 1986.

[129] A. C. Gade, "Acoustics in Halls for Speech and Music," in *Springer Handbook of Acoustics*, New York, Springer Science+Business Media, LLC, 2007, pp. 301-350.

[130] Y. Ando, "Concert Hall Acoustics Based on Subjective Preference Theory," in *Springer Handbook of Acoustics*, New York, Springer Science+Business Media, LLC, 2007, pp. 351-386.

[131] J. S. Bradley and G. A. Soulodre, "Objective measures of listener envelopment," *J. Acoust. Soc. Am,* vol. 98, no. 5, pp. 2590-2597, 1995.

[132] Furuya et al., "Effect of early reflections from upside on auditory envelopment," *J. Acoustic. Soc. Jpn. (E),* vol. 16, no. 2, pp. 97-104, 1995.

[133] M. Morimoto et al., "The role of reflections from behind the listener in spatial impression," *Applied Acoustics,* vol. 62, no. 2, pp. 109-124, 2000.

[134] D. Griesinger, "Spaciousness and Envelopment in Musical Acoustics," in *AES Convention 101*, Los Angeles, 1996.

[135] P. Power et al., "Investigation into the Impact of 3D Surround Systems on Envelopment," in *AES Convention 137*, Los Angeles, 2014.

[136] R. Mason and F. Rumsey, "A comparison of objective measurements for predicting selected subjective spatial attributes," in *AES Convention 112*, Muchen, 2002.

[137] J. Meyer, Acoustics and the Performance of Music, Braunschweig: Springer, 2009.

[138] A. C. Gade, "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part I: Methods and Laboratory Experiments," *Acustica,* vol. 69, pp. 193-203, 1989.

[139] A. C. Gade, "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part II: Field Experiments and Synthesis of Results," *Acustica,* vol. 69, pp. 249-262,

1989.

[140] W. Woszczyk and J. Hong, “Listening Test Site Documentation, Studio 22 of New Music Building 3D Research Laboratory of the Graduate Program in Sound Recording,” McGill University, Montreal, 2014.

[141] S. Oode et al., “Vertical Loudspeaker Arrangement for Reproducing Specially Uniform Sound,” in *AES Convention 131*, New York, 2011.

[142] B. Martin et al., “Immersive content in three-dimensional recording techniques for single instruments in popular music,” in *AES Convention 138*, Warsaw, 2015.

[143] W. Howie and R. King, “Exploratory microphone techniques for three-dimensional classical music recording,” in *AES Convention 138*, Warsaw, 2015.

[144] “Surround Techniques,” DPA Microphones, [Online]. Available: http://www.dpamicrophones.com/en/Mic- University/Surround%20Techniques.aspx. [Accessed 29 June 2015].

[145] L. Simon and R. Mason, “Spaciousness Rating of 8-Channel Stereophony-Based Microphone Arrays,” in *AES Convention 130*, London, 2011.

[146] R. Kassier et al., “An Informal Comparison Between Surround-Sound Microphone Techniques,” in *AES Convention 118*, Barcelona, 2005.

[147] S. Kim et al., “An Examination of the Influence of Musical Selection on Listener Preferences for Multichannel Microphone Techniques,” in *AES 28th International Conference*, Piteå, 2006.

[148] S. Siegel and N. Castellan, Non parametric statistics for the behavioural sciences, New York: MacGraw Hill Int., 1988.

[149] J. Riedmiller and N. Tsingos, "Recent Advancements in Audio - How a Paradigm Shift in Audio Spatial Representation and Delivery Will Change the Future of Consumer Audio Experiences,"*2015 Spring Technical Forum Proceedings.*

[150] K. Hiyama et al., "The minimum number of loudspeakers and its arrangement for reproducing spatial impression of diffuse sound field," in *AES Convention 113*, Los Angeles, 2002.

[151] T. Sporer et al., "Localization of Audio Objects in Multi-channel Reproduction Systems," in *AES 57th International Conference*, Hollywood, 2015.

[152] W. Howie et al., "Listener preference for height channel microphone polar patterns in three-dimensional recording," in *AES Convention 139*, New York, 2015.

[153] B. Martin and R. King, "Mixing Popular Music in Three Dimensions: Expansion of the Kick Drum Source Image," in *Innovation In Music (InMusic'15)*, Cambridge, 2015.

[154] D. Griesinger, "The Science of Surround," New York, USA, 1995.

[155] N. Zacharov et al., "Next Generation Audio System Assessment using the Multiple Stimulus Ideal Profile Method," in *8th International Conference on Quality of Multimedia Experience*, Lisbon, 2016.

[156] W. Howie et al., "A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound," in *AES Convention 141*, Los Angeles, USA,

2016.

[157] M. Hietala, *Perceived differences in recordings produced with four surround microphone techniques,* Jyväskylä: Univ. of Jyväskylä, 2007, pp. 1-45.

[158] F. Camerer and C. Sodl, "Classical Music in Radio and TV - a Multichannel Challenge," 30 March 2015. [Online]. Available: http://www.hauptmikrofon.de/stereo-3d/orf- surround-techniques.

[159] M. Paquier et al., "Subjective assessment of microphone arrays for spatial audio recording," in *Forum Acusticum 2011*, Allborg, 2011.

[160] A. Sitek and K. B, "Study of Preference for Surround Microphone Techniques Used in the Recording of Choir and Instrumental Ensemble," *Archives of Acoustics,* vol. 36, no. 2, pp. 365-378, 2011.

[161] N. Peters et al., "Recording Techniques and their Effect on Sound Quality at Off-Center Listening Positions in 5.1 Surround Environments," *Canadian Acoustics,* vol. 41, no. 3, pp. 37-49, 2013.

[162] M. Chapman et al., "A standard for interchange of Ambisonic signal sets," in *Ambisonics Symposium*, Graz, 2009.

[163] J. Daniels, "Evolving Views on HOA: From Technological to Pragmatic Concerns," in *Ambisonics Symposium 2009*, Graz, 2009.

[164] N. Barret, "The Perception, Evaluation and Creative Application of Higher Order Ambisonics in Contemporary Music Practice," IRCAM, 2012.

[165] P. J. Power, Future Spatial Audio: Subjective Evaluation of 3D Surround Systems, Salford: University of Salford, 2015.

[166] A. Heller and E. M. Benjamin, "The Ambisonic Decoder Toolbox: Extensions for Partial- Coverage Loudspeaker Arrays," in *Linux Audio Conference 2014*, Karlsruhe, 2014.

[167] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.,* vol. 60, no. 10, pp. 807-820, 2012.

[168] A. Heller et al., "Is My Decoder Ambisonic?," in *AES Convention 125*, San Francisco, 2008.

[169] R. Mason and F. Rumsey, "An assessment of the spatial performance of virtual home theatre algorithms by subjective and objective methods," in *AES Convention 108*, Paris, 2000.

[170] H. Lee, "2D-3D Ambience Upmixing Based on Perceptual Band Allocation," *J. Audio Eng. Soc.,* vol. 63, no. 10, pp. 811-821, 2015.

[171] S. Kim et al., "A Cross-Cultural Comparison of Salient Perceptual Characteristics of Height-Channels for a Virtual Auditory Environment," *Virtual Reality,* vol. 19, no. 3, pp. 149-160, 2015.

[172] S. Kim et al., "Reproducing Virtually Elevated Sound via a Conventional Home-Theatre Audio System," *J. Audio Eng. Soc.,* vol. 62, no. 5, pp. 337-344, 2014.

[173] K. Matsui and A. Ando, "Binaural Reproduction of 22.2 Multichannel Sound with

Loudspeaker Array Frame," in *AES Convention 135*, New York, 2013.

[174] J. Francombe et al., "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preferences," *J. Audio Eng. Soc.,* vol. 65, no. 3, pp. 212-225, 2017.

[175] M. Parmentier, "Object-based audio: The Next Big Turn," *J. Audio Eng. Soc,* vol. 63, no. 7/8, pp. 659-660, 2015.

[176] W. Howie et al., "Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio," in *AES Convention 142*, Berlin, 2017.

[177] A. Ando, "Conversion of Multichannel Sound Signal Maintaining Physical Properties of Sound in Reproduced Sound Field," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 6, pp. 1467-1475, 2011.

[178] "Loudness Metering: 'EBU Mode' metering to supplement loudness normalization in accordance with EBU R 128," EBU - TECH 3341, Geneva, 2011.

[179] M. Schoeffler and J. Herre, "The relationship between basic audio quality and overall listening experience," *J. Acoust. Soc. Am.,* vol. 140, no. 3, pp. 2101-2112, 2016.

[180] S. Olive, "Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study," *J. Audio Eng. Soc.,* vol. 51, no. 9, pp. 806-825, 2003.

[181] W. S. Snow, "Basic Principals of Stereophonic Sound," *Journal of the SMPTE,* vol. 61, no. 5, pp. 567-586, 1953.

[182] H. Fletcher, “Auditory Perspective – Basic Requirements,” *Electrical Engineering,* vol. 53, no. 1, pp. 9-11, January 1934.

[183] J. Steinberg and W. B. Snow, “Auditory Perspective – Physical Factors,” *Electrical Engineering,* vol. 53, no. 1, pp. 12-17, 1934.

[184] S. Oode et al., “Dimensional Loudspeaker Arrangement for Creating Sound Envelopment",” IEICE Technical Report, EA2012-46, 2012.

[185] S. Bech, “Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment,” *J. Audio Eng. Soc.,* vol. 40, no. 7/8, pp. 590-610, 1992.