Sponsoring Committee: Professor Agnieszka Roginska, Chairperson
Doctor Olivier Warusfel
Professor Pablo Ripollés
Doctor Tomasz Żernicki

ACOUSTIC AND PERCEPTUAL FACTORS AFFECTING PLAUSIBILITY IN SOUND DESIGN FOR AUDIO AUGMENTED REALITY EXPERIENCES

Marta Gospodarek

Program in Music Technology
Department of Music and Performing Arts Professions

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
Steinhardt School of Culture, Education, and Human Development
New York University
2024

ABSTRACT

In Augmented Reality (AR) environments sound design faces the unique challenge of seamlessly integrating virtual elements into real environments while preserving the cohesion of the user's auditory experience. An important measure of the perceived realism of the auditory layer is plausibility. The concept of plausibility in AR audio experiences is multifaceted, incorporating both acoustic and perceptual factors. While acoustic factors originate mostly from the Virtual Acoustic Environment (VAE) design and its similarity to the real acoustic space around the user, perceptual factors stem from real auditory references originating from the environment. The core empirical study of this dissertation aims to investigate the factors posed above and at the same time validate a new methodology for plausibility evaluation in AR. During the study participants rated the plausibility of pairs of loudspeakers (real or virtual) emitting sound consecutively from different positions in the room, mimicking real-life scenarios where similar but not identical sources coexist.

Results showed that the primary factors influencing plausibility judgments during motion appeared to be the coherence between self-motion and auditory cues, as well as the alignment between visual and auditory cues. In contrast, during the standing phase, participants were confined to a single static perspective of the sources and as a result, they relied more heavily on comparing the two sources presented together in each trial. The analysis also indicated that plausibility perception was affected by manipulating the real reference's properties. Furthermore, all evaluated sound attributes contributed to the plausibility perception. The highest correlation with plausibility was observed for blur and localization error, highlighting the importance of congruency between visual anchors and sound. Additionally, the analysis showed that plausibility ratings were mainly correlated with distance estimation errors for virtual and sometimes real sources and were driven by the inaccuracy of the sound intensity values.

The findings contribute to broadening the understanding of the plausibility of sound design and sound perception in Extended Reality (XR) environments. Furthermore, the conclusions can be utilized by AR sound designers to inform their design choices. Finally, the proposed novel methodology proved to be effective in evaluating plausibility in the AR context and can be implemented in other studies to further the understanding of plausibility.

ACKNOWLEDGEMENTS

This dissertation would not be possible without the support of many amazing people who I was privileged to know along the way.

First and foremost, I express my deepest gratitude to my doctoral advisor, Prof. Agnieszka Roginska. I will be always thankful for showing me this path. Her unwavering support, guidance and encouragement through the highs and lows of my doctoral journey, as well as her inspiration in navigating the challenges of working mum, are deeply appreciated.

I extend my deep gratitude to my second advisor - Dr. Olivier Warusfel. I am very thankful for his patience and generosity in sharing his time and wisdom, his kindness, and detailed approach to research.

My warm appreciation goes to the rest of my doctoral committee members and to all NYU professors I had the privilege to learn from and work with. My appreciation goes to Prof. Pablo Ripolles for his patient assistance with statistical matters. To Dr. Tomasz Żernicki, for his assistance along the way. To Prof. Ken Perlin, I am grateful for introducing me to the wonders of XR and expanding my creative horizons in sound. To Prof. Morwaread Farbood for her valuable feedback. I am also indebted to Prof. Anat Lubetzky for introducing me to new areas of spatial audio impact.

My deep gratitude goes to Theresa Leonard, for inviting me to the life-changing adventure in Banff and showing me possibilities I would never think of.

To Prof. Andrzej Miśkiewicz, the best master thesis advisor I could imagine who helped me write my first publication which opened so many doors. To all the Professors at the Music University in Warsaw, especially Prof. Andrzej Bohdanowicz for teaching me the art of sound engineering.

I also wish to acknowledge the MARL PhD students who preceded me: Andrea Genovese, Rachel Bittner, and Andrew Telichan. Your friendship and support were invaluable during my time in New York and later. I am also thankful to everyone else from MARL with whom I interacted, especially Ana Elisa, Willie, Yu, Mark, Charlie, and Peter.

To all the master students I had the pleasure of working with over the years, including Denis, Makan, Tatiana, Christie, Greg, Scott, Simon, Yu, and all others who participated in the 3D Audio Interest Group and Immersive Audio Group, it has been an

enriching experience collaborating with you. I extend my gratitude to everyone from FRL with whom I shared the excitement of discovering co-located VR and collaborating on various projects, especially to Sebastian, Zhu, Kris, Karl, and Corinne.

To my colleagues at IRCAM, Vincent and Pierre, thank you for your assistance with Matlab scripts and never-ending measurements.

I am also deeply grateful to everyone who patiently participated in my listening tests in Paris and New York; your contribution was indispensable to the completion of this thesis.

To my friends Iza, Magda, Ksenia, Marta, Irenka, Magda, Dulce, Jorge, and Alex, thank you for your unwavering support, calls, and visits, despite the challenges of distance and time.

To my grandparents, aunts, siblings, and their families Maciej, Mateusz, Matylda, Marcin, Michał, Daria, Kaja, Marysia, and all my nieces - you helped me in so many ways, there is not enough space to write them all.

To my Mum and Dad, there are not enough words to say how grateful I am and how much in this journey I owe you.

To my children, Kubuś and Łucja, for showing me the proper perspective on work and for letting me enjoy doctoral studies for a little longer.

Finally, to my husband Adam, if there is any value in this work, half of the credit goes to you for the countless ways you supported me during our life together. Thank you for being part of it.

TABLE OF CONTENTS

LIST OF FIGURES xii

CHAPTER

LIST OF TERMS xvi

I INTRODUCTION 1

1. Sound Design for Augmented Reality 2
2. Statement of Purpose 3
3. Research Questions 5
4. Methodology 7
5. Motivation 8
6. Dissertation Outline 10
7. Related Articles 11

II VIRTUAL ACOUSTICS ENVIRONMENT 14

1. Introduction 14
2. Sound Source Modeling 15
   2.1 Influence of Directivity of the Sound Source on Perception 16
3. Room modeling 17
   3.1 Physical Based Approaches to Room Modeling 19
   3.2 Convolution-based Approaches to Room Modeling 20
   3.3 Algorithmic Approaches to Room Modeling 21
   3.4 Acoustic Modeling and Perception 22
4. Listener Modeling with Head-Related Transfer Functions 24
   4.1 Influence of Individualized HRTF on Perception 25
5. Other Factors in Designing Virtual Acoustics System 26
   5.1 Latency and Its Perceptual Salience 26
   5.2 3DoF vs 6DoF Environments 27
   5.3 Audio-Visual Interaction 28

III PERCEPTUAL EVALUATION OF SOUND ATTRIBUTES IN XR CONTEXT 30

1. Methods for the evaluation of audio in AR and VR - plausibility and presence 30
   1.1 3DoF Environments 32
   1.2 6DoF Environments 33
2. Identification of Perceived Sound Quality Attributes in VR 37
   2.1 Introduction 37
   2.2 Elicitation Methods for Sound Quality Evaluation 38

2.3 Spatial Attributes in Literature 40
2.4 Techniques Used for Audio Production in VR Application 43
2.5 Experiment 44
2.6 Analysis and Discussion 48
2.7 Conclusion and Future Work 52

IV SOUND DESIGN PRACTISES IN XR 55

1. Goals and Challenges of Sound Design for Mixed Reality Experiences 55

2. Case Study: Sound Design and Reproduction Techniques for Co-Located Narrative VR Experiences 57
2.1 Introduction 57
2.2 Background 59
2.3 Design Factors and Principles 61
2.4 *"Cave"*: A Case Study 68
2.5 Survey 73
2.6 Discussion 75
2.7 Conclusion 78

V EXPERIMENT: EVALUATION OF PLAUSIBILITY AND OTHER SOUND ATTRIBUTES IN AAR CONTEXT 79

1. Research Questions 80

2. Goal and Purpose of the Study 81

3. Overview 83

4. Room and Loudspeaker Directivity Measurements 85

5. Implementation of Virtual Acoustics Environment 88
5.1 Direct Sound and Directivity Modeling 88
5.2 Auralization Based on GA Simulation 90
5.3 Auralization Based on SRIR Synthesis 94
5.4 HOA to Binaural Decoding and Equalization 96
5.5 Compensation Filters 96
5.6 Calibration and Objective Comparison 97

6. Subjective Listening Tests 98
6.1 Experimental Setup 99
6.2 Stimulus Choice and Preparation 99
6.3 Conditions 100
6.4 Participants 101
6.5 Invitation Procedure 102
6.6 Collection Method 102

VI RESULTS OF THE EXPERIMENT - WALKING PHASE 107

1. Analysis Method 108
1.1 Fixed and Random Effects 109
1.2 Summary of Dependent Variables 113

2. Results analysis 115
2.1 Plausibility 116
2.2 Blur 117
2.3 Localization Error 120
2.4 Loudspeaker Recognition Rate 122

2.5 Externalization 124
2.6 Timbre Difference 126
2.7 Reverberation Difference 127
2.8 Plausibility Difference 128
2.9 Speed of Walking 129

3. Discussion 131
3.1 What Are the Main Factors That Influence the Plausibility of an Audio-only Augmented Reality Scene and How Do They Affect Evaluation? 131
3.2 How Other Attributes of Sound Were Affected by the Method of Playback and Loudspeaker Position? 133
3.3 Secondary Effects 139

4. Summary 142

VII RESULTS OF THE EXPERIMENT - WALKING VS STANDING PHASE 151

1. Comparison of Walking and Standing Phases 151
1.1 Plausibility 152
1.2 Blur 153
1.3 Localization Error 158
1.4 Loudspeaker Recognition Rate 159
1.5 Externalization 163
1.6 Timbre Difference 166
1.7 Reverberation Difference 167
1.8 Plausibility Difference 170
1.9 Amplitude of Yaw Movement - Standing Phase 171
1.10 Influence of Yaw Movement on Evaluation in Standing Phase 173
1.11 Influence of Height Difference 175
1.12 Influence of Trial Index 178
1.13 Correlation between Plausibility and Other Attributes 179

2. Discussion 182
2.1 Influence of Participant's Movement on Subjective Evaluation 182
2.2 What Is the Influence of Secondary Effects on Sound Evaluation? 193
2.3 What Behavioral Measures Tell About the Interaction of Movement and Perceptual Evaluation? 196
2.4 Is the Proposed Methodology an Effective Method for Evaluating Plausibility and Other Sound Attributes in 6DoF AR Environments? 199
2.5 Which Attributes Contribute to the Plausibility Perception? 203

3. Objective Analysis 204
3.1 Energy of Time Segments for Each Measurement Point 204
3.2 Total Energy of the Signal 207
3.3 Direct-to-Reverberant Ratio 207
3.4 C_{50} 209
3.5 C_{80} 211
3.6 $[1 - IACC_{E_3}]$ - Apparent Source Width 211
3.7 Diffuseness 214
3.8 Direction of Arrival 216
3.9 Elevation 218

4. Correlation Between Perceptual and Acoustic Factors 218
4.1 Energy of Time Segments 219
4.2 Plausibility and Blur 221

4.3 Distance Perception 224
4.4 Interaction of Self-Motion and Auditory Cues 226
4.5 Source Orientation 227
4.6 Visual Cues 229
4.7 Early Reflections 230
4.8 Non-Individualized HRTFs 231

5. Limitations and Future Work 232

VIII CONCLUSIONS AND FUTURE WORK 235

1. Summary of Contributions 235

2. Implications for Sound Design in AR 241

3. Future Directions 242
3.1 Varying the Signal 242
3.2 Exploring Simplifications of Auralization Methods 243
3.3 Stimuli Types 244
3.4 Participants Selection 244
3.5 Behavioral Measures in Plausibility Evaluation 244
3.6 Sound Source Orientation 245
3.7 Non-Individualized HRTFs in 6DoF 245

4. Conclusions 245

REFERENCES 247

APPENDICES

A INFORMATION NOTICES 261

B SIMULATIONS DIAGRAMS 266

C STATISTICAL ANALYSIS - WALKING PHASE 270

1. Plausibility 270

2. Blur 272

3. Localization Error 275

4. Externalization 278

5. Loudspeaker Recognition Rate 280

6. Timbre Difference 283

7. Reverberation Difference 284

8. Plausibility Difference 285

9. Speed 287

D STATISTICAL ANALYSIS - WALKING AND STANDING PHASES 289

1. Plausibility 289

2. Blur 294

3. Externalization 299

4. Localization Error 301

5. Loudspeaker Recognition Rate 305
6. Timbre Difference 307
7. Reverberation Difference 309
8. Plausibility Difference 311
9. Yaw Movement - Standing Phase 313

LIST OF FIGURES

1 Factors influencing plausibility perception 6

2 Model of virtual acoustics rendering system from Huopaniemi (1999) 7

3 Original source and image sources from Allen & Berkley (1979) 20

4 Impulse response divided into 3 segments from Yilmaz, (2010) 21

5 Rate of appearance of spatial sound attributes 53

6 Number of verbal descriptors elicited during experiment 53

7 Audience avatars in the VR experience *"Cave"*. (© and Art: Kris Layng, 2018) 69

8 Headset with prototypes of nearfield open-ear device from Bose. (Photo: Eric Chang, 2018) 72

9 Acoustic and perceptual factors of plausibility 84

10 Studio 5 at IRCAM 86

11 Setup for measurements - Studio 5 at IRCAM 87

12 Directivity of the PMX 5 loudspeaker 88

13 Direct sound simulation 89

14 Directivity index of loudspeaker PMX5 90

15 The results of directivity model simulation: the speaker measurements and simulation deconvolved with on-axis signal 91

16 Measured RT_{60} and its FDN synthesis 94

17 Participant wearing AKG K1000 headphones with Vive Tracker during listening test 105

18 Screenshot of the graphic interface used for question about localization 106

19 Predicted means and 95% confidence intervals for plausibility (scale: 0 - not at all plausible, 6 - very plausible), blur (scale: 0 - very focused, 6 - very blurry) and localization error ratings in walking phase according to loudspeaker position and rendering method (***P $<$ 0.05). 118

20 The influence of walking speed on plausibility, blur ratings, and percentage of correctly recognized loudspeaker positions (***P $<$ 0.05) 119

21 Stimulus position reported by participants during the walking phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Positions [0,0] and [0,7.2] mark the starting and ending points of the walking path. 145

22 Predicted means and 95% confidence intervals for probability of correct recognition of loudspeaker position in walking phase. Black brackets indicate statistically significant differences between methods while red brackets indicate statistically significant differences between the two orders of playback. 146

23 Distribution of loudspeaker recognition errors in walking phase. The pointed loudspeaker indicates which loudspeaker was the closest to the position marked by the participant. For better clarity, the vertical scale was limited to the 0-5% range. 146

24 Predicted values and 95% confidence intervals for externalization ratings in walking phase according to the rendering method and loudspeaker position (***P $<$ 0.05) 147

25 Predicted means and 95% confidence intervals for timbre, reverberation, and plausibility difference ratings according to loudspeaker pair and rendering method pair(***P $<$ 0.05) 148

26 Histogram of speed of walking data 149

27 Histogram of mean of speed for all the trials 149

28 Influence of trial index and playback method on the walking speed of participants 150

29 Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05) according to the rendering method, loudspeaker position, and phase. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry). 154

30 Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05) according to the rendering method, loudspeaker position and phase. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry). 155

31 Predicted means and 95% confidence intervals for plausibility ratings according to the rendering method pair (***P < 0.05). 156

32 Stimulus position reported by participants during the walking phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Positions [0,0] and [0,7.2] mark the starting and ending points of the walking path. 160

33 Stimulus position reported by participants during the standing phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Position [0,2.1] marks the standing point during the experiment. 161

34 Predicted means and 95% confidence intervals for percentage of correctly recognized loudspeaker positions in both phases according to the loudspeaker position and phase. 163

35 Distribution of loudspeaker recognition errors according to the target loudspeaker and the phase. The pointed loudspeaker indicates which loudspeaker was the closest to the position marked by the participant. For better clarity, the vertical scale was limited to the 0-10% range. 164

36 Predicted probability and 95% confidence intervals for an answer ”outside of the head” for externalization question according to the loudspeaker position and phase (plot on the left) and rendering method (plot on the right) (***P < 0.05) 166

37 Predicted means and 95% confidence intervals for timbre, reverberation, and plausibility difference ratings according to the phase and loudspeaker pair (plots on the left) and the rendering method pair (plots on the right) (***P < 0.05) 168

38 Influence of loudspeaker position on amplitude of yaw movement of participants 172

39 Influence of playback method on the amplitude of yaw movement of participants 173

40 Influence of amplitude of yaw movement of participants on plausibility, blur, externalization ratings, and localization error in standing phase. 174

41 Histogram of the absolute value of height difference between participant and loudspeaker 176

42 Influence of height difference between participant and loudspeaker on plausibility and blur ratings and localization error. 177

43 Influence of height difference between participant and loudspeaker on externalization and timbre difference ratings. 178

44 Predicted means and 95% confidence intervals for blur ratings, and localization error. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry). 180

45 Influence of trial index on the amplitude of yaw movement of participants 180

46 Room setup depicting angles and distances from the listening point to each of the loudspeakers during the standing phase 187

47 Energy of time segments for each measurement point for all loudspeakers (0-5 ms, direct sound, 5-80ms - early reflections, >80ms - late reverberation). The vertical line denotes the point at which participants were positioned during the standing phase. 206

48 Total energy of the signal for each measurement point for all loudspeakers. The vertical line denotes the point at which participants were positioned during the standing phase. The black line indicates the theoretical values of the energy evolution according to the distance law. 208

49 DRR for each measurement point for all loudspeakers (sum of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase. 210

50 C_{50} for each measurement point for all loudspeakers (mean of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase. 212

51 C_{80} for each measurement point for all loudspeakers (mean of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase. 213

52 ASW for each measurement point for loudspeaker D 215

53 Diffuseness calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The circles indicate values at the standing point. Means along the path are represented on the right. 216

54 Direction of arrival calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The thick lines indicate the estimation of direction at the standing point. 217

55 Source elevation calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The circles indicate values at the standing point. Means along the path are represented on the right. 218

56 Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05). The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry). 220

57 Percentage of loudspeaker recognition errors for different rendering method pairs in standing phase 226

58 Elements of the direct sound segment sound path for real situation, simulations and measurements 267

59 Elements of the early reflections segment sound path for real situation, simulations and measurements 268

60 Elements of the reverberation segment sound path for real situation, simulations and measurements 269

LIST OF TERMS

3 Degrees of Freedom (3DoF)

Technology able to sense and react to 3-dimensional rotation of a device using gyroscopes. 3DoF tracking system responds to the rotation of the user's head (yaw, pitch, and roll).

6 Degrees of Freedom (6DoF)

Expansion on 3DoF by the addition of sensors able to track the XYZ position of a device within a space. A 6DoF tracking system can respond to a user walking inside a space.

Ambisonics

A full-sphere surround sound format which in addition to the horizontal plane, entails sound sources above and below the listener. The order of the Ambisonics indicates the number of channels (varying as the square of the order) and microphone capsules of the Ambisonic microphone. The higher the order the better the spatial resolution. In contrast with channel-based formats (stereo, 5.1, etc.), the number and position of the recording microphones or reproduction loudspeakers are not mandatory, as long as they are regularly distributed in space.

Augmented Reality (AR)

An interactive experience of a real-world environment enhanced by computer-generated perceptual information.

Binaural Room Impulse Response (BRIR)

HRIRs measured in the presence of reverberation.

C50

C50 is related to the attribute clarity or intelligibility of speech and represents the ratio of the early sound energy (between 0 and 50 ms) and the late sound energy (that arrives later than 50 ms).

C80

C80 is related to the music clarity and represents the ratio of the early sound energy (between 0 and 80 ms) and the late sound energy (that arrives later than 80 ms).

Direct-to-Reverberant Ratio (DRR)

The direct-to-reverberant ratio quantifies the ratio of direct sound to reverberant sound in an acoustic environment, with a higher DRR indicating a stronger dominance of direct sound over reverberation..

Extended Reality (XR)

An umbrella term that encompasses all immersive technologies, including virtual reality (VR), augmented reality (AR), and mixed reality (MR). It refers to any environment that combines real-world elements with virtual or digital content, thereby extending the user's perception of reality.

Feedback Delay Network (FDN)

Algorithmic approach for reverberation modeling using parallel delay lines connected recursively through a feedback matrix.

First Order Ambisonics (FOA)

Ambisonics of first order. It consists of four channels (W, X, Y, Z), with the W channel representing the omni-directional component (sound pressure), and the X, Y, and Z channels representing the directional components (sound velocity).

Geometrical Acoustics (GA)

Geometrical acoustics is a branch of acoustics that models sound propagation by treating sound waves as rays, allowing for the prediction of sound behavior based on the principles of ray optics, such as reflection, refraction, and diffraction, particularly in environments where wavelengths are much smaller than the dimensions of objects and structures..

Head-related Impulse Response (HRIR)

A stereo acoustic filter which describes the path of a sound source to the human ears from a defined location.

Head-related Transfer Function (HRTF)

An acoustic transfer function between a point sound source in the free-field and a listener's ear canal.

Higher Order Ambisonics (HOA)

Ambisonics of order second and higher.

Interaural Cross-Correlation Coefficient (IACC)

Measure used to quantify the similarity of sound signals arriving at the two ears. It assesses the degree of correlation between the two signals, providing information about the spatial properties of a sound field, such as the perceived width or spaciousness of a sound source.

Plausibility

Plausibility refers to the perceived realism or believability of a sound or auditory event. It encompasses how well a sound conforms to our expectations based on our understanding of the auditory environment and our past experiences with similar sounds.

Room Impulse Response (RIR)

The response of an acoustic channel (open space, room) to an impulse sound signal emitted by a source and captured by a receiver, used to characterize the acoustic properties of a space as sound is reflected by hard boundaries. It can be measured or modeled.

Spatial Room Impulse Response (SRIR)

Impulse response that characterizes acoustic properties of a space recorded in Ambisonics format.

Virtual Reality (VR)

The computer-generated simulation of a three-dimensional image and sonic environment that can be interacted with in a seemingly real or physical way by a person.

CHAPTER I

INTRODUCTION

Over the past decade, the rapid advancement of Virtual Reality (VR) systems has forced a significant shift in the priorities of the audio industry. This evolution is reflected in a joint effort among audio professionals and scientists to enhance immersive audio technologies. Recognizing the limitations of traditional stereo systems in creating truly convincing spatial audio layers, there has been a return of interest in 3D audio formats such as binaural and Ambisonics, originally conceptualized nearly six decades ago. These formats are being explored and exploited in new application fields to meet the demands of modern XR environments (Roginska & Geluso, 2017).

The fast progress of VR technology over recent years is now paralleled by a surging interest in Augmented Reality (AR) technologies. AR systems, designed to seamlessly integrate virtual elements into the real environment, impose more strict requirements for both visual and auditory components.

The growing market demand for commercial hardware supporting AR audio and visual capabilities signifies a significant step towards the realization of high-fidelity AR systems. Yet, despite these advancements, several technological hurdles remain to be overcome to ensure a seamless and immersive user experience. Key challenges in the audio domain include accurately identifying the acoustic characteristics of the user's environment, crafting convincing acoustic environments with limited processing resources, and enhancing the fidelity of spatialization algorithms.

Before resolving these technical obstacles, a deep understanding of auditory perception within the context of AR experiences is critical to drive further advancements and refinements of AR technology. This understanding serves as the foundation for the ongoing improvement and evolution of AR systems.

1 Sound Design for Augmented Reality

The evolution of technology coincides with the emergence of innovative artistic expressions in XR realms. Within this dynamic landscape, sound design plays a pivotal role, serving as the bridge between technological advancements and artistic expression. Unlike sonic art, which stands as an independent form of creative expression, sound design is focused on crafting audio experiences with a purpose beyond themselves (Gibbs, 2007).

While sound design is traditionally associated with the realms of film and television, its scope extends far beyond, encompassing diverse fields and applications. In VR, sound design is instrumental in constructing immersive auditory layers that transport users into virtual realms seamlessly. Conversely, in AR, sound design faces a unique challenge: seamlessly integrating virtual elements into the real environment while maintaining the integrity of the user's auditory experience.

This challenge in AR underscores the importance of understanding the factors influencing plausibility of sound design in this context.

2 Statement of Purpose

The AR audio experience aims to create a perfect illusion of the virtual sources being part of the real environment. The quality assessment of these experiences can be approached using two different paradigms. One is aimed to quantify the subject performance during a given task, and from it conclude the quality of the rendering system. The advantage of this method is that it gives more objective measurements. Another approach is to evaluate realism which is subjective and challenging to measure.

Evaluating the realism of an AR experience is a multimodal task as many factors contribute to it such as the quality of the rendering system, the task of the user, modes of interaction, or the user's personality. Thus, it is necessary to precisely define the measure of the quality being assessed. One of the essential measures of perceived realism is authenticity which is commonly defined as the perceptual identity of real and virtual events. Authenticity sets a very difficult goal that might not be feasible to achieve in most rendering systems. The auditory system has high perceptual accuracy in discriminating level differences as little as 1 dB when immediate comparison is provided. However, in most real-life situations, the exact same reference to a virtual source is not available. That is why plausibility seems to be a more appropriate measure of audio experience quality. Lindau suggests the definition of plausibility as "a simulation in agreement with the listener's expectation towards a corresponding real event" (Lindau & Weinzierl, 2012) which applies to AR. Taking all of that into account, plausibility seems to be an appropriate measure of the overall user experience in audio AR.

Plausibility is a complex percept that can be affected by both acoustic as well as perceptual factors (refer to Figure 1). Acoustic factors relate to the design of a Virtual Acoustic Environment. Any virtual acoustic system consists of three base modules: modeling of the source, modeling of the environment, and modeling of the listener (as shown in Figure 2). In the case of AR audio where the goal is to achieve a seamless connection between real and virtual elements of the environment, modules aim to mimic the characteristics of reality, in particular the real source, the real space the user is in, and the particular listener. A very accurate rendering of the sound layer needs a lot of detailed measurements and computational resources which usually are not available on consumer devices. This is why current research is dedicated to discovering efficient and simpler methods for dynamic room auralization, capable of maintaining very high sound quality. Consequently, identifying the most critical aspects of VAE for sound plausibility and determining which acoustic parameters to prioritize to ensure high plausibility of the simulated sound layer are key objectives of this dissertation.

The plausibility of sound is not solely determined by acoustic factors; perceptual factors also play a significant role in shaping plausibility judgments. For instance, as plausibility evaluates sound in comparison to an internal reference, it is reasonable to assume that this reference can be adjusted by real sounds heard during the AR experience, consequently influencing the evaluation. User movement is another factor that can significantly influence plausibility judgments. In the majority of AR experiences, users are expected to move in 6 Degrees of Freedom (6DoF). Depending on the nature of the user's movement, their perception of sound may be affected. Movement provides users with access to a broader range of acoustic cues, potentially altering their perception of sound. Also, the interaction between self-motion and changing auditory cues might influence the interpretation of the sound.

Another aspect of plausibility perception that remains mostly unexplored is its relationship to other sound attributes. Sound perception has been extensively studied, revealing a number of attributes crucial to overall quality judgment. However, the link between these attributes and the perception of plausibility remains an open question. It is unclear whether certain attributes hold more importance for plausibility or if all attributes contribute equally. Further research is needed to explore these relationships and better understand the factors influencing the plausibility of sound in AR experiences.

3 Research Questions

The main research question addressed by this dissertation is the following:

What perceptual and acoustic factors influence the perception of plausibility in Audio Augmented Reality environments?

The proposed study will aim to answer a set of subquestions which can be divided into several areas of interest posed below.

Perceptual Evaluation

- How does a subject's freedom of movement affect the perceptual evaluation of an AR sound scene?
- What is the correlation between plausibility and other perceptual attributes of sound?

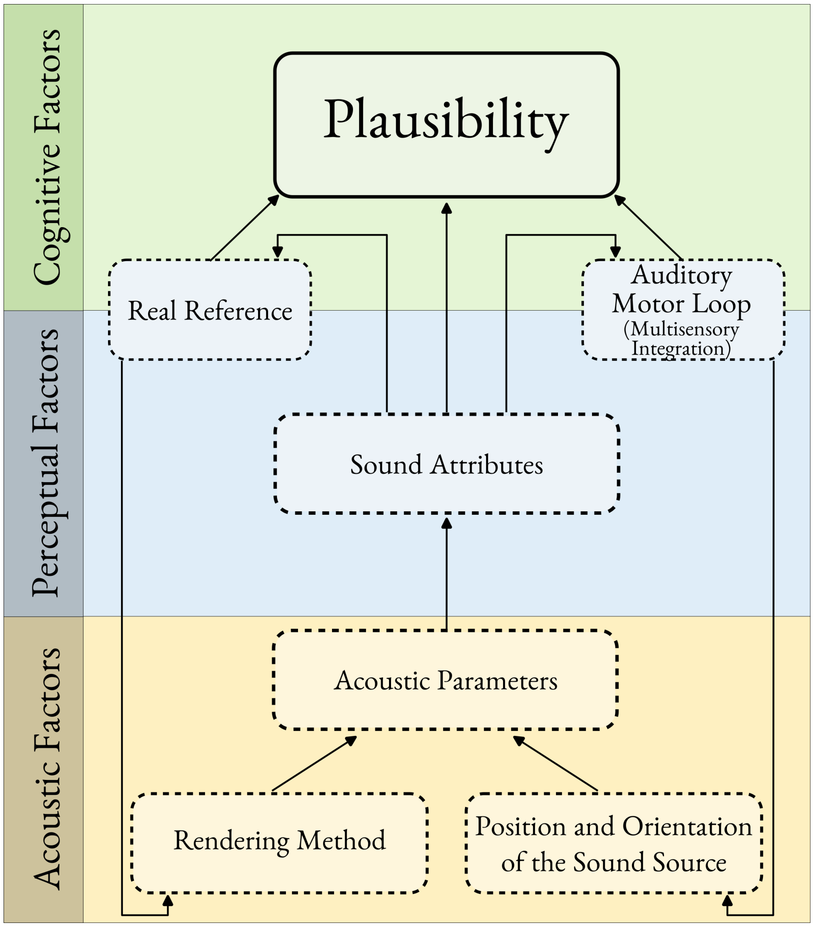


Figure 1: Factors influencing plausibility perception

- Do the properties of real reference affect plausibility judgment?

Acoustics

- How do objective measures of acoustical parameters correspond to subjective evaluation of acoustic processing?
- How does the position of the source in the room and orientation influence the assessment of the auralizations?

Methodology

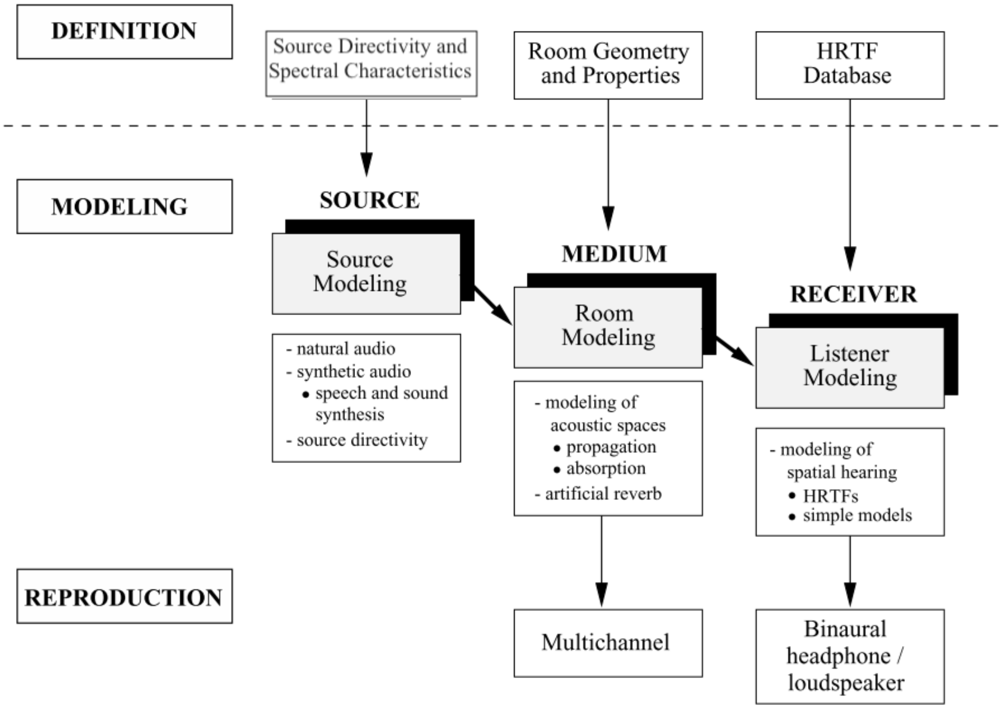


Figure 2: Model of virtual acoustics rendering system from Huopaniemi (1999)

- Is the proposed methodology an effective method for evaluating plausibility in 6 Degrees of Freedom (6DoF) AR environments?
- How do the participants' speed of walking and amplitude of yaw movement affect the evaluation?

4 Methodology

In order to address the questions stated above, the core of this dissertation is an experimental study focused on evaluation of virtual and real sources in AR context. This study proposes a new methodology where participants compare virtual sources to the real reference which is heard from another location. Another localization of the reference sound provides a lot of information about the source and space but it does not require a perfect match between the simulation and the real loudspeaker. This approach avoids the ceiling effect when evaluating plausibility but at the same time allows for evaluation in conditions similar to real-life AR scenarios. Besides that, the dissertation study focuses on the influence of different acoustic simulations on the perceived realism of virtual sound sources within an AR environment. Each simulation is developed with a very limited number of room and source measurements. The first method is based on geometrical acoustic modeling of the room. The method combines a real-time image-source algorithm for the simulation of early reflections and an FDN for the rendering of late reverberation. In the second method, the acoustical properties of the room are first characterized by a single room impulse response performed with a 4th-order Ambisonics microphone. This one measurement is then transformed in real-time to represent different user listening perspectives.

5 Motivation

It is only recently that the hardware capabilities allowed to design dynamic sound modeling systems with sufficient quality to begin investigation on the perception in the AR context. Present versions of audio software for AR are largely underdeveloped in terms of implementation for high-quality rendering. This is due mainly to hardware constraints, but also to difficulties with the effective implementation of acoustic modeling. As of today, there is no fast and efficient way of obtaining individualized Head-related Impulse Response (HRIR) (although many different approaches are being researched). Implementation of source directivity is also a challenge, especially for complex radiation patterns such as musical instruments. Acoustical modeling systems are also limited in terms of the computational power required to calculate accurate acoustic models in real-time. However, before all of these obstacles are solved there is a need to define what kind of simplifications can be employed without reducing the perceived quality of the sound simulation. In general, the number of perceptual studies in AR environments is relatively small. There is little understanding and research on the psychoacoustic and acoustic factors that impact the perception of realism. It is still not clear what are the most important parameters which make sounds appear to be part of our natural environment. This research will allow us to gain a better understanding of acoustic factors affecting plausibility and how the perception of them is changed depending on the auditory-motor interaction of the listener. In addition, there are currently no established methodologies for the study of sound perception in 6DoF environments. This study will seek to address this gap by validating new methods for perceptual evaluation of the audio AR environment quality. This research will not only gain insight into designing more realistic experiences in virtual environments but also expand our lexicon of audio AR. As realism is directly correlated with immersion and engagement, understanding perceptual factors in the AR context can lead to better accessibility of experiences.

Understanding plausibility in the AR context is also a first step in defining sound design aesthetics for these kinds of experiences. As without sound realism, the illusion of virtual auditory objects being part of the real environment is broken, it seems absolutely necessary to understand plausibility as the first step towards high-quality sound design. Depending on the application, sound design in AR aims to create not only an engaging audio layer but also a sound art of aesthetic value. That is why this research can contribute to broadening creativity and developing new approaches to sound design and by this – open new avenues for expression.

6 Dissertation Outline

The dissertation begins in Chapter II with a presentation of literature background focused on Virtual Acoustic Environments. The areas discussed include design methods incorporating source, room, and listener (spatial audio cues) simulation. The chapter discusses the technical implementation of different methods as well as their influence on sound perception.

Chapters III and IV include previous relevant research conducted and published by the author and colleagues during the doctoral program at NYU's Music and Audio Research Lab. Chapter III focuses on the topic of sound evaluation in the context of XR. The first part discusses previous literature on methodology of studies on sound plausibility evaluation. In the second part, It presents the background literature on perceptual sound evaluation and attribute elicitation methods. Later, it describes a preliminary experiment aimed toward the investigation of appropriate attributes that comprehensively describe auditory perception in VR and highlight its specific characteristics.

Chapter IV focuses on the sound design for XR environments with a focus on the creation of a highly plausible and immersive audio layer. The first section reviews the factors and principles behind the implementation of audio systems for co-located narrative VR experience. The following part of this chapter illustrates a case-study discussion around the implementation of the audio reproduction system for a short narrative VR art piece.

The core of the dissertation is contained in Chapters V-VIII. It presents a scientific experiment carried out for this dissertation. The main goal of the study is to investigate the perception of plausibility in Audio Augmented reality environments as previously discussed. Chapter V describes the principles of the experiment including the technical implementation of two rendering methods for AR and the experimental design of the subjective listening tests carried out in two phases. Chapter VI presents a statistical analysis of the data collected during the first phase of the experiment. Chapter VII focuses on the statistical analysis of both phases followed by an objective analysis of acoustic parameters of the auralization methods and measurements. The analysis is then discussed to find the links between subjective ratings and acoustic characteristics of the auralizations. The section is followed by a discussion on perceptual and acoustic factors affecting plausibility.

The last Chapter VIII presents the contributions of the dissertation, discusses the high-level conclusions relevant to sound design for XR environments, and suggests future research directions.

7 Related Articles

The research process leading to this dissertation has resulted in the creation of the following articles:

Sound Evaluation in the Context of XR Environments

- **Gospodarek, M.**, Warusfel, O., Ripollés, P., & Roginska, A. (2022). Methodology for perceptual evaluation of plausibility with self-translation of the listener. In: *Audio Engineering Society Conference: 2022 AES International Conference on Audio for Virtual and Augmented Reality.*, Audio Engineering Society

- **Olko, M.**, Dembeck, D., Wu, Y.-H., Genovese, A. F., & Roginska, A. (2017). Identification of perceived sound quality attributes of 360° audiovisual recordings in VR using a Free Verbalization Method. In: *Proceedings of the 143rd AES Convention.* Audio Engineering Society

- Reardon, G., Calle, J. S., Genovese, A., Zalles, G., **Olko, M.**, Jerez, C., Flanagan, P., & Roginska, A. (2017). Evaluation of Binaural Renderers: A Methodology. In: *Proceedings of the 143rd AES Convention.* Audio Engineering Society

Sound Design for XR Environments

- **Gospodarek, M.**, Genovese, A., Dembeck, D., Brenner, C., Roginska, A., & Perlin, K. (2019). Sound design and reproduction techniques for co-located narrative VR experiences. In: *Proceedings of the 147th AES Convention.* Audio Engineering Society

- Genovese, A., **Gospodarek, M.**, & Roginska, A. (2019). Mixed Realities: a live collaborative musical performance. In: *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio; Ilmenau, Germany,* pp. 159–164

- Gochfeld, D., Brenner, C., Layng, K., Herscher, S., Defanti, C., **Olko, M.**, Shinn, D., Riggs, S., Fernández-Vara, C., & Perlin, K. (2018). Holojam in Wonderland : Immersive Mixed Reality Theater. In: *Leonardo, 51(4),* 362–367.

- Lubetzky, A. V, Kelly, J., Wang, Z., **Gospodarek, M.**, Fu, G., Kuchlewski, E., & Hujsak, B. (2019). Head Mounted Display Application for Contextual Sensory Integration Training: Design, Implementation, Challenges and Patient Outcomes. In: *Proceedings of the 13th International Conference on Virtual Rehabilitation.*

CHAPTER II

VIRTUAL ACOUSTICS ENVIRONMENT

1 Introduction

This chapter provides an overview of methods for designing virtual acoustic environments and explores studies focused on sound perception within such settings.

Every virtual acoustic system comprises three fundamental modules: source modeling, environment modeling, and listener modeling (illustrated in Figure 2 in Chapter I). In the context of AR audio, where the aim is to seamlessly integrate real and virtual elements within the environment, these modules strive to emulate the characteristics of reality, including the real source, the physical space the user occupies, and the individual listener. The closer each module aligns with reality, the more immersive and convincing the user experience becomes. Furthermore, these modeling components should exhibit dynamic responsiveness, meaning they must update in real-time to reflect changes in the sound scene, particularly when sources or the listener are in motion.

The sections below discuss the methods for modeling each of the modules and their influence on the perception of sound.

2 Sound Source Modeling

Source modeling techniques usually entail the generation of a sound and the modeling of its radiation pattern. Sound can be reproduced from pre-recorded material or generated using different sound synthesis techniques. The audio material then needs to be processed to recreate the directivity characteristics of the real source. Dynamic directivity allows obtaining a different spectrum of sound depending on the rotation of the source in relation to the listener. In most implementations, sound sources are modeled as omnidirectional point sources. It is a sufficient method for some applications but to achieve a higher level of realism more detailed rendering is required. Radiation patterns vary in complexity depending on the type of source. For common sources such as voices or loudspeakers, the energy is distributed mostly in the frontal hemisphere and high frequencies are attenuated with the increase of angular distance from the front (Monson et al., 2012). Modeling of the directivity is especially challenging for musical instruments where the pattern is not only frequency dependent but also changes with the intensity and articulation or the performance (fingering for woodwind instruments, presence of a mute for brass instrument, etc.) (Karjalainen et al., 1995).

There are several ways of implementing source directivity (Rindel et al., 2004). Previous studies have suggested the use of multichannel source directivity auralization (Otondo & Rindel, 2005). In this method, the first step is to characterize the source with multichannel anechoic recordings. After that, the source's radiation sphere is divided into angular segments. Each segment is assigned to one microphone position. The resulting RIR is then convolved with the corresponding channel of the multichannel recording. The disadvantage of this method is that it leads to abrupt changes in level and spectrum when the source orientation is modified. To alleviate this problem another solution was proposed (Postma & Katz, 2016). Instead of using multichannel recording, an overlapping beamforming approach to multichannel source decomposition is suggested. In this method, the source's radiation sphere is divided into 12 beam patterns. The beams have minimal overlap while remaining equal gain sum in all directions. The directivity is applied by controlling gains for each of the 12 components.

Most of the current 6DoF audio renderers implement directivity using a simple parametric function (Google, 2018; Valve, 2019). Zero- and first-order directivity patterns (omnidirectional, cardioid, bi-directional, etc.) can be computed by calculation of the weighted sum of the omnidirectional and dipole pattern (Southern & Murphy, 2009). Yet, in this approach, the directivity is applied to all frequencies with the same weights.

More advanced approaches include creating the acoustic radiation pattern in the spherical harmonics domain for each harmonic partial of every played tone. An assessment of the complexity of the acoustic radiation pattern can be conducted based on the number of excitation points utilizing the centering algorithm (Shabtai et al., 2017).

2.1 Influence of Directivity of the Sound Source on Perception

The study by Postma, Demontis, and Katz (2017) included perceptual listening tests where participants evaluated omnidirectional, static, and dynamic directivity applied to the anechoic recording of an actor's performance in theater space. The ratings included: plausibility, distance, apparent source width, and listener envelopment. The results showed that dynamic directivity was rated significantly higher for plausibility, distance, and apparent source width in comparison with static voice directivity and omnidirectional sources. While it seems intuitive that implementation of dynamic directivity should improve plausibility, the question arises if including the subjects' movement in 6DoF will affect sensitivity to source directivity changes. The study by Robotham, Rummukainen, and Habets (2019) focused on answering this question by finding just noticeable difference (JND) for static and dynamic (with subjects' movement) evaluation of source directivity. In the study directivity was independent of frequency thus the change of directivity caused modification of loudness rather than timbre. The results showed that for the static presentation of steady signal – noise, the JND was -0.6dB while in the 6DoF condition, this difference was not audible which suggests that the threshold is higher. For non-steady signals, the JND difference remained at -2.6dB and seemed to not change between static and dynamic conditions.

3 Room modeling

The goal of room modeling in virtual acoustics systems is to simulate the sound propagation behavior in acoustic space. Sound wave initialized by the source interacts with the space around. When it arrives at the boundary, it reflects and diffracts. As a result, the sound wave does not arrive at the listener at one time but instead, it bounces off the obstacles creating reflections. At first, the reflections come to the receiver at distinct time intervals but slowly their energy decreases while the echo density increases, creating a diffuse reverberation. That is why the RIR is usually divided into three sections: direct sound, early reflections, and late reverberation (see Figure

4). The reflections amount and type depend on the room's shape, and the dispersion and absorption coefficients of the surfaces' materials. The direct sound segment of the RIR is usually modeled separately to account for the different positions of the source and receiver. The late reverberation is generally considered as diffused so the individual reflections are not differentiated and are uniformly distributed around the listener. That is why the late reverberation time-frequency envelope is considered independent from the position in the room (Barron & Lee, 1988). The early reflections part of the RIR is generally the most challenging to simulate. As the spatio-temporal pattern of the early reflections is perceptually salient, the timing and the direction of arrival for each reflection need to be properly rendered. Besides that, the pattern of the reflections changes depending on the position in the room of the source and receiver. Very accurate rendering of early reflections requires very detailed measurements or high-cost calculations and usually is not practically feasible. That is why different ways of approximation are investigated in order to simplify the rendering without losing perceived spatialization quality.

There are three basic approaches to modeling the acoustics of the room for virtual environments:

1. based on physical models

2. based on the convolution with the premeasured RIR

3. based on algorithms e.g. delay networks

The sections below provide a quick overview of each of the methods.

3.1 Physical Based Approaches to Room Modeling

The acoustical modeling of sound propagation can be obtained under two main frameworks: wave-based and geometric-based. In wave-based techniques, sound propagation is physically modeled using spatio-temporal approximations. The wave-based modeling is able to provide more accurate results but at the same time, it is much more computationally expensive (Savioja & Svensson, 2015). The geometrical acoustics (GA) aims to compute the various propagation paths based on the initial model of the space, receiver, and source. In GA sound is assumed to propagate as rays and all of the properties of sound waves are neglected. This assumption is valid in high frequencies where the sound wavelength is much smaller than the dimensions of the elementary surfaces of the room. The problem arises in low frequencies where the approximation is not as relevant. Despite that, this technique is widely used in practical implementations of room acoustics as being faster and more efficient (Elorza, 2005).

A number of different approaches exist within the GA framework. One of them is the image method (IM) which replaces the physical boundaries of the environment with an equivalent infinite lattice of image sources (Allen & Berkley, 1979). It aims to find purely specular reflection paths between the source and receiver (see Figure 3). The main assumption is that all of the boundaries are perfectly flat and rigid. Diffraction and interference of sound waves are neglected. The main advantage of the method is its relatively simple implementation and sufficient quality, especially for the early reflections segment of the IR. The ray-tracing method is an alternative algorithmic implementation. The method tracks the rays emitted by the sound source as they reflect off the surfaces of a modeled space. Ray-tracing method is more efficient than IM although it creates geometrical-paths approximations that can cause delay errors. On the other hand, acoustic radiosity is a method that may be used to account for sound scattering on non-perfectly flat and rough surfaces. The method assumes perfectly diffuse reflection from all surfaces of the enclosure(Miles, 1984; Nosal et al., 2004).

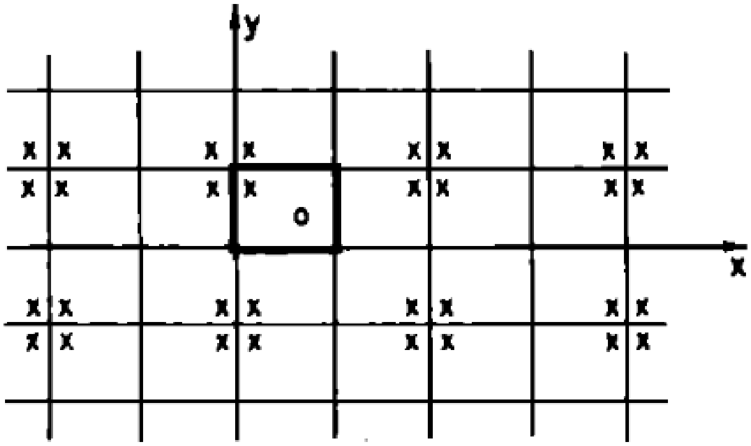


Figure 3: Original source and image sources from Allen & Berkley (1979)

3.2 Convolution-based Approaches to Room Modeling

The convolution-based methods aim to reconstruct the acoustics of the space from premeasured RIRs. The types of RIRs used previously for acoustic rendering in virtual systems include omnidirectional RIRs (Pörschmann & Wiefling, 2015), BRIRs (Pörschmann et al., 2017; Neidhardt & Knoop, 2017; Neidhardt et al., 2018; Werner et al., 2021) and SRIRs (Nowak & Klockgether, 2017; Engel et al., 2019). With this technique, the reference IR is usually divided into two or three segments: direct sound and reverberation or direct sound, early reflections, and late reverberation (see Figure 4).

The direct sound is simulated separately to ensure the proper sound positioning by using convolution with HRIRs. The reverberation is usually left unchanged from the reference IR. As the spatio-temporal structure of the early reflections segment is perceptually relevant and most difficult to simulate, different methods of simplification are used for this segment (e.g. keeping the temporal structure intact and using convolution with HRIRs to obtain one spatial pattern of the reflections independent from the position in the room (Pörschmann & Wiefling, 2015)).

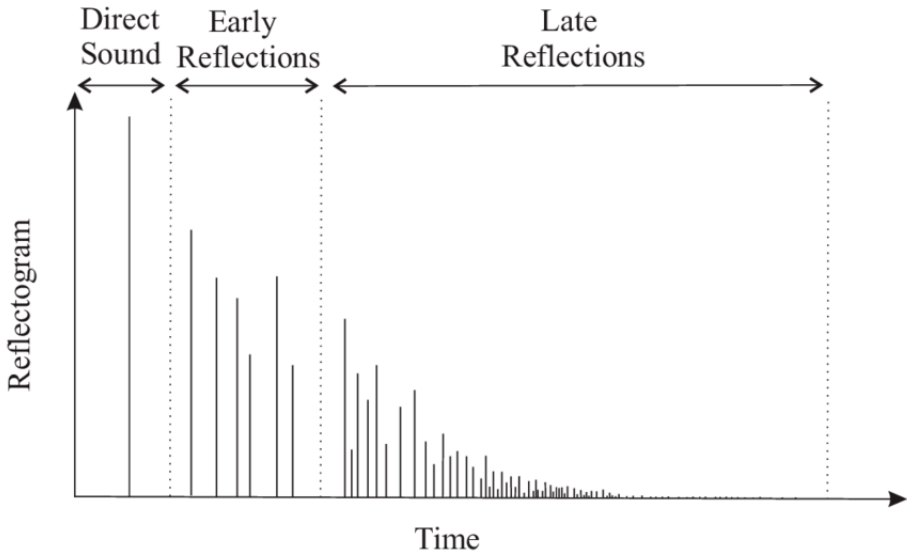


Figure 4: Impulse response divided into 3 segments from Yilmaz, (2010)

3.3 Algorithmic Approaches to Room Modeling

In the typical rooms, the reflections build up until the *mixing time* establishing the diffuse reverberation. Late reverberation of the room is a diffuse sound field independent from the source and listener positions. Because the individual reflections are no longer noticeable the reverberation can be approximated by a system of delay lines.

FDNs are designed with parallel delay lines connected recursively through a feedback matrix. The state-of-the-art FDN implementation was proposed by Jot and Chaigne (1991). The authors suggested using a set of multiband absorptive filters connected with delay lines in order to control the frequency-dependent reverberation time. The advantages of FDN algorithmic reverb are the simple design, low computational complexity, and high quality of reverberation which can be easily tuned to the real room reverberation characteristics. Although there are no new standards established, new methods are proposed that aim to improve the FDNs (Lee et al., 2012; De Sena et al., 2015).

3.4 Acoustic Modeling and Perception

One of the aims of acoustic research is to investigate the interaction between objective and subjective spatial characteristics of sound. Based on that it is possible to draw conclusions about the influence of different objective parameters of rendering on the perception of virtual acoustics. Research in this area is very extensive. The section below gives a quick overview of the most important aspects.

Direct sound allows localizing the source in a space. Specular reflections that arrive 5-10ms after the direct sound can create a perceived image shift and increase the apparent source width (Olive & Toole, 1989). These first reflections are not perceived as a separate event (Wallach et al., 1973) and as a consequence, they modify the timbre of the direct sound due to comb-filtering (Bech, 1995). The time delay between the direct sound and the first perceptually distinct reflection influences the perception of space dimensions and presence. Besides that, the temporal pattern affects the perceived size of the environment (Kaplanis et al., 2014). The spatial, temporal, and spectral pattern of early reflections modifies the envelopment of the space (Barron & Marshall, 1981). However, with the increase in the density of the reflections, perception is influenced more by the statistical parameters of the late reverberation. In large rooms, the perception of room size is governed more by the reverberation time but in the small rooms by the pattern of early reflections (Yadav et al., 2013). The reverberation improves the externalization of sound sources even in situations when only the early reflections are implemented (Begault et al., 2001). All of these findings suggest that early reflections pattern and statistical properties of the reverberation tail have an influence on the perception of acoustic space.

The quality of the simulated acoustic environment might be especially crucial for an AR system in which the user can see the real surroundings. The visual cues create a particular set of expectations. Lindau's definition of plausibility states that plausibility depends on how much a simulation meets the expectations of the listener. Previous research proved that congruency between visual and auditory stimuli (in this case acoustic space) influences significantly the plausibility (Werner et al., 2016). Results of the experiment by Neidhardt suggest that simulations without any reverberation were degrading the plausibility of sound during the subject's walk (Neidhardt et al., 2018). Different studies focused on investigating the accuracy of the spatial rendering which is necessary to achieve a high level of realism and immersion. Picinali, Wallin, Levtov, and Poirier-Quinot (2017) compared different techniques for the implementation of reverberation. Each of them presented a different level of spatial accuracy. The study was focused on the influence of reverberation accuracy on localization and realism. The methods used included multichannel Ambisonics-based, stereo, and mono reverberation methods. The results showed that there was no perceived difference between four of five reverberation techniques which suggests that the high complexity of reverberation does not always correspond with the improvement of perceptual attributes. Similar results were obtained by Engel et al. (2019). The research aimed to investigate the trade-off between the complexity of reverberation and simulation quality. The reverberation was based on SRIR of zeroth to the fourth order of Higher Order Ambisonics (HOA). Results suggested that the order of the HOA reverberation soundfield did not have any influence on the perceived realism of the simulation.

4 Listener Modeling with Head-Related Transfer Functions

In section 2 the importance of the source directivity modeling has been emphasized, i.e. the way the source radiates the acoustic energy in space. Conversely, each receiver has its own directivity characteristics. An omnidirectional microphone will respond equally to a sound wave coming from any direction, whereas a cardioid microphone will favor its frontal direction. Similarly, our head presents a very specific directivity function. It originates from the scattering of sound waves on our shoulders, head, and ear pinnae and has to be precisely modeled in order to provide convincing spatial cues to our perception. Head-related Transfer Function (HRTF) characterizes the auditory spatial cues of a person for a defined sound source position. It includes interaural time difference (ITD), interaural intensity difference (IID), and the difference in the signal spectrum which is introduced by the reflections of the pinna, head, shoulders, and torso of a listener. Obtaining a complete set of HRTFs of a person requires a long measurement procedure. Microphones are inserted into ear canals to capture the measurement signal from sound locations around the head. Much of the research focuses on different ways of acquiring HRTF measurements in a fast way, but at present, there is no widely accessible solution for obtaining high-quality HRTFs efficiently (Guezenoc & Séguier, 2018).

4.1 Influence of Individualized HRTF on Perception

Because of the constraints described above, most of the binaural rendering software uses generic HRTFs measured from a model of the human head. Yet, the differences between the actual HRTFs of a person and the generic HRTFs can cause different kinds of distortions of spatial auditory perception. These include front-back confusions, problems with externalization, and localization on the vertical plane (Begault et al., 2001; F. L. Wightman & Kistler, 2018). The distortions not only affect the spatial attributes of sound but can also influence the sense of presence (Väljamäe et al., 2004). However, previous studies have shown that there are ways to improve the perception of sound with non-individualized HRTFs: sound localization with non-individualized HRTFs can be greatly improved by training with correct answer feedback (Andéol et al., 2014; Mendonça et al., 2012). Other studies have found that applying head-movement can decrease substantially problems with front-back reversals (Wallach, 1940; Blauert & Butler, 1985) as well as externalization Wersényi (2009). The implementation of a tracking system along with the user's movement should alleviate the front-back confusion and improve externalization.

Previous studies have indicated that generic HRTF may be sufficient when paired with the temporally and spatially matched visual cues (Berger et al., 2018) for VR audio reproduction, as well as AR audio reproduction (Werner et al., 2021).

5 Other Factors in Designing Virtual Acoustics System

There are several other factors that should be taken into account when designing a virtual acoustics environment. The sections below describe their implementation and perceptual importance.

5.1 Latency and Its Perceptual Salience

Latency in virtual displays defines the amount of time that passes between the action of the user (i.e. body movements, head rotation) and the corresponding visual and auditory feedback. In the case of the audio layer, the latency defines the time between the listener's head movement and the corresponding change in the spatial audio reaching the listener's ears. It is constituted by the time needed for updating the position data, sending it to the receiver (usually the computer that handles the calculations), rendering audio based on the new position, and sending the audio to the playback system. If the delay exceeds the detectable threshold it can cause multisensory distortions coming from the disparity between the perceived sound image and the expectations from proprioception (the sense of our own movement). They can affect user response time (D. S. Brungart et al., 2004), and sound localization (Sandvad, 1996), thus influencing plausibility. D. S. Brungart, Simpson, and Kordik (2005) investigated the detection threshold for the head tracker latency and found out that the delay below 80 ms in the auditory scene with one virtual source is not reliably detectable by an average listener. However, their result showed also that this threshold is reduced by approx. 25 ms when there is a low-latency reference signal provided at the same location as the virtual source. This situation occurs in AR environments where the listener is exposed to real sounds that do not have any delay. We can assume that the threshold will be different for VR and AR systems. It has to be noted that there was a significant disparity between the subject ratings and some of the subjects were able to detect threshold values as low as 30 ms. Authors suggest that a threshold value below 30 ms might be needed to make sure that it is not detectable for all of the listeners and different listening conditions. On the other hand, Yairi, Iwaya, and Suzuki (2007) conducted two experiments that aimed to investigate the detection threshold and JND of the system latency. Subjects were asked to rotate once and judge if they were able to detect the delay. The values for detection threshold and JND averaged over nine subjects were 45 ms and 59 ms respectively.

5.2 3DoF vs 6DoF Environments

The freedom of the user's movement in three-dimensional space is common for AR and VR systems. In 3 Degrees of Freedom (3DoF) environments, the rotation of the user's head is tracked whereas in 6 Degrees of Freedom (6DoF) environments, the rotation and position of the user are tracked. Because of that, all of the elements of the audio rendering system need to be implemented in real-time and modulated according to the change of rotation and location of the user and the source. The perception of the auditory layer is different with the ability to self-translation and requires highly responsive rendering systems (Neidhardt et al., 2018).

5.3 Audio-Visual Interaction

Even though in an audio-only AR experience no visual virtual objects are presented to the subjects, participants still can see their actual surroundings thus their judgments are affected by visual cues. Perceptual phenomena like *colavita effect* where visual stimuli change the perceived location of a simultaneously presented sound indicate that the influence of visual cues on the perception of sound may be crucial (Colavita, 1974). The importance of visuo-auditory integration mechanisms can be assumed to be especially significant in AR environments where the virtual sources blend with the real environment (Begault, 1999). The influence of visual cues on audio perception in AR systems may include:

- the physical space around the listener creates a certain type of expectations for the acoustic rendering, i.e., the type of reflections expected thus affecting the plausibility judgment (Larsson et al., 2001; Thery et al., 2017; Postma et al., 2017; Bailey & Fazenda, 2018; V. Martin et al., 2022)

- the presence of the visual object representing the sound source may affect the overall judgment of plausibility of the scene (Bailey & Fazenda, 2018) and possibly improve localization which might be distorted when using generic HRTFs (Colavita, 1974)

- the visual cues may affect the perception of other attributes of sound i.a. acoustic distance or loudness (D. Brungart, 1998; Postma & Katz, 2017)

CHAPTER III

PERCEPTUAL EVALUATION OF SOUND ATTRIBUTES IN XR CONTEXT

The initial section of this chapter offers a review of the literature related to sound plausibility assessment. It begins by clarifying the distinction between two principal measures of experience quality: plausibility and presence. Subsequently, it describes prior studies on plausibility evaluation, categorized into research conducted in 3DoF and 6DoF, with particular emphasis on experimental methodologies employed.

The subsequent segment of this chapter presents a previously published study focused on exploring sound attributes of significance in the VR context. The conclusions of the study relate to the question of the correlation between plausibility and other sound attributes. Before exploring this correlation, it is essential to investigate which sound attributes are critical for sound perception within XR contexts. The insights obtained from this research informed the experimental design of the dissertation study.

1 Methods for the evaluation of audio in AR and VR - plausibility and presence

One of the essential measures of the experience quality commonly used in literature is *presence* which relates to the feeling of "being there" in a virtual environment (Wagner et al., 2009). The sense of presence involves several components: the sense that we are located in and act from within the virtual environment (VE), and the sense that we are concentrating on the VE and ignoring the real environment. On the other hand, immersion describes the extent to which the computer displays are capable of delivering an illusion of reality to the senses of a human participant (Slater & Wilbur, 1997). Important parameters of immersion include the extent of the field of view, the number of sensory modalities that the system simulates, the quality of rendering in each sensory modality, the extent of tracking, the realism of the displayed images, the frame-rate, and the latency. The concept of 'presence' refers to the mental state in which a user behaves and feels as if physically present within the computer-mediated environment (Draper et al., 1998). The sense of presence is a subjective experience and only quantifiable by the user experiencing it. Immersion and spatial aspects contribute to the sense of presence, but it is the quality of interaction which plays a major role in eliciting presence.

Slater divides the sense of presence into two concepts: place illusion and plausibility. Place illusion relates to the feeling of actually being in a virtual place while plausibility relates to the impression that what is happening is real. In the case of AR, place illusion is less important as the surroundings are already real. On the other hand, the virtual elements overlaid on this environment should seem to be really happening which can be evaluated by plausibility (Slater, 2009). Lindau suggests another definition of plausibility as "a simulation in agreement with the listener's expectation towards a corresponding real event" (Lindau & Weinzierl, 2012) which also applies to AR. It is contrasted with authenticity which requires an immediate comparison with a real event and consequently, needs a very high level of accuracy which may not be necessary for real-life scenarios. The authenticity is associated with an *external* reference to which the listener is immediately comparing the virtual event. On the other hand, when evaluating plausibility, the listener is relating to the *inner* reference which is caused by the memory of previous experiences. Moreover, setting the listener's expectation as the measure corresponds well with the majority of AR system applications. Real surroundings on which the virtual objects are imposed, create a shared set of expectations toward the auditory experience of the scene. Taking all of that into account, plausibility seems to be an appropriate measure of the overall user experience in audio AR. The sections below present existing research on plausibility in VR/AR environments. It focuses on the methodology used in the studies.

1.1 3DoF Environments

This section presents the methodology of plausibility studies on sound with a dynamic simulation where the rotation of the listener's head was tracked. The study by Lindau and Weinzierl (2012) proposed a Yes/No paradigm to investigate the plausibility of binaural rendering of sound. The stimuli was presented by speakers or through binaural rendering with BRIRs recorded with dummy head in the experimental space. After listening to each stimulus, subjects were choosing if the stimulus was real or not. Signal detection theory was employed to analyze the data. The method proved very efficient for testing plausibility of high quality rendering systems.

The same methodology was implemented in the study by Pike, Melchior, and Tew (2014). The aim was to repeat Lindau's experiment in smaller space which is usually considered more challenging to simulate. The short training session was introduced. The results showed that there was a small but significant difference between ratings of real and virtual speakers. Yet, subjects indicated that it was challenging to identify the simulation and most of the time they were guessing.

Bailey and Fazenda (2018) employed the similar experimental design in the VR context. This time, subjects were presented with 3D models of the room they were seated in on VR headsets. The stimuli were played through real speakers or through headphones. Two types of simulation were employed: convolution with SRIR and image-source acoustic modeling. Both simulations yield relatively low ratings of the plausibility.

Study by Engel et al. (2019) aimed to test the influence of rendering accuracy on realism of the simulated scenes. Subjects were not seated in the simulated room. Instead, they were shown an image of the rendered space and position of the sound source. After listening to the pair of different stimuli they answered question: "Considering the given scene, which example is more appropriate?" Results indicated that there was no influence of rendering accuracy on the plausibility of simulated sounds.

1.2 6DoF Environments

Most of the studies evaluating the plausibility of different rendering parameters were performed with a static listener. Perceptual investigation of audio attributes with self-translation in AR systems is not an easy task. The experience is very multimodal, and the way the subjects move determines the sound they are evaluating. There are several methods proposed in previous research for these kinds of tests.

Bergstrom, Azevedo, Papiotis, Saldanha, and Slater (2017) evaluated the plausibility of a VR experience with a string quartet using "color matching" theory. The subjects were first exposed to the scene using the most realistic settings of all of the rendering variables and then tasked to match all of the parameters to the plausibility level of the first exposure starting with the lowest level. The rendering parameters included: gaze (if the virtual musicians were reacting to the audience with the direction of looking), sound spatialization (if the sound was spatialized), reverberation, environmental sounds (if the subjects could hear ambiance from a virtual window). Results showed that subjects were first adjusting the environmental sounds, then gaze parameter, reverberation, and the spatialization as the last. The method seems to be less efficient for a larger amount of stimuli.

Neidhardt and Knoop (2017) tested the plausibility of different audio scenes using the HTC Vive system. Participants could see only the walking path and edges of the tracked area in the HMD. The stimuli were processed using measured BRIRs and acoustic simulation using a simple "shoe-box" model with two reverberation times, one similar to the actual room and another one with a longer RT60. In the first part of the experiment, participants were rating plausibility on a scale of 0-100. Results did not bring any conclusion as the ratings were not normally distributed and it was not possible to find any trend. It is probably due to the lack of any reference to which subjects could adjust their evaluation. What is important is that the rendering was covering only 180° in front of the listener and when rotating outside this angle, the illusion was completely broken. In the second part of the experiment, the authors decided to take another approach to the evaluation of plausibility using modified *Yes/No paradigm*. This time, subjects were answering two questions:

- Did you get the impression of walking towards/past a sound source?

- Would you call this experience a plausible illusion of a sound source?

This method brought much more conclusive results. Participants gave higher ratings to the scenes which used measured BRIRs than to the simulation. Longer reverberation seemed to reduce the drawbacks of virtual acoustics. Also, subjects were asked to give reasons for their ratings which gave insight into attributes related to the evaluation of plausibility.

Neidhardt et al. (2018) tested the plausibility of sound synthesis in a 6DOF environment with different sets of BRIRs convolved with the stimuli. Subjects were tasked to physically walk towards the speaker and then answered several questions about plausibility, externalization, continuity of the sound change, and impression of walking towards the source. Results showed that the lack of reverberation influenced the most ratings of plausibility. On the other hand, a simplified version of reverberation (taken from only one point in the room) did not change the plausibility significantly. Besides that, plausibility was correlated with externalization, continuity, and impression of walking towards the speaker which may indicate that this method is an appropriate way of measuring the overall quality of the audio AR experience.

Wirler, Meyer-Kahlen, and Schlecht (2020) conducted an evaluation of mixed reality audio scenes with a new approach *transfer-plausibility* which stands in between authenticity and plausibility. Subjects instead of comparing the virtual stimuli to the same real sound have to detect the simulated sound alongside multiple real sources playing simultaneously. The sounds were simulated using non-individualized dynamic binaural rendering with varying scene complexity controlled by the number of speakers from 1 to 8. Subjects were seated in the center of the speakers. Results showed that this methodology allowed to omit ceiling effect which happens when the quality of rendering is not perfect. High complexity yielded low detection rates even for low-quality virtual stimuli. The detection rate was highest when two sources were playing simultanously. This approach has the advantage of being similar to real-life scenarios where the simulated sound is always presented alongside a real audio environment. The drawback of this approach is that it does not conclude about the quality of the rendering when the source is presented individually (not being shadowed by other sources).

Werner et al. (2021) presented an extended study on BRIR synthesis for a moving listener in the AR context. They conducted several listening tests to evaluate the system in terms of plausibility and other spatial attributes. The participants had to walk a given path and listen to different types of stimuli. After walking they rated several audio parameters on the scale including externalization, ability to localize the auditory event, the stability of the position of the reproduced audio object, the coloration during movement, and the overall impression.

In conclusion, most of the previous studies on plausibility in the 6DoF environment tasked participants with walking along predetermined paths and answering a short questionnaire about the perception of sound. The spatial attributes evaluated included plausibility, localization accuracy, externalization, continuity, stability, overall impression. Most of the studies did not use real references in the test. As the plausibility relates to the inner reference and listener expectations, it is not necessary to introduce the real sounds along with the simulations. Yet, in the real-life use cases of AR systems, the real sounds will almost always be present in the sound environment. Thus, it seems to be beneficial to include a real reference in the experimental design of studies on the plausibility of sound in AR. The next chapter will describe the methodology of the study which is designed based on the conclusions from the literature review.

2 Identification of Perceived Sound Quality Attributes in VR

This section presents the article titled ”Identification of Perceived Sound Quality Attributes of 360° Audiovisual Recordings in VR Using a Free Verbalization Method” written by Marta Gospodarek (previously Olko), Dennis Dembeck, Yun-Han Wu, Andrea Genovese, and Agnieszka Roginska, published in the Proceedings of the 143rd AES Convention (2017). It is reproduced with the permission of the Audio Engineering Society.

2.1 Introduction

The need for spatial audio reproduction in novel contexts like VR applications or 360° degree video has been growing along with the recent developments in the gaming and multimedia industry. Delivering a truly immersive experience in VR systems requires high visual quality, intuitive user interaction, and authenticity of the perceived sound. New tools for 360° audio recording, post-production, rendering and playback in VR are facilitating the production pipeline available for artists, engineers, and customers. To appropriately evaluate and compare the quality of different VR audio productions, comprehensive subjective assessment tests need to be employed.

Compared to static spatial audio experiences, such as binaural audio and surround sound systems, sound for head-tracked 360° experiences (as in VR) involves a different order of perceptual dimensions related to the possibility of shifting the point of listening perspective. The experience of sound in 360° is closer to a natural way of listening; thus, the list of factors that influence naturalness of the auditory sensation is assumed to be larger than in common playback systems. The conceptual differences between static channel-based audio and dynamic object audio may significantly influence how listeners evaluate the sound quality of traditional multichannel sound compared to the upcoming 360° audio formats. As a result, it may not be appropriate or sufficient to employ the same evaluation attributes used to rate static spatial experiences when judging dynamic audio presentations.

This paper illustrates a preliminary experiment aimed toward the investigation of appropriate attributes that comprehensively describe auditory perception in VR and are able to highlight its specific characteristics. Specifically, the focus is to study subjects' verbal elicitations and identifications of relevant auditory attributes within a dynamic binaural audio reproduction of a 3-degrees-of-freedom VR system. Discovered attributes can facilitate the future creation of judgment scales and assessment methods. Results and methods are compared with previous literature concerning the elicitation of sound attributes.

2.2 Elicitation Methods for Sound Quality Evaluation

In usual perceptual studies, before asking listeners to evaluate the spatial features of an audio signal, the attributes of sound quality need to be defined first by an experimenter. When a field becomes increasingly established, there is a higher possibility for the attributes to be validated, well-developed, and accurate in describing certain features.

The experience gained from conducting experiments provides information to improve and refine the scales used, while listeners can sometimes be trained to focus on desired attributes of a given stimulus (Bech, 1992). Unlike some well-established fields that are more consistent with their terminology, the words, and concepts used to describe sound are more likely to vary from individual to individual ((Shaw & Gaines, 1989)). As a result, differences between verbal constructs provided by an experimenter and elicited constructs provided directly by the subjects may occur, especially with non-trained subjects who account for the majority of the population.

In several instances of studies on reproduced sound quality evaluation, subjects are asked to rate relatively vague pre-defined terms (Toole, 1985; Woszczyk et al., 1995; Rumsey, 1998). The major problem with provided attribute scales is that the subject is limited to responding in the ways predefined by the experimenter. In addition, some listeners might not be able to accurately map and connect their complex auditory perception using separable attributes. It is also hard for researchers to clarify which exact isolated attribute they want the listener to rate unless they provide extreme stimuli as an example. In the paper published by Colomes et al. in 2010, (Colomes et al., 2010), the issue of unclear definitions in traditional single-axis test methodologies, such as BS.1116 (ITU-R BS.1116-1, 1997) and MUSHRA (ITU-R BS.1534-1, 2003), is demonstrated. The paper aimed to validate the idea of sound families by comparing the results of a free categorization method and a multidimensional scaling method. The authors concluded that the use of sound families helps to minimize the bias created by the vague definition of sound attributes. Verbal elicitation tasks are designed to minimize the experimenter bias (Kelly, 1991). By encouraging the expression of personal sensations towards the stimulus under evaluation, the differences between the way each subject defines certain attributes can be put into context. In the paper published by Guastavino and Katz in 2004, 26 subjects were presented with live recording materials in 1-D, 2-D (added speakers behind the listener) and 3-D (added speakers at height) configurations and were allowed to describe the perceptual impact of each stimulus freely. A semantical analysis, conducted by the researchers of all the phrasings generated by the free verbalization, served to group synonyms into several semantic themes. This method permits to gather information about how listeners subjectively perceive certain phenomena and describe them as spatial attributes using their own mental and verbal constructs and associations.

2.3 Spatial Attributes in Literature

Over the years, different approaches have been employed to identify the spatial attributes of sound in different reproduction systems. The attributes elicited were then used in subjective tests on the quality of various reproduction systems like surround, stereo headphones, or Wave Field Synthesis.

Although in the past there were several attempts to create a common lexicon of spatial sound attributes, in literature the terms used to describe spatial sound attributes are open to different kinds of interpretation. In general descriptive terms, Berg and Rumsey indicated that spatial attributes stand for “the three-dimensional nature of sound sources and their environments”. In order to satisfy two of the important requirements for psychological research, validity (“the test measures what it claims to measure”) and reliability (“the repeatability of the measurement”), previous literature should be put in relevant context when making decisions on which spatial attribute to apply for rating a given setting.

In practice, the choice and definition of relevant attributes for judging spatial perception present a certain degree of variance according to the system being tested. In the paper written by Zacharov and Koivuniemi, source width and spatial impression are said to be the two spatial terms that repeatedly appeared in several spatial quality evaluation experiments done on mono, stereo, 5-channel, and periphonic speaker systems. However, sometimes they were brought up in slightly different forms (Berg, 2002; Mason & Rumsey, 2000). In another paper published in 2010, Kamekawa and Marui pointed out that the typical spatial attributes used in some of the multichannel surround sound system evaluation are localization (the seeming location of the sound sources), depth (the seeming spatial distance between the listener and the sound source), width (the width of the whole sound image), envelopment (the surround feeling from the side of and behind the listener) and presence (the feeling of "being there"). In the case of a stereo headphone system, Lorho (2005) indicated that five clusters of sound attributes were found after examining the dissimilarity between individual attributes elicited by subjects. The first category consists of spatial-related attributes such as width, reverb, and room size. The second cluster contains attributes concerning the timbral aspect of sound, e.g. clarity, brightness, and treble. The third cluster includes attributes related to various kinds of perceptual experiences, with three occurrences of the term noise. Moving on, the low-frequency emphasis is the core concept of the fourth cluster, which includes nine occurrences of the attribute bass. Finally, the fifth cluster is relatively close to the previous category and contains attributes of different sound natures. In another paper, based on auditory virtual environment playback system, Silzle stated that sound attributes elicited by listeners, which can also be called quality features, corresponded to quality elements on the service provider side. In addition, the evaluation results on quality features represent the quality of the listener's experience.

Differently, well-established standards for sound quality evaluation, such as IEC 60268 and EBU 562-3, defined three spatial attributes for sound quality evaluation. These are spaciousness (closed vs spacious), distance (distant vs near) and location of sources (unstable vs stable). Later versions of this standard also suggested three factors relating to spatial attributes: 1) image localization, which stands for how well-defined the spatial location of the reproduced sound sources is; 2) image stability, which depends on several factors - including pitch and loudness - and is also a function of the listener's position and head movement; 3) width homogeneity, which indicates if the stereophonic image is distributed uniformly between loudspeakers.

Previous research on elicitation of spatial sound attributes was performed using surround, binaural reproduction systems or virtual acoustic environments. This paper describes an experiment that is the first attempt to elicit attributes of spatial sound in the 360° audio format played back binaurally with head-tracking. The 360° format introduces new dimensions to the perception of the sound. The listener is provided with a full sphere in which object audio elements can be positioned and then delivered through speaker matrixes or binaurally through headphones. The signal delivered is commonly reproduced either within a spherical sound-field representation (Ambisonic) or as a speaker-independent sound object (Object-based audio). That is to say, any direction around the listener should be treated equally within an experimental investigation, as opposed to traditional multichannel surround sound which is tied to discrete channel outputs and possesses the concept of a main “front” image (Horsburgh et al., 2011).

2.4 Techniques Used for Audio Production in VR Application

Currently, there are two major flexible audio representations used for VR application — sound-field representations, also known as scene-based, and object-based representations. Susal, Krauss, Tsingos, and Altman described sound-field representations as “physically-based approaches that encode the incident wavefront at the listener location”. Ambisonics is the common method for representing all the wavefronts in the spherical space around the listener (Furness, 1990). In fact, it is relatively more similar to traditional channel-based techniques compared to object-based representations, since the spatial information is directly encoded in the audio signal rather than stored as separated metadata. Scene-based audio is ideal for VR applications because of a more convenient process for acoustic capture, offline content creation, and post-production (Shivappa et al., 2016). An ambisonic microphone is a tool that provides the ease of direct capturing of a spherical sound-field surrounding. New hybrid software tools combine the two capturing philosophies and allow artists and producers to design ambisonic scenes by encoding signals captured with spot microphones into ambisonic sound-fields. Those possibilities introduce new dimensions of modification of the sound scene and, as a result, might introduce new aspects of the perception of the sound quality.

2.5 Experiment

The purpose of this experiment was to extract a vocabulary of auditory differences and similarities in the stimuli presented to the subjects. Subjects composed their own attributes that were later gathered and reviewed by the researchers. In a previous study of related research, Berg and Rumsey generated spatial attributes by asking subjects to describe how one out of three stimuli was different from the other two, and how those two stimuli are similar to each other. Each subject was allowed to listen to every stimulus as many times as they wanted. The process was repeated until no more new attributes could be generated.

There are two major advantages of the triadic method. First, it prevents the researchers from asking the subjects for opposite expression directly. In other words, this method aims to guide the subjects to come up with phrases opposite in meaning naturally, by instructing them to describe the similarities and differences between the three stimuli (Choisel & Wickelmaier, 2006). However, an obvious disadvantage of grouping stimuli in triads is that the relatively small differences between two of the stimuli will be neglected if they are always presented with a distinct counterpart. Therefore, an alternative method of comparing the stimuli in pairs, which allows subjects to focus on small differences, is suggested.

An elicitation process was conducted where subjects generated their own bipolar constructs based on a triad of A/B pair comparisons of the recorded stimuli. In order to analyze this data, the verbal descriptors were grouped together in categories based on the Verbal Protocol Analysis and the semantical analysis. These groupings were then inspected for repeated or common verbal attributes used to identify the stimuli. Finding these common attributes was the desired goal of this study.

2.5.1 Subjects

Eighteen subjects with normal hearing, aged between 23 and 42 with a median age of 25, participated in the experiment. All subjects were expert listeners and students of New York University's Music Technology program. All of them listen to music actively several times a week. 11 subjects were native English speakers, 7 subjects were fluent in written and oral English as their second language.

2.5.2 Stimuli Generation

Four individual musical performances were prepared for playback on a Samsung S7 smartphone and GearVR device. There were three versions/mixes of each video, with each version composed of a different audio mix while using the same visual. Each subject was presented with two out of the four video stimuli chosen by randomization. The stimuli were presented in three separate pairs to elicit differences and similarities between each version. Stimuli generation for the subjects to reflect upon was divided into three separate stages: recording, mixing, and encoding.

Recording The recording process took place in the Dolan Studio at New York University. The 360° visuals were captured using a Giroptic 360° camera. The audio was recorded using both soundfield and object-based capturing techniques. To capture the soundfield recordings, the Sennheiser AMBEO VR microphone was used for all of the stimuli recordings, except for the percussion trio recording. In this case, the double MSZ technique was used (see (Geluso, 2012). All soundfield devices were placed in the center of the room, surrounded by the performance ensembles. The 360° camera was also positioned in the perspective of the soundfield recording devices. Various spot microphones (object-audio elements later encoded in Ambisonics by the renderer) were placed on individual musicians to capture the performance from a close perspective.

Mixing The three audio mixes for each video stimulus was rendered in ProTools HD using the Facebook Spatial Workstation-OSX v2.0 Beta2 plugin and were as follows:

- soundfield microphone only
- spot microphones and artificial reverb
- soundfield microphone and spot microphones

Two different reverberations were applied to the stimuli audio mixes by randomization. The first one utilized the Facebook Spatial Workstation plugin by activating the "Room" parameter. Through this parameter, room acoustic modeling is available to synthesize artificial reverberation in three-dimensional space with the ability to adjust the reverberation mix level and reflection order. The second reverberation method utilized was a stereo convolution reverberation, which was applied during the encoding stage.

The loudness of each stimulus was measured using the Facebook 360 Loudness meter. All stimuli were normalized to an integrated measurement of -15 LUFS.

Encoding The 360° videos and eight channel spatial audio mixes were rendered and synchronized using the Facebook 360 Spatial Workstation Encoder. In order for the subjects to compare mixes in an A/B format, the three different mixes for each stimulus were rendered in pairs (ab, bc, ac). Subjects were then able to compare two different mixes within one video file.

2.5.3 Reproduction

The video stimuli were uploaded to the Facebook 360 application and played back on the Samsung GearVR using Sennheiser HD 650 headphones. The Facebook 360 application allowed for 360° visual playback and auditory binaural rendering of the eight-channel encoded mixes. The subjective testing took place in an acoustically treated research lab at New York University. Subjects were equipped with the GearVR while seated in a chair that allowed full 360° rotation. The playback of the video stimuli was streamed from a saved library within the Facebook 360° application. The loudness level of the playback was adjusted on the GearVR by the subjects at the beginning of each test to suit their loudness preferences and kept consistent throughout the experiment session.

2.5.4 Elicitation Process

The goal of the elicitation process was to acquire verbal descriptors from the subjects personal vocabulary. The four video stimuli, each having three different mix versions presented in pairs, were randomly assigned to the subject. The stimuli versions, labeled A, B, C, were then uploaded to the Facebook 360° application on the Samsung Gear VR. Subjects were first allowed to navigate the stimuli to experience all of the given A/B pairs. The duration of each stimulus averaged 30 seconds. Subjects viewed the pairings in order and were allowed to review and repeat the playback of each mix pair as desired. Participants were then instructed to listen for similarities and differences of the auditory experience in each version and subsequently instructed to write down the perceived experiential similarities and differences in their own format.

Once the subjects had finished viewing the video stimuli, they began dissecting verbal descriptors from their own documentation. They were asked to read all of their notes and create bipolar scales from each of the descriptive words they used. Subjects were encouraged to search for the words which are opposite in meaning and the most precise in the description of their perception. This created a list of bipolar constructs that were then gathered and processed by the researcher.

2.6 Analysis and Discussion

2.6.1 Constructs Elicited

The total number of constructs elicited by all subjects was 231. The minimum number of constructs generated by a single subject was 7, while the maximum number of constructs generated by a single subject was 20. The median value of the number of constructs elicited by subjects was 12.5.

2.6.2 Verbal Protocol Analysis

The first step in the analysis of results was to reduce redundancy of the obtained verbal descriptors when the same identical words were used by several subjects. After removing repeated instances of grading scales, 166 bipolar constructs were left.

Verbal Protocol Analysis (VPA), proposed in the paper of (Samoylenko, 1996), was employed in the analysis of results. In that paper, verbal descriptors describing timbre were analyzed on three levels: logical sense, stimulus relatedness, and semantic aspects. A similar analysis was used in this experiment to divide obtained descriptors into more general classes. The third level of analysis, which focuses on the semantical aspects of verbal units, was employed in this study. Verbal descriptors were categorized into attitudinal and descriptive. Attitudinal descriptors express the emotional relation to the sound (emv) and naturalness (ntl). Descriptive constructs were divided into those describing auditory modality only (UMD) or multiple sensory modalities (PMD).

From all of the scales obtained during the experiment, 9% was attitudinal, and 91% was descriptive. Attitudinal descriptors were related to the preference, overall evaluation of the stimuli, and naturalness of the sound. Noticeably, there were several constructs describing naturalness of the sound change during head movement. From the descriptive features, 82% were unimodal and 18% were polymodal. Unimodal verbal descriptors were describing characteristics of auditory modality only. These constructs, which were a majority of all the obtained descriptors, were related to the general perception of the sound in the 360° scene.

It should be noted that grouping of the descriptors is a difficult task. Categorization based on semantical analysis is largely biased by the interpretation of the researcher. In order to reduce the bias, the categorization of the descriptors was conducted by researchers and a panel of experts. A panel of five experts, including some of the authors, was formed to read each of the scales carefully and to group them based on similar words usage, meaning, and comments of the subjects. The created

Table 1

Attributes elicited during experiment describing sound in relation to the head movement

Attribute	**Scale**
Change of sound during head movement	How noticeable is the horizontal and frontal change in response to head movement
Sound balance during head movement	The signal is attenuated/not attenuated during head movement The amplitude change during head movement is/is not expected
Localization during head movement	Sound sources are easy/hard to localize during head movement Localization seems correct/incorrect during head movement
Width during head-movement	Width of the sound image is steady/changing during head movement
Depth during head movement	Depth or distance of the sources from the listener is steady/changing during head movement
Externalization during head movement	The changes in sound during head movement are happening inside/outside of the head
Clarity during head movement	Sound sources are present/absent when turning head toward the source Sound sources are focused/unfocused when turning head toward the source

groups of attributes were compared with the attribute definitions from previous studies effectuated on the spatial sound.

During the test, subjects were encouraged to comment on each of the scales to allow more precise interpretation of them. The attributes that defined the grouping of the descriptors during the analysis were as follows (the reference source for each attribute is shown in brackets): *Clarity* (EBU 3286–E, 1997), *Externalization* (Durlach et al., 1992b), *Spatial impression* (EBU 3286–E, 1997), *Depth perspective* (Kamekawa & Marui, 2010), *Timbre* (EBU 3286–E, 1997), *Sound image width* (Kamekawa & Marui, 2010), *Location accuracy* (EBU 3286–E, 1997), *Sound balance* (EBU 3286–E, 1997), *Punch* (Fenton & Wakefield, 2012), *Immersion/Presence* (Guastavino & Katz, 2004), and *Freedom from noise* (EBU 3286–E, 1997). The rate of appearance of the verbal descriptors assigned to each attribute is shown in Figure 5. There were no differences in the distribution of verbal descriptors elicited between native and non-native English speaker subjects.

Two categories of verbal descriptors related to polymodal sensations were found: audio-video congruency and perception of sound during head movement. Figure 6 shows the number of unimodal and polymodal descriptors elicited by subjects. The number of polymodal descriptors is relatively small in comparison to unimodal. Audio-video congruency was described by subjects in four different aspects: sense of space (if the sense of space in sound was matching the space in the image), localization (if the localization of the sound sources was matching the image), distance (if the distance of the sound sources from listener was matching the video), and time synchronization between sound and image.

The other category of polymodal descriptors was related to the sound change during head movement. This category relates to the initial motivations behind the paper, to find new descriptive attributes for subjective perception of dynamic audio/video experiences in VR. The groups of scales identified during the analysis are reported in Table 1.

Verbal descriptors indicate that changes in the sound during head movements are perceived separately to the overall sound impression and might be a crucial element in the evaluation of the quality of sound in 360°. The results of the experiment are not robust enough to provide definitions to the new attributes with clear confidence. More research is required to validate the perception of sound during head movement.

2.7 Conclusion and Future Work

This preliminary study was the first attempt to investigate sound quality attributes in 360°. Verbal descriptors elicited by subjects and analyzed using the Verbal Protocol Analysis, and were divided into three main groups: attributes of sound quality describing the general impression of the sound environment, attributes describing sound in relation to the head movement, and attributes describing audio and video congruency.

Verbal descriptors identifying attributes of sound quality, relating to the general impression of the sound environment, were found to be the same as in the similar research on static spatial sound reproduction. Head-tracking allowed listeners to compare the change of sound from different positional perspectives. As a results, inconsistencies between head perspectives were noted by subjects. The study highlighted a number of verbal descriptors, describing the relation between sound and

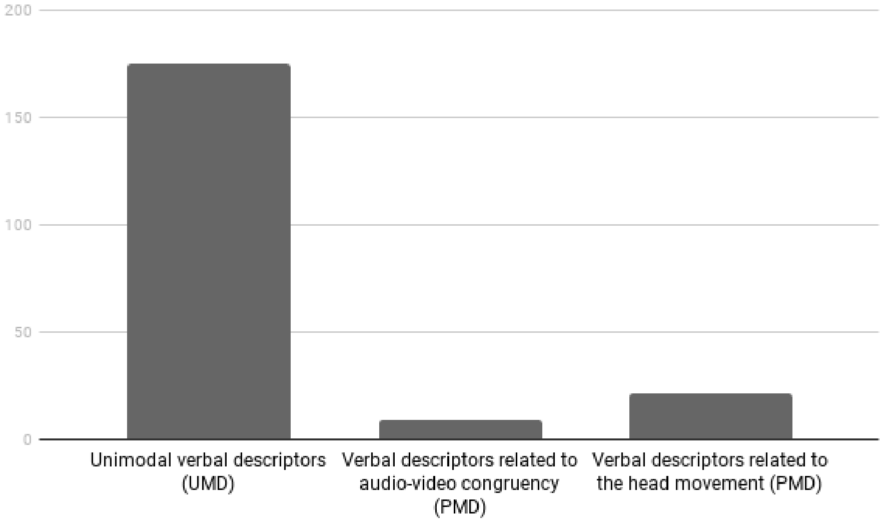


Figure 5: Rate of appearance of spatial sound attributes

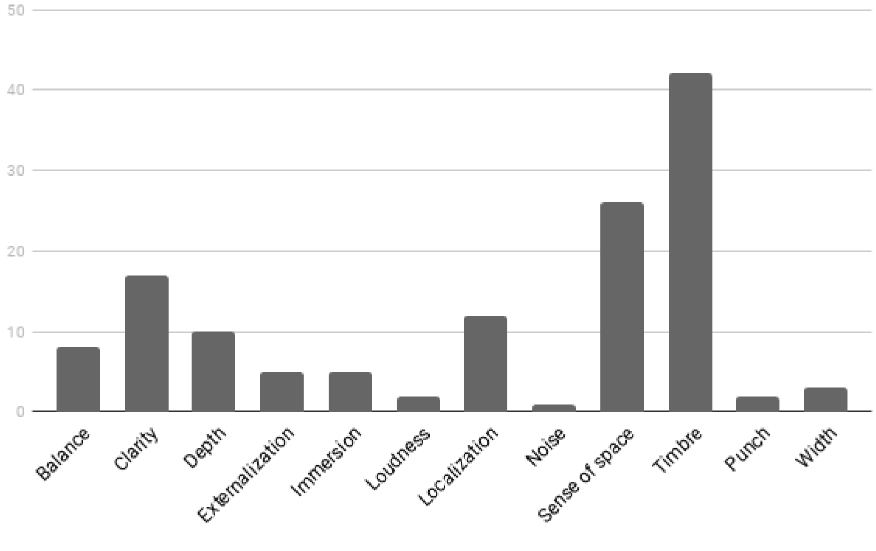


Figure 6: Number of verbal descriptors elicited during experiment

head movement in various aspects. The elicited scales were related to attributes stability and change during head movement. Overall, the consistency of sound between different positions in 360° environment seems to create a new fundamental aspect of sound evaluation for these types of experiences, relevant for upcoming VR and AR multimedia content.

The main limitation of this study is that the conducted experiment only comprised an elicitation stage. Due to constraints, subjects were not asked to use the elicited scales for numerical qualitative rating of the stimuli, which would allow a more robust statistical analysis of verbal descriptors and more precise identification of the attributes. Next studies aimed toward defining attributes of 360° sound should involve methods that allow statistical validation of obtained attributes, such as the Repertory Grid Technique. Other constraints including hardware limitations, low quality of videos, same recording space used in experiments, might have limited the number of attributes elicited in this study. More diversified stimuli might facilitate obtaining a bigger variety of verbal descriptors.

Nevertheless, this exploratory study should be regarded as a first attempt to explore the issue and to propose an experimental strategy to be applied to the new multimedia VR/AR devices that employ spatial audio. The experiment revealed also that the evaluation of 360° sound format is much more time-consuming than the evaluation of stereo or surround formats because of the infinite number of listener positions inside the scene. That should be taken into consideration in future test designs.

CHAPTER IV

SOUND DESIGN PRACTISES IN XR

The majority of this dissertation is dedicated to theoretical discussions on plausibility and the factors influencing it. This chapter offers a slightly different perspective on plausibility. It revolves around the practical aspects of sound design for a specific XR experience which aims at achieving high plausibility of the sound layer. The discussion on practical challenges in designing the plausible audio layer provides a valuable perspective on the broader plausibility issues addressed in other sections of this dissertation.

The chapter starts with a short discussion of the goals and challenges of sound design in XR experiences. The second part reproduces a previously published article describing a case study focused on sound design for co-located narrative VR experiences.

1 Goals and Challenges of Sound Design for Mixed Reality Experiences

The main goal of the sound design for XR experiences is to enhance immersion and presence in the experience (Tatlow, 2024). In order to achieve that the audio layer needs to correspond with visual cues and engage the user on the sensory level. Consequently, the use of spatial audio is critical to provide a three-dimensional auditory space around the user which is necessary to achieve immersion. The audio objects need to be properly spatialized to ensure that the localization of sound sources is accurately aligned with the position of visual cues. Acoustic properties of the space are required to be consistent with the virtual world (in case of VR) or with the real environment around the user (in case of AR) (Serafin et al., 2018). Furthermore, sound design plays a crucial role in evoking emotions and enhancing storytelling (Felnhofer et al., 2015). It provides additional context about actions and environments, supplementing the visual layer. Moreover, sound design for VR supports the interactivity of the experience by giving auditory feedback to the user's actions. In summary, the plausibility of sound is a necessity to achieve the most important goals of sound design in XR, ensuring the delivery of a convincing user experience.

However, practical implementation of sound design for XR experiences may encounter a number of challenges that need to be considered when creating the audio layer. Technical limitations on VR/AR platforms, including constraints on audio processing capabilities, spatialization techniques, and hardware compatibility, may restrict sound design options. It is essential to balance audio quality with performance optimization to ensure smooth playback on VR devices, which often have limited processing power and memory resources. Seamless integration of audio with other elements of the VR experience, such as visuals and interactive mechanics, requires meticulous coordination and collaboration among designers, developers, and audio engineers. Additionally, the immersive nature of VR requires extensive testing and iteration of sound design to ensure that audio cues are plausible, coherent, and enhance the overall user experience.

2 Case Study: Sound Design and Reproduction Techniques for Co-Located Narrative VR Experiences

This section presents the article titled ”Sound Design and Reproduction Techniques for Co-located Narrative VR Experiences” written by Marta Gospodarek, Andrea Genovese, Dennis Dembeck, Corinne Brenner, Agnieszka Roginska and Ken Perlin, published in the Proceedings of the 147th AES Convention (2019). It is reproduced with the permission of the Audio Engineering Society.

2.1 Introduction

Virtual Reality (VR) is expanding very fast in the fields of gaming and entertainment and numerous cinematic productions experiment with headsets to deliver new sorts of experiences. However, most of these works are designed to engage a single user at a time and do not usually invoke a sense of social gathering, a quintessential feature of cinema and theater.

Technological developments observed in recent years now enable the creation of different types of VR production that allow large co-located audiences to experience a shared virtual environment presented in specially-designed entertainment spaces (Layng et al., 2019; Gochfeld et al., 2018). These productions have the goal of bringing back the social aspects of cinema and theatre, which is achieved by designing an experience where the participants are able to see and hear each other as virtual avatars, spatially coherent with their actual physical location. Thus, a cognitive impression of being in a shared experience is established, enabling a sense of presence and awareness of communication between fellow users.

The audio layer is especially important in VR experiences as it affects the subjective senses of *immersion*, *plausibility* and *presence* (Kobayashi & Ueno, 2015), which are key to the success of co-located immersive VR. Spatial audio techniques, which allow users to match the position of sounds with their respective visual cues, can, in fact, improve these quality metrics (Brinkman et al., 2015), while poor audio production can negatively affect them (Zhao et al., 2017).

As of today, there is not much literature on the sound design theory behind this particular style of creative production. The following section reviews the factors and principles behind the implementation of audio systems for co-located narrative VR, whether cinematic or theatrical. We propose the use of hybrid reproduction systems made of both loudspeakers and a transparent hearing device (such as nearfield speakers or transparent earphones) in order to address the audio challenges involved.

The second part of this paper illustrates a case-study discussion around the implementation of the audio reproduction system for a short narrative VR art piece, *”Cave”*. The experience gained by the authors through working on this production helped to inform and validate the design principles discussed, as well as identify the technical variables that may affect particular choices. A short survey was conducted to gain formative insights on the effectiveness of the system and to illustrate the challenging aspects that need to be addressed in future work.

2.2 Background

Immersive co-located VR is a new type of production category which merges some elements of gaming interaction with the linear narrative elements of theatre and cinema. The defining element is the assumption that multiple participants are located in the same room and experience the same virtual content (through virtual or mixed reality devices) under their own individual perspectives and points-of-view, while also being able to see each other in the virtual space. Each participant is rendered in the shared virtual scene as a virtual avatar (usually humanoid), spatially matching their physical location and orientation in real-time, by means of motion-tracking technology.

Since simulating the social setting of a crowd inside a theatre is a goal for these systems, it is important that participants are treated as audience members and feel present as such in the space (Diemer et al., 2015). To this goal, the audience is usually placed in "seats" from which unique first-person views are dynamically rendered and the narrative content is placed onto a virtual "stage".

To technically achieve a multi-user visual reproduction, the headsets are connected using a network-synced infrastructure that allows for the simultaneous delivery of the content for all participants (Churchill & Snowdon, 1998). While the narrative content may or may not be linearly progressing (usually it is), the rendering of the audience members' avatars (e.g. their head rotation and off-axis shift) needs to be actively updating close to real-time. Each client device reports its 3D location and orientation to a server, and receives the location of every other device with their respective timestamps (Herscher et al., 2019). The rendering is finally facilitated at each device through time synchronization signals that make sure that there are no differences in perceived time between participants.

The implementation of such cinematic, or theatrical, experiences can exist under different variations. One particular experimental production, “Holojam in Wonderland”, shown at the 2017 New York’s Future of Storytelling Festival, was portrayed as an “Immersive Mixed-Reality Theatre”(Gochfeld et al., 2018). Two live-rendered actors and four audience members shared a virtual reality stage where a theatrical narration took place in a shared environment. While the actors represented the story characters, the audience was represented by avatars of butterflies, and all were allowed to move in 6 degrees of freedom (DOF), explore the virtual world, and interact with a semi-linear progression of events.

The sound was implemented through a quadraphonic loudspeaker system with an additional overhead speaker. The actors’ dialogue was presented in dual form as live free-field speech alternated to pre-recorded dialogue lines played from the overhead speaker. This choice served the narrative purpose of simulating one actor’s change in size both visually and aurally. No headphones were used as it was necessary for the free-field dialogue to be heard without the effects of occlusion and attenuation of the sound path to the ears, although it is possible to achieve transparent headphone reproduction using hear-through microphones (Rämö & Välimäki, 2014).

2.3 Design Factors and Principles

2.3.1 Sound Principles for Co-Located VR Theater

Each kind of VR experience requires a different approach to sound design and sound production due to the potentially different modalities of the medium used for the storytelling. Most design implementations in VR are based on the game-audio framework (in case of interactive experiences (Horowitz & Looney, 2014)) or on the cinematic framework (in 360° videos (Paterson & Kadel, 2019)). Co-located VR theater entails a set of design requirements for the audio layer which differs from the other types of VR productions:

Transparent Hearing To enable communication within the audience, it is important that users can hear each other during the experience. The use of headphone playback is not appropriate in this context as it impairs free-field listening abilities of the participant. Even open-back headphones are shown to produce occlusion and attenuation effects at the ear canal (Gupta et al., 2018), making the blend between real and virtual sources more difficult to achieve. Although it is possible to equalize this effects with an attentive individualized calibration, a more flexible solution is to employ different kinds of non-obstructive sound reproduction devices such as hear-through earphones or headphones supplemented with microphones that enable transparent hearing (Rämö & Välimäki, 2014). A loudspeaker-only reproduction method would also provide transparent hearing, but likely interfere with other requirements.

Spatial Sound The auditory localization of sound objects has to match the visual localization of the sound sources in order to achieve immersion and presence inside the experience (Brinkman et al., 2015). Spatial audio techniques need to be used to ensure proper perception of the sound localization and adequate proximity effects between the far and the near auditory fields.

Cinematic Sound Design The sound layer has to support the storytelling and reflect a cinematic style of sound design, supporting the full spectrum of sounds which make a compelling experience. The implementation of sound for co-located cinematic VR is merging the approaches from traditional cinema and game audio. The VR narrative is linear, meaning it is played identically for every performance. This format creates an opportunity to design sounds which perfectly match the visual action, without the need for sound randomization which is necessary in games (Horowitz & Looney, 2014). On the other hand, the experience is also interactive. The user has the ability to modify their orientation and position inside the scene, which means that their point of listening can change.

Individual Audio Mix When each member of the audience's "virtual seat" corresponds to their position in physical space, the sound mix delivered also must be matched to that position and orientation, and thus differs for each member of the audience. As a result, each member of the audience receives an individual sound mix which represents their point of listening.

2.3.2 Proposed Reproduction System

The biggest challenge is to allow audience members to hear each other while delivering a high quality spatial audio layer. This paper suggests the employment of a hybrid reproduction system for immersive co-located VR experiences. The proposed system consists of a transparent hearing device and a loudspeaker array.

2.3.3 Spatial Audio Over Headphones

Spatial audio content is more easily and flexibly deliverable through binaural audio techniques. Binaural audio technology allows to reproduce spatial sound by encoding auditory cues into a stereo audio signal, thus changing the perceived localization of object sound sources (Begault & Trejo, 2000).

The cues which depend on the anthropomorphic measurements of the person's head, pinna, and torso are unique for every individual. Head Related Transfer Function (HRTF) characterizes the auditory spatial cues of a person for a defined sound source position. It includes interaural time difference (ITD), interaural intensity difference (IID), and spectral modulations. The limitation of binaural sound in most VR productions is the use of non-individualized HRTFs, which can cause distortions in perceived sound image such as front-back confusions, distortions in localization on the vertical plane, and weak externalization (Guezenoc & Séguier, 2018). Adequate reverberation (coherent to the visual environment) and head-tracking can, to some extent, mitigate the drawbacks of using non-personalized HRTFs. Headphone playback is the most common way of delivering spatial audio, mostly because it can ensure a perfect separation between the two binaural channels. The drawback of headphone reproduction is that even open-back headphones introduce significant attenuation of real-world sound sources (Gupta et al., 2018). The resulting coloration takes away from the plausibility of the experience, and in the case of the immersive co-located experience where the interaction between the audience members during the experience is crucial, a system which enables delivery of audio signals without impairing user's normal free-field hearing is necessary. One of the solutions to that problem are earphone drivers coupled with acoustically transparent earpieces (A. Martin et al., 2009). Another way of delivering the audio are nearfield open ear devices mounted in front of the ears, oriented towards the entrance of the ear canal. However, the small size of transducers in this type of reproduction device often leads to a non-linear frequency response and attenuation in the low frequency region (Gutierrez-Parera et al., 2015). To mitigate the frequency response problem, a hybrid reproduction system with loudspeakers is proposed.

2.3.4 Loudspeaker Playback

Loudspeaker playback is broadly used in cinema production. Surround speaker systems enhance the perception of envelopment and spaciousness of a sound scene, and enable designers to create an impression of movement of the sound sources around the listening space. A sub-woofer speaker provides energy at low frequencies, which are especially important in cinematic sound design where emotional impact is greatly enhanced by the use of low frequency sound effects (Whittington, 2007). The limitation of a speaker-only system is that the subjective localization of sound sources is not very accurate. Furthermore, it is hard to create a convincing virtual source positioned closer to the listener than the physical position of the loudspeaker. Besides that, in surround speaker setups the sweet spot is usually limited to the central seating position (Rumsey, 2007). Adding more speakers to the setup can enhance the immersion and allow for a more accurate trajectory of movement of the sound sources, but the rendering of near-field sound sources is still limited. The use of wavefield synthesis techniques would allow designers to create a very realistic soundfield around the listening area, but its implementation is very expensive and requires acoustic treatment of the performance space (Boone et al., 1995).

The proposed use of a hybrid reproduction system can take advantage of both types of reproduction methods and deliver high-quality convincing and cohesive sound. Hear-through earphones enable transparent hearing and deliver an individual mix of binaural audio to each user, providing an accurate match with the visuals. The speaker system improves the experience by providing a full frequency spectrum of sound and enhances the 3D auditory scene with far-field sounds, which can improve the externalization (Mueller-Tomfelde, 2002).

2.3.5 Technical Challenges

Delay An audio signal played simultaneously through earphones and loudspeakers will not reach a listener at the same moment in time. Signals from loudspeakers arrive to a listener delayed, and the delay will depend on the distance of the listener from the speaker. This issue might be especially important if the sounds played through the device have a short temporal structure (significant amount of transients) where the delay can be perceived by ear (Scharine et al., 1999). Delay adjustments at the binaural device, for some of the seating positions, might be necessary in larger spaces. Contrarily, this issue is not salient for sounds with longer temporal structures.

Distance Attenuation For each doubling of distance from the source, the intensity of a signal in free field decreases by 6 dB (Shinn-Cunningham, 2000). Depending on the distance of each listener to each speaker in a chosen configuration, the signal may be attenuated to a different degree. This seat-dependency must be taken into account during the mixing stage to ensure a proper sound level for each of the more sensitive positions.

HRTFs Rendering When listening to loudspeakers, listeners perceive sound through their own natural HRTFs. The situation is different with spatial sound on nearfield devices: a listener would be usually delivered sound processed through generalized HRTF filters, which are likely non-ideally tuned to their personal spatial cues response. This might cause a problem if too similar sounds were to be played through both the earphones and the loudspeakers, as the HRTFs may color the signal spectrum and create a timbre mismatch between the two delivery methods (Takanen et al., 2012a). In an ideal situation, individualized HRTF filters, measured in the listening room for each seating positions, would deliver the highest possible spatial audio quality. But this is simply unfeasible, a more efficient workaround is to keep the content distinct for the two reproduction systems.

Room Acoustics When playing back sound on speakers, the room acoustics influences the end signal as it reaches listeners' ears. The acoustic character of the room might significantly differ from the one given to the virtual sound layer played at the earphones (for example, when using artificial reverberation). The acoustic mismatch might negatively impact the perceptual auditory integration between the two reproduction elements, affecting the smoothness and capability of immersion into the experience. For this reason, the exhibition room should be acoustically treated to reduce reflections. If this cannot be accomplished, it is desirable to adjust the reverb of the virtual content to be similar or even slightly longer than the actual room reverb, in order to minimize the perceived reverberation mismatch between the two reproduction systems.

2.3.6 Sound Design

The sound-design style for co-located cinematic VR is based on traditional cinematic approaches with the addition of spatial audio processing. The main stems necessary for film audio soundtracks include dialog, music, and sound effects (foley, sfx, backgrounds and ambiances). In contrast to the stereo or surround cinematic mix, there are more audio formats available to the sound designer in shared narrative VR. The audio layers can be reproduced using different spatial audio techniques, even concurrently: as audio objects using binaural rendering, as Ambisonics files that capture the whole sphere of sound around the listener (Gerzon, 1973), as traditional stereo on headphones, as surround formats through VSS processing on headphones (Pike & Melchior, 2013), or as a channel-based mix played back on speakers.

Each sound layer has different requirements in terms of spatial processing and diffusion (see Table 2). The dialogue and foley require very precise scene placement, achievable with binaural rendering. Distance cues such as level attenuation, direct sound to reverb ratio, as well as radiation pattern, need to be added to ensure a realistic sound change during the character's movement (Shinn-Cunningham, 2000). Background and ambiance sounds are usually more diffused. They can be reproduced in Ambisonics format on headphones to allow the rotation of the soundfield according to the listener's orientation, as static stereo tracks, or in surround format on speakers. The music can be reproduced as either diegetic or non-diegetic using different spatial audio formats (Neumeyer, 2009). When using binaural techniques, the music will be perceived as coming from within the virtual space, thus diegetic. When instead using stereo or surround speaker playback it will more likely to be perceived as coming from the background.

Table 2

Suggested audio techniques for audio layers.

Layers	Binaural	Ambisonics	Surround	Stereo
Dialogue	✓			
Foley	✓			
Ambiances		✓	✓	✓
Music	✓	✓	✓	✓

2.4 *"Cave"*: A Case Study

An early version of the proposed sound system was implemented for a six-minute virtual reality co-located narrative piece called *"Cave"* (Layng et al., 2019; Herscher et al., 2019), that took place in a single, multi-user, virtual environment. The story involved one main character, one supporting character, and a virtual mammoth. The experience was prepared for a 30-member virtual audience, separated into two groups in the thrust stage format (Fig. 7). The audience could see each other as avatars inside

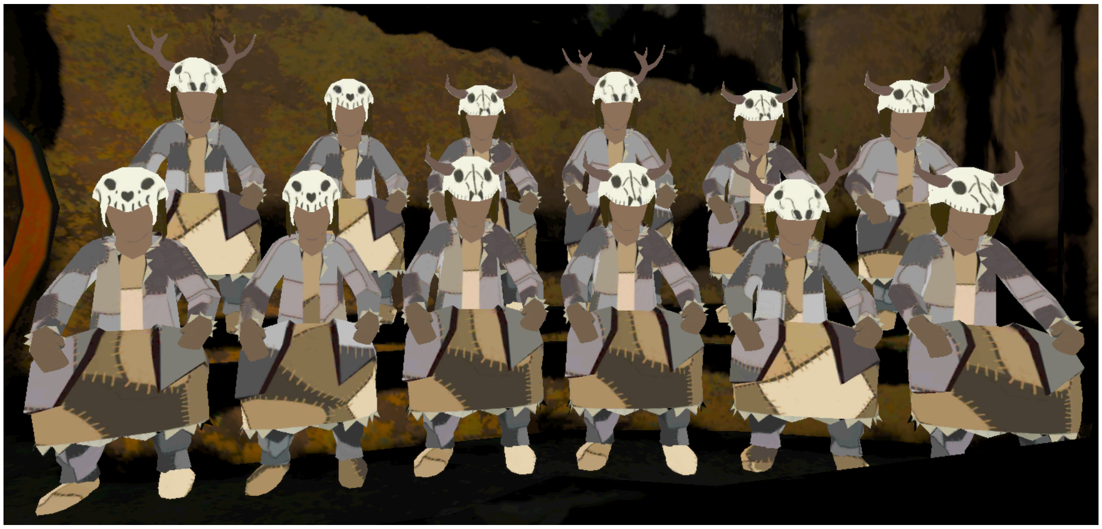


Figure 7: Audience avatars in the VR experience *"Cave"*. (© and Art: Kris Layng, 2018)

the virtual experience while the position and orientation of their heads were tracked using the headsets' IMUs, so that the avatars' heads moved inside the virtual space accordingly. The VR experience was built using the Unity game engine (*Unity Real-Time Development Platform | 3D, 2D VR & AR Visualizations*, 2019), and was executed on standalone headsets for each audience member, using a smartphone as the control unit. The game networking service Photon (*Multiplayer Game Development Made Easy | Photon Engine*, 2019) was used for sending all data and signals between all devices (Herscher et al., 2019).

2.4.1 Design Choices

The audio layers used in the experience consisted of dialogue, music, sound effects, and ambiances. The sound effects and the dialogue materials were treated as point source objects, connected to a visual component in the three-dimensional space. While the original content of these materials was in mono format, a dynamic binaural rendering engine tool (Steam Audio) transformed the sound objects into responsive stereo binaural streams, responding to the spatial relationship between characters and listeners. Additional distance cues were tuned separately for the foley, while the dialogue track had to maintain constant intensity in order to keep it intelligible. As the rendering was individual per-device, each audience member was able to get a unique sound perspective into the scene.

The ambiance sounds were created from both a mix of stereo recordings and sound objects. Important and constant background sounds, such as the wind noise in the entrance to the cave or water stream, were positioned at diffuse point sources within the scene, while more general and sporadic sounds, such as drops of water, were rendered in stereo and not given specific spatial positions. The music track was exported as a non-spatialized stereo in order to give it a sense of separation from the dialogue and sound effect layers.

2.4.2 Audio Workflow

The audio implementation for the project was done in the Unity Engine using the Steam Audio plugin (Valve, 2019) for sound spatialization. The IMU tracking data was utilized to affect the individualized mix for each of the participants. The audio stems were designed and edited in Pro Tools (Avid, 2019) following a traditional linear workflow, as in film post-production, using a 2-D video rendering of *”Cave”* provided for reference for all editing. The mixing stage was split between Pro Tools and Unity. All equalization, compression, and limiting was applied in Pro Tools before adding to Unity to ensure that all stems blended well together before being spatialized. Those were mixed “dry” so that the reverb could be rendered in Unity based on the listeners’ locations. Once the stems were imported into Unity, they were attached to the corresponding visual object or rendered in stereo format. Spatialization, reverberation, and additional equalization for stems were programmed per-user using the Steam Audio plug-in, while the Unity audio mixer was used to introduce general changes in the intensity of the audio layers. Finally, the synchronization between the visual and audio layers was implemented using the *timeline* playback tool, for both visuals and audio tracks.

2.4.3 Reproduction System

The sound reproduction system consisted of one single speaker with subwoofer placed in the middle of the stage, and a prototype nearfield open-ear device produced by Bose Corporation specifically for *”Cave”*, which allowed transparent hearing. The devices were mounted in front of the ears and were oriented to project towards the entrance of the ear canal (Fig. 8).

The speaker unit was used mostly for the sound effects of the virtual mammoth and it was made sure that the physical position of the speaker matched the virtual position of the mammoth’s avatar as seen by all audience members. In this implementation, one speaker was sufficient because the visual object for which the sound effects were rendered was static. With more moving elements, more speakers would be necessary to ensure proper localization of the sounds chosen to come from the loudspeakers. A system-wide calibration was performed to achieve a proper blend

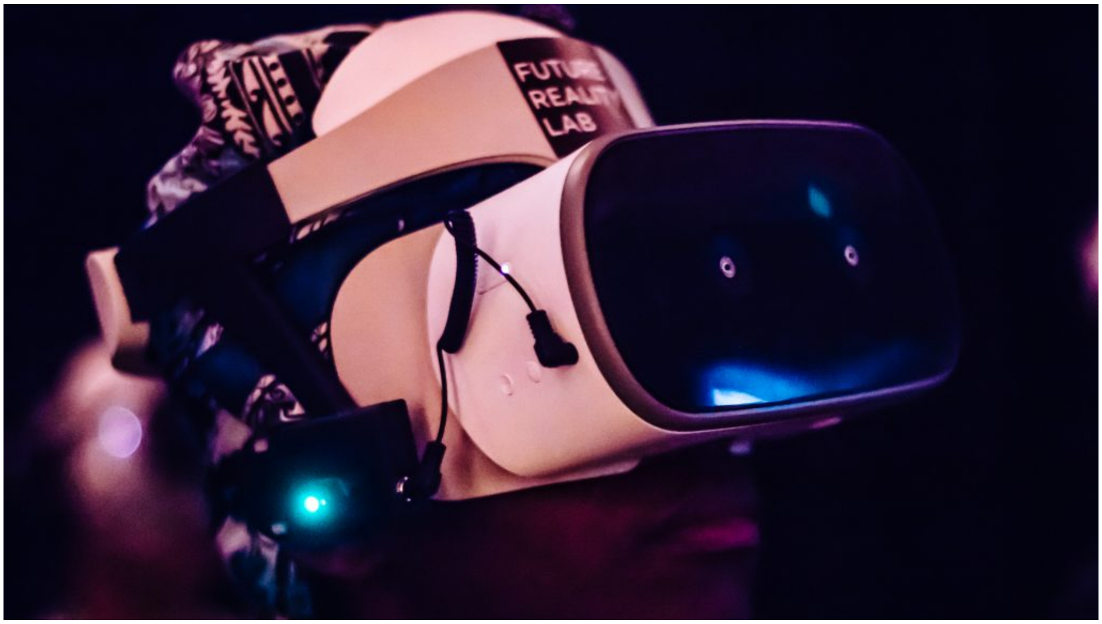


Figure 8: Headset with prototypes of nearfield open-ear device from Bose. (Photo: Eric Chang, 2018)

between the nearfield devices and the loudspeakers, and to ensure that the levels would be comfortable for each member of the audience.

Table 3

Judgments based on the user's familiarity with VR. The questions were given on a Likert scale (1, Strongly Disagree to 7, Strongly Agree).

Impact of Familiarity with VR on Audio Experience

Item	Low Familiarity	Medium Familiarity	High Familiarity
I enjoyed the *"Cave"* experience	5.88 (1.12)	5.72 (1.26)	5.92 (1.19)
I enjoyed the audio elements of the experience	5.76 (1.02)	5.77 (1.18) 5	6.06 (1.0)
I understood some audio elements were spatialized (placed in the room)	5.32 (1.54)	5.66 (1.50)	5.97 (1.38)
I felt the audio spatialization helped me feel immersed in the experience	5.79 (1.27)	5.90 (1.25)	6.14 (1.10)

2.5 Survey

To explore the efficacy the sound implementation, we developed a questionnaire offered to all 1,927 users immediately after watching the experience. We received 374 responses (a 19% response rate), of which 317 were complete and used to provide richer insights into user experiences.

The questions took one of four formats: i) 7-point Likert-type scale, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree), ii) single choice response, potentially including an “Other” option with short text entry, iii) multiple selection response, potentially including an “Other” option with short text entry, and iv) Open-ended text response.

2.5.1 Respondent Profile

Over half of the participants reported their age as under 35, with 57 (18%) reporting 18-24, and 117 (37%) reporting 25-34; among those over 35, 65 (21%) reported 35-44, 42 (13%) reported 45-54, 18 (6%) reported 55-64, 10 (3%) reported 64+, and 8 (<3%) preferred not to give an age. Participants were asked where they had been seated in the audience from a list of 5 areas; they came from a relatively even distribution of the areas with the fewest responses from the right front row (57, or 18% of the sample) and the most responses from the left front row (76, 24% of the sample).

Participants reported how long they had used virtual reality technology, and a single self-reported value for level of expertise with virtual reality. Based on these responses, we created three categories of familiarity with VR: High familiarity participants (86, 27%) had used VR for a year or longer, and rated themselves a 6 or 7 (Extremely proficient); Low familiarity participants (88, 28%) had used VR for less than a year, and rated themselves a 3, 2, or 1 (Not at all proficient); and Medium familiarity participants (143, 45%) provided any other combination of time and rating of expertise.

In sum, this sample of conference attendees comprised relatively young, technically-savvy professionals. Although participants were not randomly selected from the audience, they viewed the experience from all areas, and had varied expertise with virtual reality.

2.5.2 User Experience

All participants were asked for their judgments of the experience as a whole, and specific questions on audio quality and spatialization. Participants enjoyed *"Cave"* and the audio elements of the experience, regardless of experience with VR (F's < 2.3, p's > .01). Participants at all levels of VR experience also reported understanding that audio was spatialized, and the spatialization contributed to feeling immersed in the experience (Table 2).

However, responses did differ based on seating for *"I enjoyed the audio elements of the experience"*. Participants in the right back row had lower reported ratings (M = 5.45, SD = 1.22) than other 4 locations (Means > 5.7), a small but significant difference, F = 3.51, p = .008, η^2 = .044.

Most participants indicated that they enjoyed the score, effects, and foley effects; a smaller but substantial number of participants reported enjoying the dialogue (Table 4).

Table 4

Percentage scores for different audio layers

Which elements (if any) did you enjoy?

Element	Percentage of participants
The score	81.1%
Effects	61.5%
Foley	61.2%
Dialogue	35.5%

2.6 Discussion

In the presented case study of sound design and reproduction for immersive co-located virtual reality theatre, the cumulative effects of real-world elements (including seating, networking, and delivery of visual elements of virtual reality) and the implementation of a hybrid reproduction system delivered an effective sound experience for this shared virtual art piece.

The results of the survey suggest that the presented approach can be sufficient for delivering an immersive audio layer given the defining elements of this particular experience, although the results are descriptive for this convenience sample, and not intended to generalize to a wider population. Participants, in general, enjoyed the audio elements of the experience. However, the audience members seated in the right back row gave significantly lower ratings of enjoyment, although they did not indicate an impact on their understanding of spatialization, and feeling that the spatialization helped them feel immersed in the experience. It is likely that these lower ratings were affected by audio glitches found to occur during several of the showings due to jittery network connections in some devices.

Participants gave lower scores to the dialogue when evaluating the different elements of the audio layers (Table 4). We noticed that the D/A converters on the headsets introduced a significant amount of sound distortion which affected mostly the quality of the dialogue rendering and might be reflected in the results of the survey. This indicates that the quality of headset's audio hardware should be taken into account when choosing a device for VR production.

Most of the participants noticed that the sound was spatialized and felt it helped them to feel immersed in the experience, which suggests that the sound implementation and reproduction was successful to enhance the immersion. However, having a control group evaluating a reference audio track should allow for more robust empirical results which was not possible within the context of this production. Also, more specific evaluation questions could bring more conclusions about the perception of sound during the experience.

The implementation described in the case study was limited to a single speaker and subwoofer. Adding more speakers surrounding the audience may further improve the immersion and allow the reproduction of more layers of audio other than sound effects, e.g. music or ambiance. Playing music tracks through the loudspeakers could indeed help with better separating the background music and the dialogues. We also noticed that some of the instrument tracks were perceived as coming from within the scene even though they were rendered in stereo.

Our implementation did not take into account the acoustics of the performance space due to the production limits. This resulted in sounds played through the speakers having different acoustic characteristics than the binaural layer. Furthermore, informal investigation revealed that even though the audience could see and hear each other inside the experience, their voices were not fully perceived as though they were coming from the same space as the action. Placing microphones above the audience may solve that problem. The sound from microphones would be processed in real-time through the same reverb processor used within the experience, to ensure consistency between the sounds of the real-world natural environment and virtual scene.

Another challenge we encountered during production was the asynchronous playback between speakers and nearfield devices. Even a small delay would cause perceptible distortions of those sounds played through both systems. We solved this problem by removing all of the transient sounds from the speaker playback, leaving only the sound effects of a longer temporal structure.

2.6.1 Future Work

Although the discussed production work helped to elucidate and expand the sound-design theory for this type of narrative VR experiences, more empirical work is required to investigate and validate the best approaches for an effective delivery of sound. While the discussion of the principles is mostly derived from professional experience and qualitative critical perspectives, the assessment of the proposed technical systems, the factors and the challenges involved can benefit from both commercial production analysis and laboratory experiments. Further insights can be gained by literature advances in similar applications such as multi-player VR interactions and spatial audio technology.

In practice, future productions would benefit from a revised questionnaire linking and addressing the accuracy of sound localization, device type (VR vs AR), seating position, sound source externalization, acoustic treatment and matching, and quality of interaction between audience peers. Having controlled conditions in a laboratory experiment would create robust conclusions about the importance of each one of the single elements which compose the proposed hybrid system.

2.7 Conclusion

This paper presented a discussion around the principles, factors, and limitations of the sound-design theory related to the novel field of co-located narrative VR experiences. A first draft of this theory, reviewing technical challenges and proposing a solution based on hybrid reproduction systems, has been derived from practical experiences within prototype productions. The experience with the production of "Cave" is discussed as a platform where some of these principles were investigated and addressed to achieve insights that inform the authors' proposed framework.

Having this base to work upon, future empirical data will help to validate, sharpen, and define the guidelines that may drive the creative choices of VR sound designers. It is reasonable to expect, that in the near future, technological advances are likely to affect and update the current conversation.

CHAPTER V

EXPERIMENT: EVALUATION OF PLAUSIBILITY AND OTHER SOUND ATTRIBUTES IN AAR CONTEXT

This chapter introduces the core part of the dissertation that concerns the experiment aimed at the investigation of plausibility perception in the context of the AAR environment. Previous chapters provided the literature background on the design methods of Virtual Acoustic Environments and methodology for sound plausibility evaluation studies. Besides that, the research focused on searching for the sound attributes important in spatial audio perception for immersive experiences was described. We presented also a case study showcasing different approaches to sound design in XR. The example was focused on delivering the highest plausibility of a spatial audio layer that is crucial to designing a truly immersive XR experience.

The present experiment explores plausibility perception in the AR context where real and virtual sources are presented together. This scenario demonstrates the most challenging case of audio implementation for XR environments as the user is constantly exposed to the real reference by the sounds of the environment which reveal the properties of the space. By allowing to move freely in the space, 6DoF environments provide more dynamic cues possibly making the comparison between real and virtual sounds even easier than in a standing scenario. The following parts of the dissertation will describe the perceptual experiment, present the analysis of the results and acoustic parameters of auralization methods, and discuss its implications.

1 Research Questions

As stated in Chapter 1 the main research question addressed by this dissertation is the following:

What perceptual and acoustic factors are meaningful in the plausibility perception of sound design for Augmented Reality environments?

The study will aim to answer a set of subquestions which can be divided into several areas of interest posed below.

Perceptual Evaluation

- How does a subject's freedom of movement affect the perceptual evaluation of an AR sound scene?
- What is the correlation between plausibility and other perceptual attributes of sound?
- Do the properties of real reference affect plausibility judgment?

Acoustics

- How do objective measures of acoustical parameters correspond to subjective evaluation of acoustic processing?
- How does the position of the source in the room and orientation influence the assessment of the auralizations?

Methodology

- Is the proposed methodology an effective method for evaluating plausibility in 6 Degrees of Freedom (6DoF) AR environments?
- How do the participants' speed of walking and amplitude of yaw movement affect the evaluation?

2 Goal and Purpose of the Study

In order to address the research questions stated above and explore the influence of acoustic and perceptual factors on sound plausibility (illustrated in Figure 9) the dissertation study implemented a novel experimental design. During the study participants rated the plausibility of pairs of loudspeakers (real or virtual) playing consecutively from different positions in the room. This approach set expectations of the user very close to the real-life scenario where similar but not identical sources exist within one environment. An immediate comparison with the real source was available but the reference was not identical with the virtualized sound as in real applications of an audio system.

The experiment was carried out in two phases in order to evaluate the influence of participants' movement on perceptual judgment. During the walking phase, participants walked back and forth following a predefined path which allowed for the evaluation of the dynamic rendering of the stimuli. During the standing phase, participants stood at a point close to the beginning of the path. The use of special transparent headphones (AKG K1000) with compensation filters allowed for direct comparison between the simulation played on headphones and real sources played through loudspeakers positioned in the room. After each trial, participants filled out a questionnaire. The questionnaire not only investigated the plausibility perception but also included several perceptual attributes taken from the Spatial Audio Quality Inventory (SAQI) (Lindau et al., 2014). The decision to include these attributes was aimed at exploring potential correlations between plausibility and other sound attributes. To facilitate such an investigation, plausibility was assessed on a continuous scale, departing from the binary ”yes/no” approach often used in the literature. Additionally, recognizing plausibility as a continuum allowed for a more nuanced evaluation, preventing data from being unnecessarily constrained. This approach accounts for scenarios where virtual sound sources may be distinguishable from real references without necessarily being perceived as implausible. Moreover, real sound sources might themselves appear less plausible in certain contexts, such as when the direct path is obstructed. Thus, the methodology included not only comparisons between virtual and real sources but also evaluations of pairs involving two real sources and two virtual sources. By analyzing these results, we aimed to investigate how the presence of a real source influences judgments of plausibility.

Another goal of the study was to validate the proposed experimental design by comparing two different approaches to the auralization of early reflections and late reverberation. The implementation of two different auralization methods allowed us also to search for correlation between objective acoustic parameters of auralizations and measurements and subjective ratings of the participants.

The first method exploited a simple 3D numerical model of the room and ran a real-time beam-tracing method to calculate the time and spatial distribution of the reflections. The late reverberation was modeled with the Feedback Delay Network (FDN). The second auralization method characterized the room with a single spatial room impulse response (SRIR) which was further manipulated in order to account for the relative listener-source distance. Both methods employed the same simulation of the direct sound propagation effect with a simple directivity model to simulate the loudspeaker radiation pattern. The two methods had different limitations and were chosen to investigate the possible influence of weaknesses of the simulations on plausibility judgment. With the first approach, the temporal structure of reflections was expected to be closer to reality as it was constantly updated according to the source and listener positions in the room. In contrast, the late reverberation was only an approximation of the measured IR as it was rendered through an FDN implementation. On the other hand, the SRIR auralization could accurately reproduce the time and frequency distribution of the late reverberation. It could also reproduce accurately the whole room effect for the measured position, but could not account for the varying time and spatial distribution of early reflections along the walking path. The temporal structure of reflections was kept from the original SRIR and did not depend on the position of the listener and source as it would in reality.

3 Overview

The dissertation experiment is described in Chapters V - VIII. Chapter V delves into the objective characterization of the experimental room and sound source. Detailed descriptions of the acoustical and software designs of two distinct auralizations, tailored

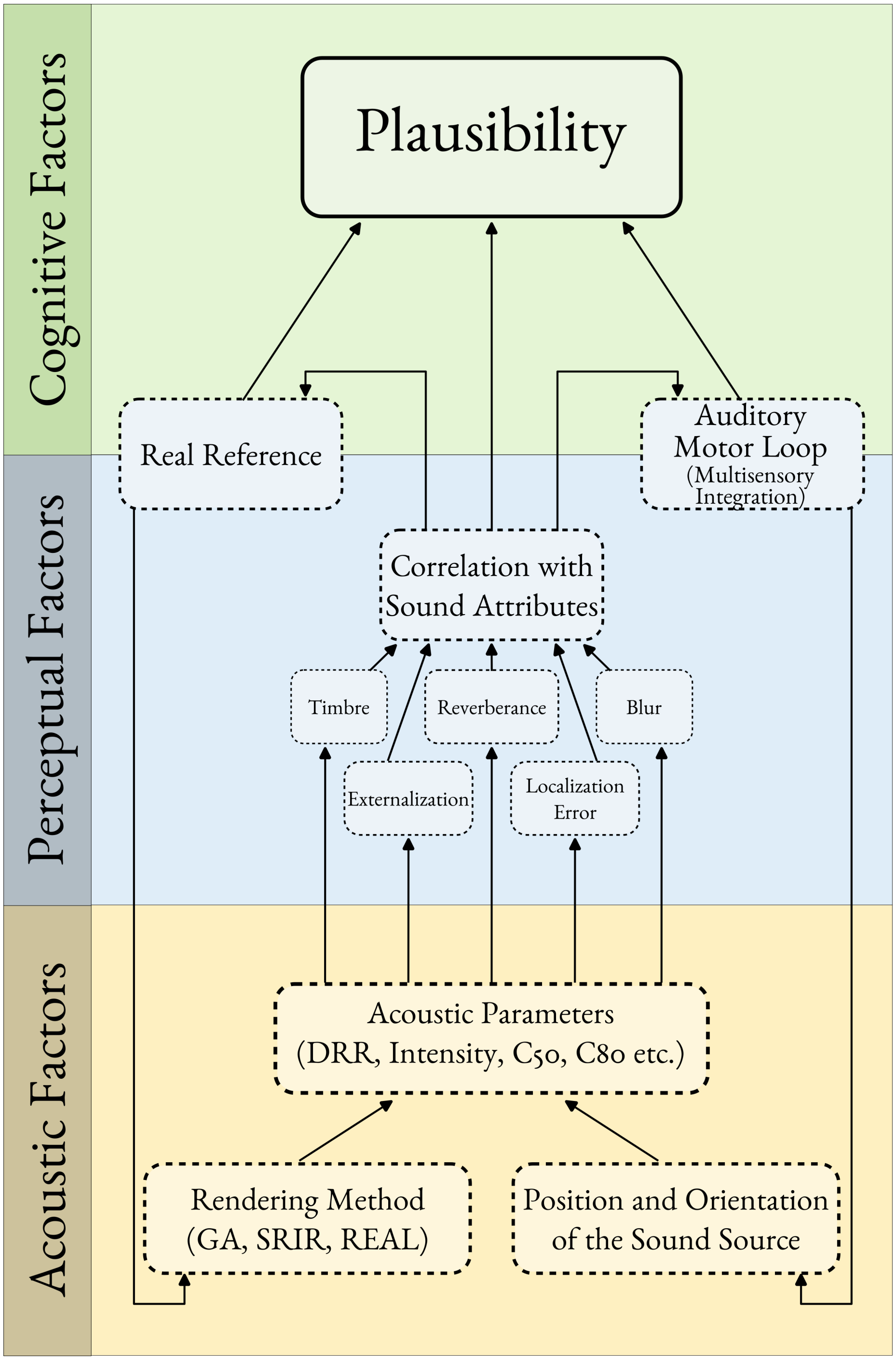


Figure 9: Acoustic and perceptual factors of plausibility

to the experimental room and loudspeaker, are provided in subsequent sections. Finally, the chapter outlines the experimental design of the perceptual study, focusing on evaluating plausibility and other sound attributes in the AAR context.

Moving to Chapter VI, the results of the statistical analysis conducted during the first phase of the study are presented. Following this, Chapter VII offers the outcomes of both the first and second study phases, alongside an objective analysis of acoustic parameters for both auralization methods and measurements. The chapter concludes with a discussion of the relationship between objective analysis findings and subjective ratings. Chapter VIII presents the summary of contributions as well as the implications of the study on sound design practices in AR.

4 Room and Loudspeaker Directivity Measurements

Acoustic simulations reconstructed the experimental room (Studio 5 at IRCAM) of size 10.68 m x 7.83 m x 4.17 m and cubature 348.71 m^3 as shown in Figure 11. The walls and ceiling are covered with a random arrangement of absorptive and semi-reflective panels. The floor is covered with linoleum on concrete. One wall has a big glass window covered with a thick curtain. During the listening sessions, the room was nearly empty. The measured reverberation time at 1 kHz was 0.28 seconds with a slight global decrease according to frequency (see Figure 16).

In order to conduct an objective comparison with the auralizations (see section 5.6), measurements in the room were performed using sine-sweep signals (Farina, 2000). One path with four different positions of the speaker was measured. The measured path is a straight line parallel to the main room axis, next to the speakers at a minimum distance of 1.5 m (see Figure 11). Measurements were taken every 30 cm which resulted in 26 measurement points. Each point was measured using the EM32 Eigenmike® (4th order 3D soundfield) spherical microphone array from *MH Acoustics* and a Neumann KU 100 dummy head. After the measurements, each of the impulse responses was denoised to allow for proper analysis and avoid any artifacts (Massé et al., 2020).

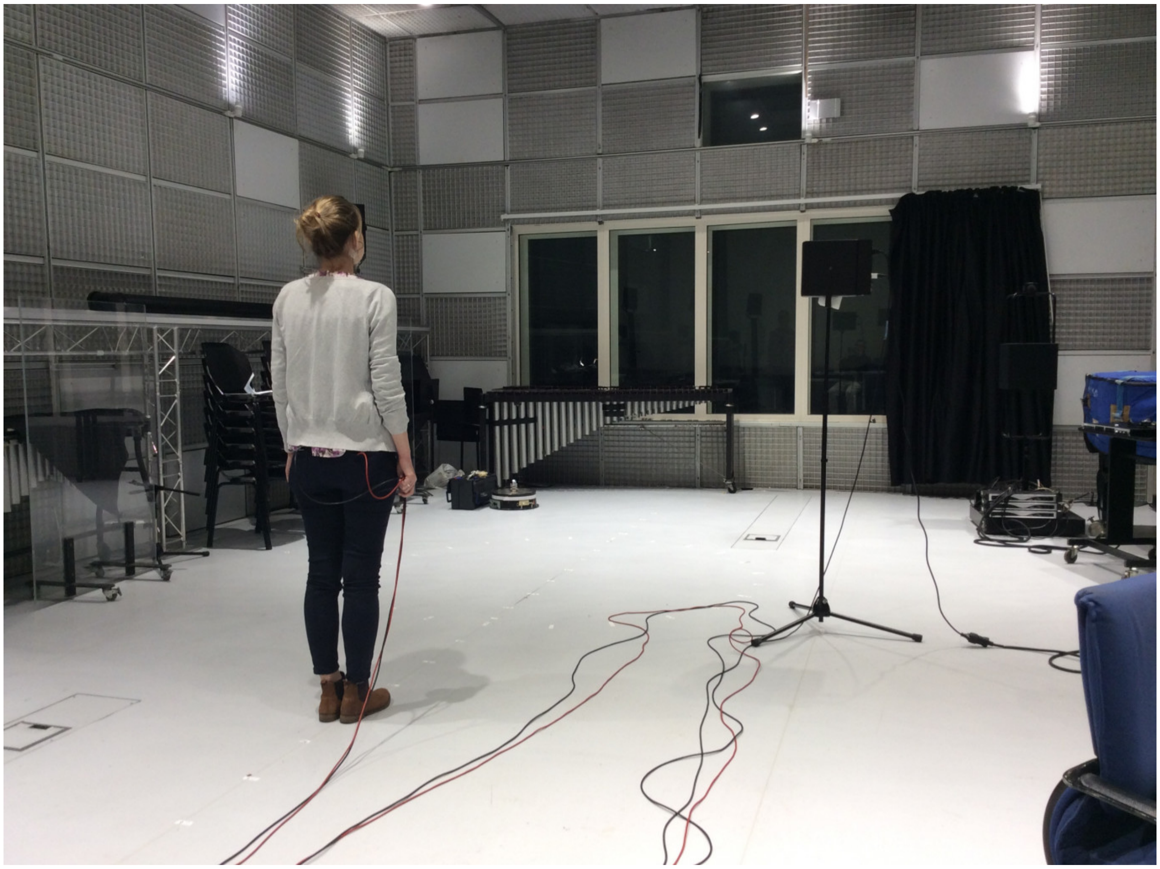


Figure 10: Studio 5 at IRCAM

The four loudspeakers used for the experiment were Amadeus PMX 5. One of them was measured in an anechoic chamber at IRCAM to characterize its frequency response and directivity. Measurements of the speaker were performed using a sine-sweep deconvolution method (Farina, 2000). The microphone was set up 1 m from

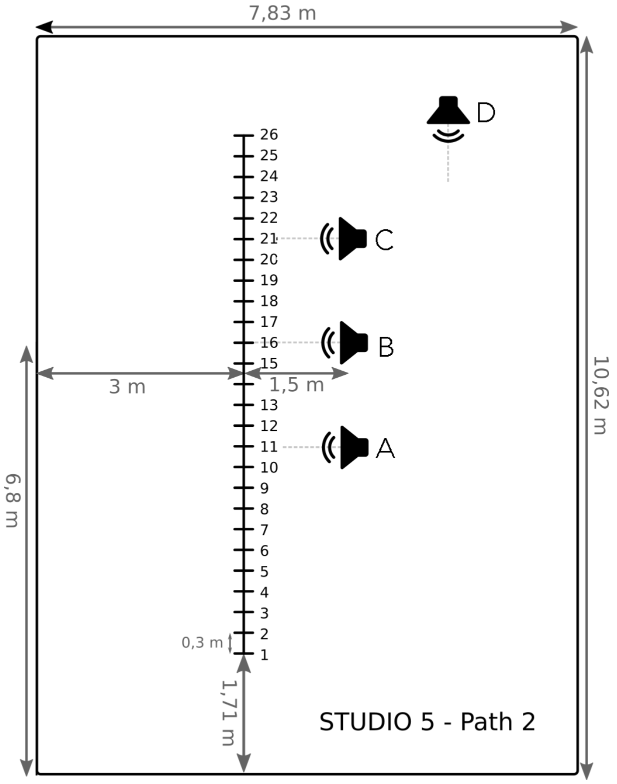


Figure 11: Setup for measurements - Studio 5 at IRCAM

the speaker. Measurements were taken every 15° around the speaker which resulted in a total of 24 measurement points. Thanks to the axisymmetrical design of the PMX 5 (dual concentric), the measurements were done only on the horizontal plane (0° elevation).

Table 5

Mixing time estimation

BRIR	SRIR	Simulation
45.5 ms	73.5 ms	80 ms

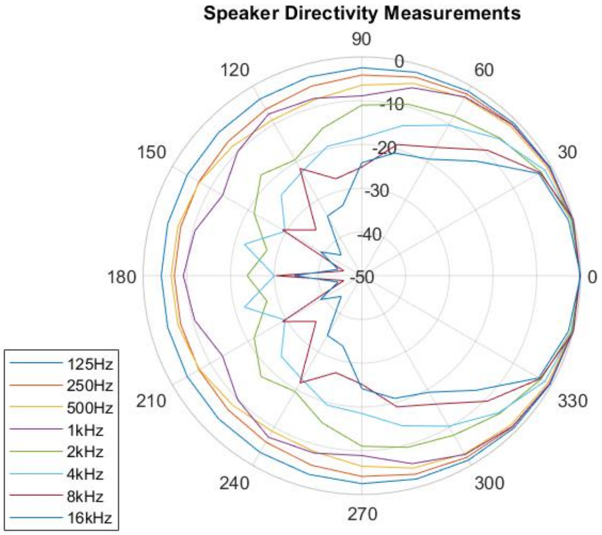


Figure 12: Directivity of the PMX 5 loudspeaker

5 Implementation of Virtual Acoustics Environment

The goal of the AR simulation is to create an illusion that the sound played on headphones is coming from the real speaker in the room. In order to achieve that, two different acoustic simulation methods were employed to recreate the acoustics of studio 5 at IRCAM and simulate the PMX 5 loudspeaker (see Appendix B).

5.1 Direct Sound and Directivity Modeling

The rendering of the direct sound for both auralizations was the same (see Figure 13). At first, the propagation delay was added to the initial stimulus based on the distance between the listener and the source. After that, the signal was filtered with the on-axis spectral response of the experimental loudspeaker, and then the source directivity model described below was applied. Next, the intensity change was imposed following the inverse-square law which states that the intensity of sound decreases by approximately 6 dB for each doubling of distance from the sound source. Finally, the signal was convolved with a proper Head-Related Transfer Function (HRTF) controlled by the rotation of the listener's head and the relative position of the source.

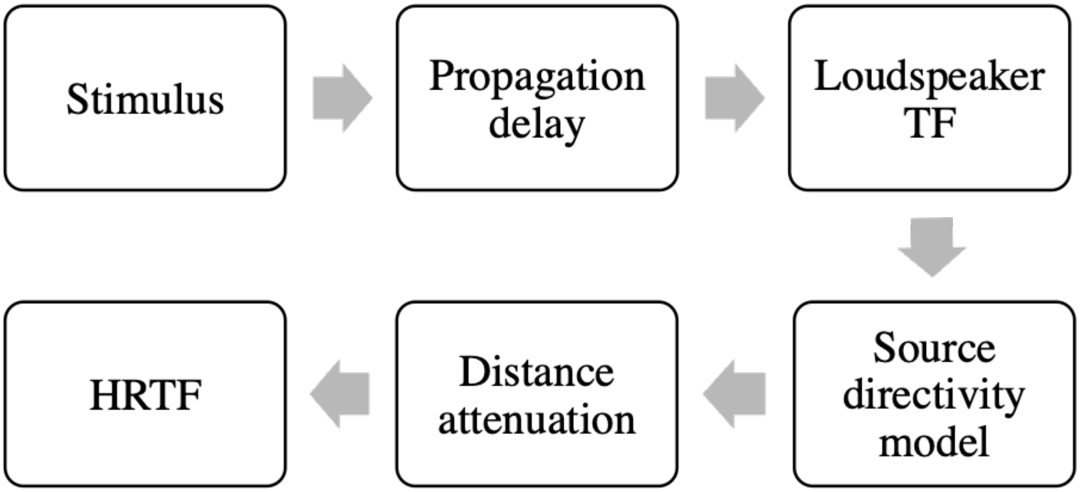


Figure 13: Direct sound simulation

The directivity of the source was synthesized through the implementation of directional filters. For simple radiation patterns such as the one of concentric loudspeakers, the simulation can be approximated with a single spatial dirac distribution bandlimited with the order of the HOA decomposition: the higher the order the narrower the beam. The directivity modeling was based on beamforming up to 4th order HOA which allowed to approach a given radiation pattern (Carpentier & Einbond, 2022). To ensure the best possible match of the model for each frequency band, the simulation was verified by comparing the fit of the curves from the measurements and simulation focusing on the range of angles from which the listener would hear the speakers (-70° 70°) along the path (refer to Figure 15). The operation was repeated in eight frequency bands (Kronlachner & Zotter, 2014). Subsequently, the total power spectrum radiated which drives the spectrum of the reverberated field was validated to ensure it is similar to the measurements. The resulting directivity index for simulation and measurements are compared in Figure 14. The real source is a bit more directive in a frontal area but at the same time less directive at the back of the loudspeaker. It means that the impression of crossing the loudspeaker when walking near it will be slightly reduced.

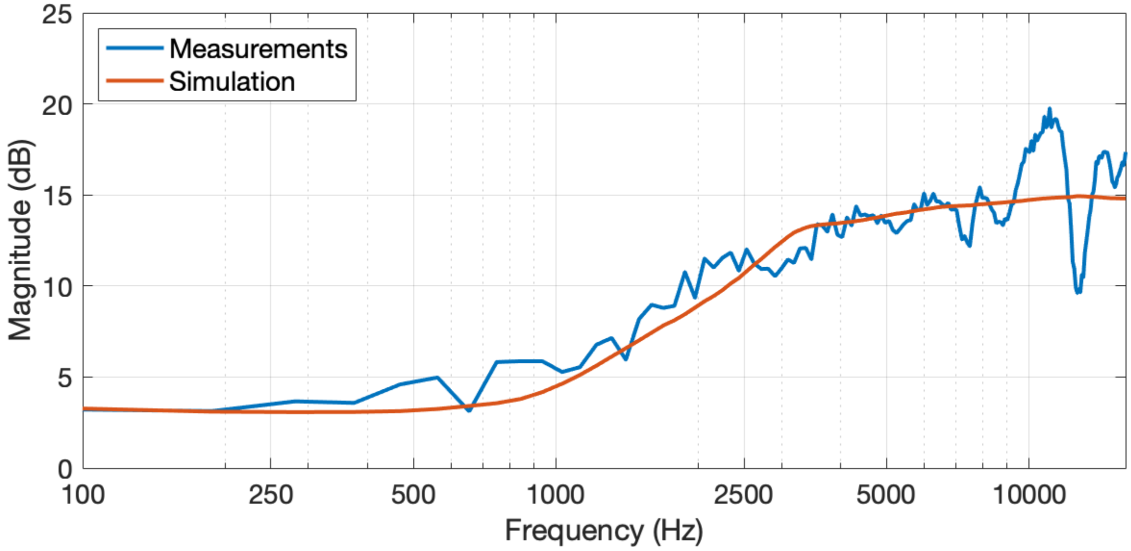


Figure 14: Directivity index of loudspeaker PMX5

The filters derived from the directivity model are applied in real-time to the direct sound segment of the impulse response based on the angle between the listener and the source continuously provided by the tracking device.

5.2 Auralization Based on GA Simulation

The first auralization method, labeled GA in the following, is based on geometrical acoustic modeling of the room. The method combines a real-time beam-tracing

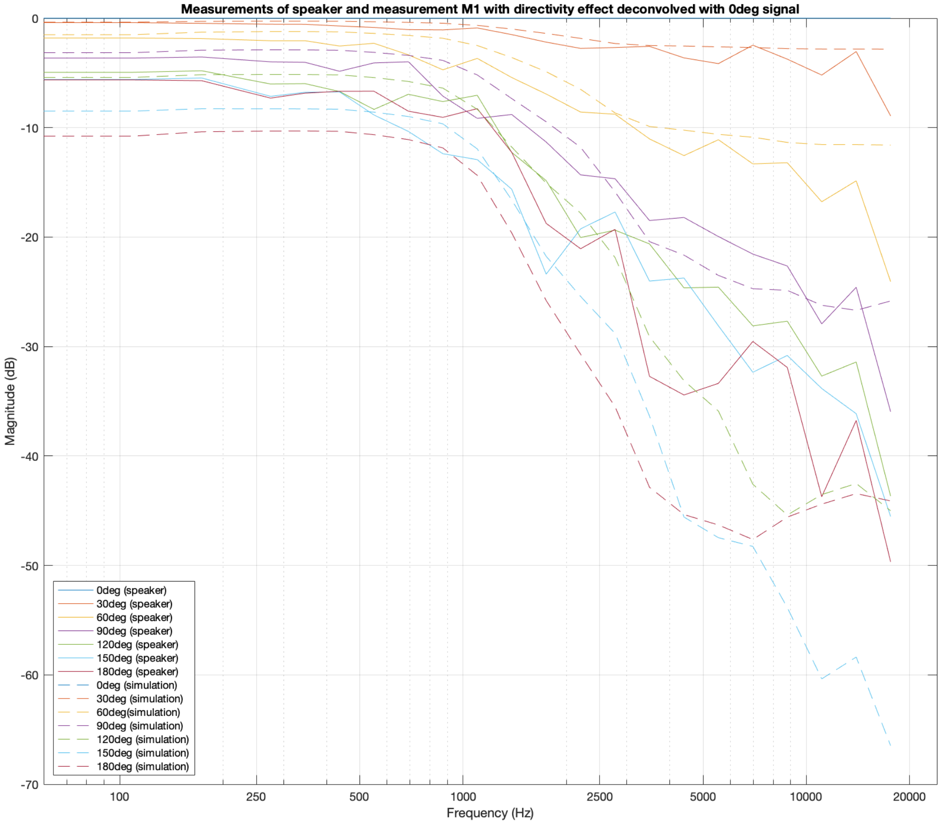


Figure 15: The results of directivity model simulation: the speaker measurements and simulation deconvolved with on-axis signal

algorithm for the simulation of early reflections, and FDN for the rendering of late reverberation. The expected advantage of the method is that it calculates the early reflections segment of the RIR based on the actual geometry of the room and according to the instantaneous position and orientation of the source and listener in the room - thus potentially giving a more accurate space-time distribution of early reflections. On the other hand, the late reverberation is an approximation of the actual RIR decay based on the reverberation time estimated in a limited number of frequency bands.

A simplified 3D model of the experimental room was designed and input to the modeler. The model was then calibrated to obtain the same frequency-dependent reverberation time as BRIRs measured with the KU 100 and averaged over six positions (1 to 6, see Figure 11). The floor material (linoleum on concrete) was assigned an absorption value based on the literature (Fediuk et al., 2021). As the absorption coefficients for wall and ceiling materials were not known, they were estimated based on the measured reverberation time, using Eyring's formula:

$$RT_{60} = \frac{0.163 * V}{S_{tot} * [-ln(1 - \frac{\sum(\alpha_i * S_i)}{S_{tot}})]}$$

where V - volume,

S_{tot} - total surface area

α_i - absorption coefficients for surfaces

S_i - surface areas with V the volume, S_{tot} the total surface area, S_i the surfaces areas and α_i their absorption coefficients. The choice of the Eyring formula was justified by the random distribution of absorptive and semi-reflective panels on the walls and ceilings as well as by the higher proportion of absorptive panels (Astolfi et al., 2008).

The rendering system was implemented using the EVERTims module of the Spat5 library running in the Max/MSP software (Carpentier, 2021). EVERTims is an open-source framework for 3D models auralization (Poirier-Quinot et al., 2017).

The modeler unit constructs a beam tree for the current scene geometry as well as the positions of the listener and the source. The beam tree is a base to generate a list of image sources sent to the auralization object. The modeler characterizes each reflection path by its direction of arrival, propagation delay, filtering due to the source directivity and frequency-dependent material properties, and air absorption. Directivity of the source in the modeler is applied to both the direct sound and image sources according to the model described in section 5.1. Image sources up to the 3rd order, and limited to reflections earlier than 100ms were implemented in the system. Each of them was encoded into a 4th-order HOA soundfield according to its direction of incidence. All together, they form a single ambisonic stream representing the early reflections segment of the RIR.

Synthesizing the late reverberation through image sources modeling would however not be efficient since the computation cost increases exponentially with the reflection order (Vorländer, 2008). The late reverberation was simulated with an 8-channel FDN, in which parameters (decay rate and modal density) were set to match the BRIRs measured in the room (see Figure 16). The incoming signal feeding the FDN was equalized according to the power spectrum radiated by the loudspeaker. FDN is characterized by a slow building up of first reflections before reaching a high-density reverberation. Hence, an anti-phase filter was used to cancel out this building-up process until 80ms (i.e., with a small overlap with the latest image source reflections). This guarantees that only the first reflections provided by the image source model are delivered to the listener (Greenblatt et al., 2010). The transition time of 80ms between the image source reflections and the FDN late reverberation was chosen to match the mixing time observed on the measured SRIRs (i.e. the time when a sufficiently high echo density is achieved). The mixing time was estimated from the analysis of the spatial coherence of the measured SRIRs (Massé et al., 2020). The eight FDN output channels were encoded to 4th-order HOA with diffuse panning. Both the first reflection ambisonic stream and the late reverberation stream were mixed before being sent to the binaural decoder.

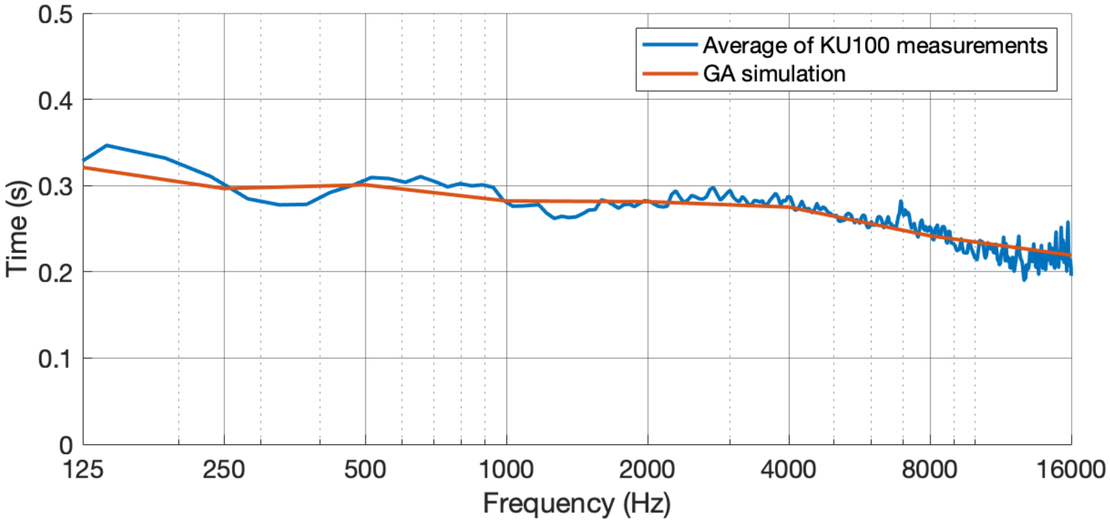


Figure 16: Measured RT_{60} and its FDN synthesis

.

5.3 Auralization Based on SRIR Synthesis

The second auralization method, labeled SRIR in the following, is based on a convolution approach (Nowak & Klockgether, 2017) using a single reference SRIR among the measurements described in section 4. The reference SRIR corresponds to loudspeaker B measured with the microphone set at position M1 (see Figure 11) which represents the maximum distance between the listener and that source for the considered path. The time and frequency envelope of the SRIR is then modified dynamically to emulate the relative source-listener distance along the walking path. The real-time modifications include time delay, level, and spectral changes, applied to different segments of the impulse response. Thanks to the SRIR encoding into the HOA domain, the rotation of the listener may be easily compensated for in real-time before being decoded in binaural mode. The expected advantage of this method is that it exploits the SRIR measured in the room thus reflecting the actual characteristics of the room acoustics. However, in contrast with the GA method, the space-time distribution of early reflections is not updated according to the position of the listener and source in the room, which limits the auralization accuracy.

5.3.1 SRIR Manipulations

The propagation delay was applied in real-time to the direct sound, early reflections, and late reverberation sections, according to the relative distance between the source and the listener. Filtering was applied to the early reflections and late reverberation segments based on Barron's revised theory (Barron & Lee, 1988; Jot et al., 2021). The revised theory takes into account the fact that whereas the reverberation level is assumed to be constant in the room, it exhibits a spatial dependency when counted from the arrival time of the direct sound. This property is expressed through the following formula, which links the frequency-dependent reverberated energy $E(f, r_n)$ observed at distance r_n with respect to the energy $E(f, r_{ref})$ measured at distance r_{ref} :

$$\frac{E(f, r_n)}{E(f, r_{ref})} = exp(\frac{-(r_n - r_{ref}) * 0.04}{RT_{60}(f)})$$

where $E(r_n)$ - the energy of time segment in distance n,

$E(r_{ref})$ - energy of time segment in reference point,

r_n - distance to the n point,

r_{ref} - distance to the reference point.

The rotation of the resulting HOA soundfield is applied after the convolution with the stimulus according to the rotation of the listener's head provided by the tracking system.

5.4 HOA to Binaural Decoding and Equalization

Both methods are delivering their output under a 4th order HOA format which needs to be decoded into binaural signals. The HOA soundfield was first decoded on a set of 24 virtual loudspeakers evenly distributed on the sphere (slightly sub-optimal compared to the theoretical 25 loudspeakers required for 4th-order HOA streams). Then, each loudspeaker channel was filtered with the corresponding HRTF of the KU 100 dummy head available from the HRTF Bili database (Carpentier et al., 2014).

5.5 Compensation Filters

Several compensation filters were applied to each segment of the synthesized RIR to compensate for measurement and reproduction chain components. The GA auralization included a filter compensating for the encoding process of FDN into 4th-order HOA and then decoding virtual speakers and binaural output. The SRIR auralization included compensation for the diffuse field response of the EM32 microphone and a compensation filter for the decoding process of SRIR into virtual speakers and binaural output.

5.5.1 Measurement of the Headphones and Implementation of the Filter

During the subjective listening tests, participants were wearing special open headphones (AKG K1000) with transducers distant from the ear pinnae, in order to guarantee high acoustic transparency of real sound sources. A filter was applied to compensate for the headphone-related transfer function (HpTF). As individualized measurements of participants were not possible to be recorded, this HpTF was averaged from measurements conducted on the KU 100 dummy head. Measurements were averaged (after multiple repositioning of the K1000 headphones) and a small compression factor was applied to limit artifacts linked to spectral spikes. In order to ensure that the signal delivered to the participant's ears is not affected by the headphone's frequency response, a filter was applied to compensate for the headphone transfer function - which includes the headphone transfer function as well as the on-ear morphology and fitting. Due to the fact that individualized measurements of participants were not possible, multiple measurements of headphone refitting on the KU 100 dummy head were taken.

5.6 Calibration and Objective Comparison

For both methods, the different time sections are simulated separately. Hence, it was possible to calibrate their energy level with regard to a reference point. This was done considering the KU 100 measurement for point M16, which corresponds to the shortest listener-to-loudspeaker distance (see Figure 11). Thanks to this, it was then possible to check the evolution of the energy levels along the walking path for each time section and to compare them with the measurements (see Figure 47 and section 4). For the GA auralization, the evolution is very close to the measurements except for the direct sound on points M1 to M6, which show a slight underestimation (up to -2.6 dB). For the SRIR auralization, the late reverberation is slightly overestimated (+1.7 dB), while the early reflection sections are underestimated for short distances (up to -2.2 dB).

6 Subjective Listening Tests

The main goal of the listening test was to evaluate audio clips played either through real speakers positioned in the room or through virtualized ones on headphones. During each trial in the walking phase, participants walked forward and back following a line drawn on the floor. While they were walking, an audio clip was played once during the way forward from a given real or virtualized loudspeaker and repeated during the way back but from a different real or virtualized loudspeaker. After each trial, participants answered a short questionnaire to rate the two audio clips in terms of their respective plausibility and other audio quality attributes. During the standing phase, participants were standing on point M07 (refer to Figure 11) while listening to the stimuli. Everything else followed the same procedure as in the walking phase.

The four loudspeakers positioned in the room were grouped into two pairs. For each trial, the audio clips were played consecutively on the two loudspeakers of a given pair (A-C or B-D, see Figure 11). Pair A-C represents loudspeakers which provide a similar listening perspective in relation to the room and to the walking path. Thus, the influence of the loudspeaker position on the plausibility and other attributes rating of the two audio clips is expected to be minimal. Pair B-D represents a very different loudspeaker perspective (in terms of distance as well as orientation). Thus the influence of the loudspeaker position on the plausibility and other sound attribute ratings of the two audio clips may be higher. Participants were instructed to maintain a straight head position but were allowed to make small movements in case they were required to see the loudspeaker that was playing.

6.1 Experimental Setup

During the experiment, participants were wearing AKG K1000 open headphones with a small tracking device attached (see Figure 17). The tracking system was implemented using the HTC Vive Pro system. A small sensor - Vive Tracker attached to the top of the headphones - allowed to track participants' rotation as well as absolute position. The system employed four infrared cameras mounted in the corners of the room to track the position of the sensor.

In order to help participants adjust their walking speed to the stimuli duration, two iPads were set on each end of the path, which displayed simple visual signs indicating the time to start walking, rotate, or stop.

6.2 Stimulus Choice and Preparation

In order to limit the test duration and the fatigue of participants, only one audio clip was used. The sound stimulus was a 10-second long excerpt from the anechoic recording of a male voice reading short sentences in English. The stimulus was processed in real-time using the above-described auralization methods for the playback on headphones or played back directly from one of the four real loudspeakers standing in the experimental room.

6.3 Conditions

There were 28 combinations of the stimuli pairs presented during the forward and backward walk during the walking phase (see Table 6) and while standing in the standing phase. The conditions included 2 pairs of loudspeaker positions (pair A-C or B-D), 2 orders of playback within a given loudspeaker pair, and 7 combinations of rendering methods: R-SRIR, SRIR-R, R-GA, GA-R, R-R, GA-GA, SRIR-SRIR. From these 28 combinations, 20 were presented twice. The eight conditions with two auralizations (GA-GA and SRIR-SRIR) were not repeated to limit the test duration. All of the trials were randomized for each participant. The average length of the experiment was 75 minutes.

Table 6

Conditions used in the listening test. The table presents the loudspeaker position and rendering method used during forth and back walk along the path for each condition.

Nr	Forth	Back	Nr	Forth	Back
1	A [SRIR]	C [R]	15	B [R]	D [GA]
2	C [SRIR]	A [R]	16	D [R]	B [GA]
3	B [SRIR]	D [R]	17	A [R]	C [R]
4	D [SRIR]	B [R]	18	C [R]	A [R]
5	A [R]	C [SRIR]	19	B [R]	D [R]
6	C [R]	A [SRIR]	20	D [R]	B [R]
7	B [R]	D [SRIR]	21	A [SRIR]	C [SRIR]
8	D [R]	B [SRIR]	22	C [SRIR]	A [SRIR]
9	A [GA]	C [R]	23	B [SRIR]	D [SRIR]
10	C [GA]	A [R]	24	D [SRIR]	B [SRIR]
11	B [GA]	D [R]	25	A [GA]	C [GA]
12	D [GA]	B [R]	26	C [GA]	A [GA]
13	A [R]	C [GA]	27	B [GA]	D [GA]
14	C [R]	A [GA]	28	D [GA]	B [GA]

6.4 Participants

Participants were recruited from students of Sorbonne University and the IRCAM network. Within the recruitment email, they were provided with a consent form and an experience information leaflet (see Appendix A). The two phases of the experiment (standing and walking) were conducted separately with three months period in between. The time between phases was due to the studio's availability. Thirty-three participants with self-reported normal hearing, with a median age of 29 (min 18, max 47, 23 men, 10 women) took part in the walking phase. Twenty-five participants with self-reported normal hearing, with a median age of 30 (min 19, max 47) took part in both the walking and standing phases of the experiment. All participants were expert listeners or students of sound engineering programs.

Inclusion criteria:

- Age between 18 and 65 years
- Normal hearing, Normal seeing (with or without correction)
- Audio expertise

Exclusion criteria:

- Presence of hearing problems leading to a significant decrease in hearing acuity
- Visually impaired

6.5 Invitation Procedure

An invitation email including the experimental procedure and consent forms was sent to all groups of subjects. When arriving at the experimental room (Studio 5 at IRCAM) for the first part of the test they were asked to read the information notice and then sign the consent form and fill out a short questionnaire about their age, sex, occupation, and expertise in audio and specifically spatial audio. Invitation notice is provided in Appendix A.

6.6 Collection Method

In the several studies on the plausibility of the virtual sound source, there are two basic approaches involved. One relies on the yes/no paradigm where the participants choose if the sound is real or not (Brinkmann et al., 2017), and the second one takes more detailed answers by using a short questionnaire (Neidhardt et al., 2018; Wirler et al., 2020). The dissertation study is focused on a comparison of the plausibility of two different auralization methods with real sources. That is why a questionnaire method seems more appropriate for this type of research as it will not only allow to quantification of the differences between plausibility ratings for different models but also correlate plausibility with other spatial attributes of sound. The questionnaire aimed to identify the perception of plausibility, localization accuracy, externalization, reverberation, and timbre differences between real and virtual sources. The attributes were chosen based on the previous studies on the subject (Neidhardt & Knoop, 2017; Neidhardt et al., 2018; Wirler et al., 2020) and Spatial Audio Quality Inventory which suggests a vocabulary for evaluation of virtual auditory environments (Lindau et al., 2014). Participants were rating the attributes of different stimuli on a visual scale with the exception of localization and externalization where the participants were using a simple graph to indicate the position of sound. The user interface for the test was implemented in Max/MSP. Participants responded to the questionnaire on a laptop and were guided with a short explanation of the different questions.

6.6.1 Questionnaire

Subjects were answering questions as written below:

- PLAUSIBILITY

  For each audio clip, rate the plausibility that it was actually played by one of the loudspeakers (6-Very plausible – 0-Not at all plausible)

- LOCALIZATION

  Drag two red circles to indicate the localization of audio clips 1 and 2. In case the sound was coming not exactly from the loudspeaker but very close to it - you can put it in the area around the speaker. If the sound was localized even further from the speaker, you can put it anywhere in the picture (refer to Figure 18).

- BLUR

  Rate how precise was the localization of the 1st and 2nd audio clip (Scale: 6-Blurred – 0-Focused). The scale for blur was reversed in comparison to the plausibility scale in order to restrain participants from assigning the same value to subsequent questions to make the evaluation process easier.

- EXTERNALIZATION

Choose the area that matches the externalization of the 1st and 2nd audio clip (Boxes: “Inside the head” - 0, “Close to the head” - 1, “Outside the head”- 2).

- TIMBRE

  Rate the difference of timbre between the two audio clips (Scale: 6 - Very different – 0 - Not different)

- REVERBERATION

  Rate the difference of room reverberation between the two audio clips (Scale: 6 - Very different – 0 - Not different)

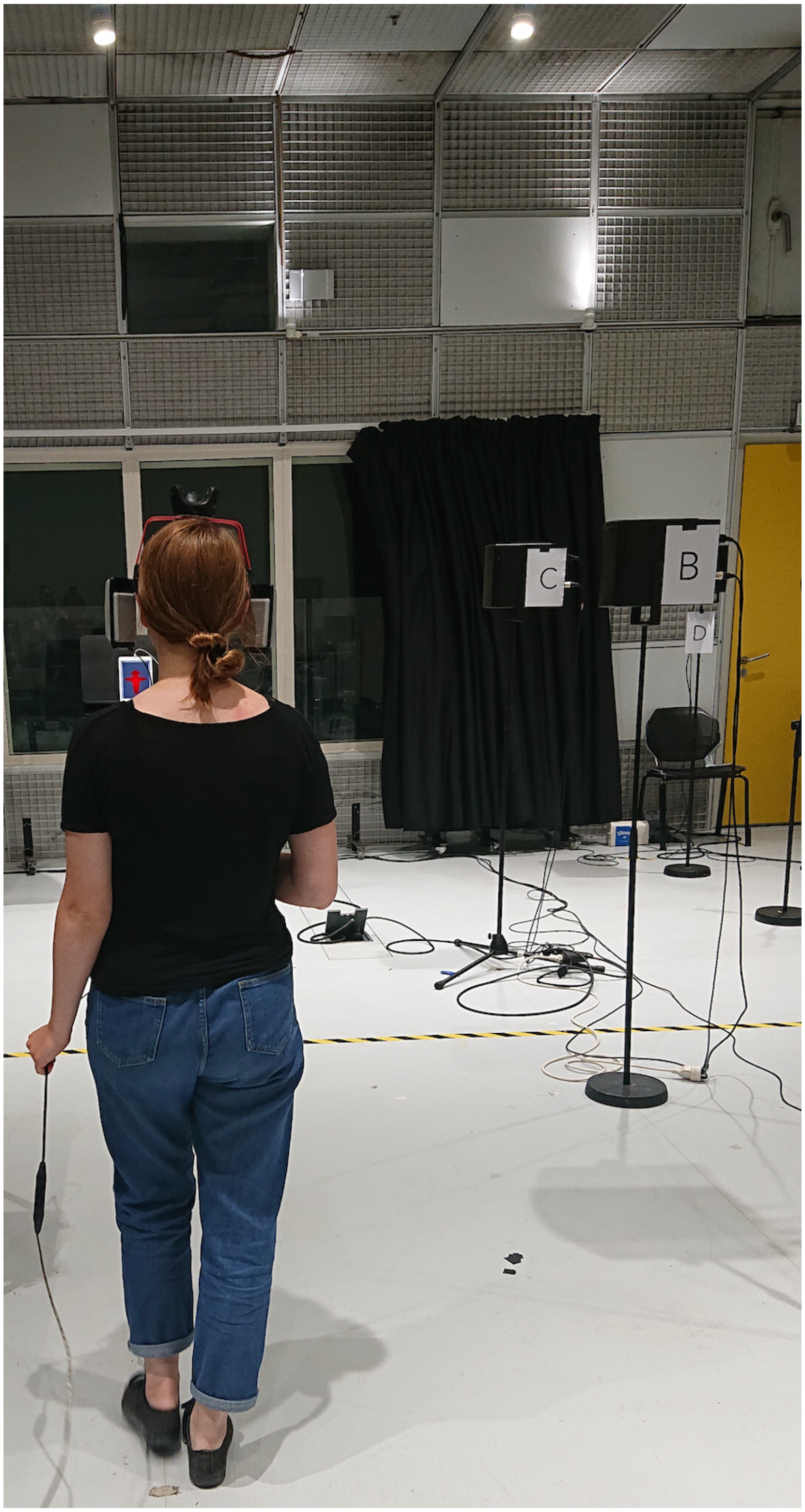


Figure 17: Participant wearing AKG K1000 headphones with Vive Tracker during listening test

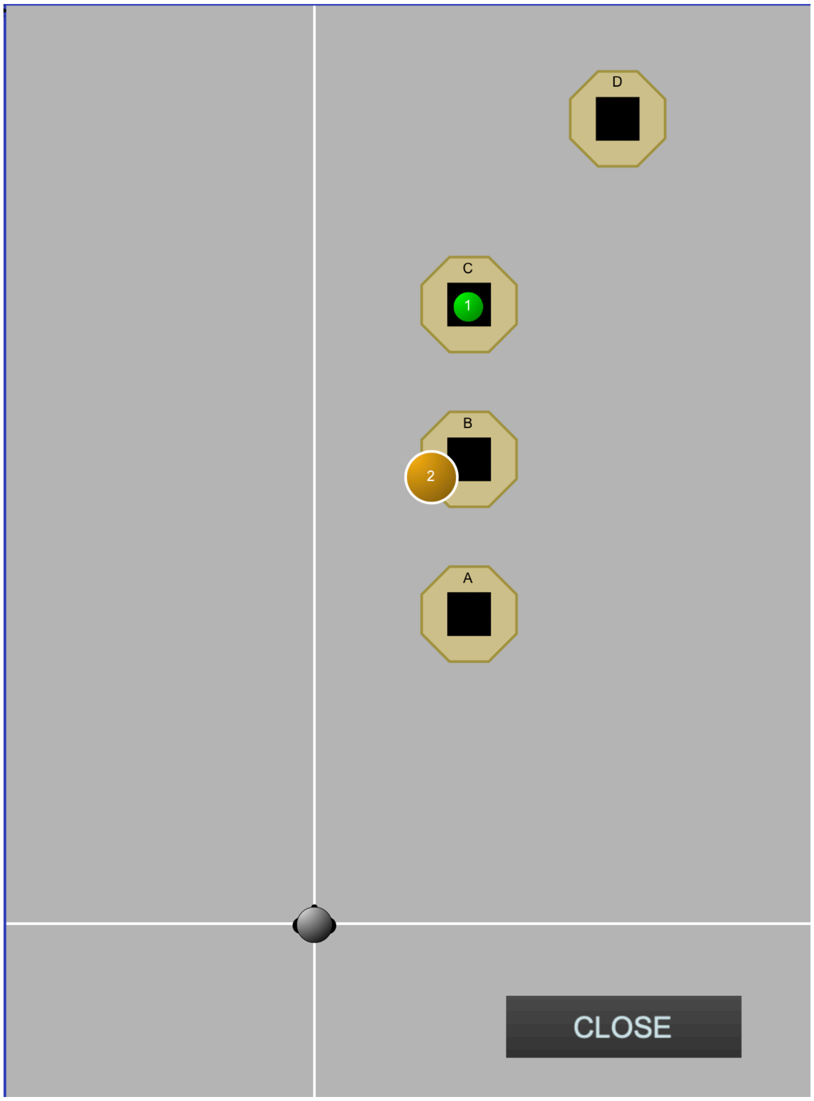


Figure 18: Screenshot of the graphic interface used for question about localization

CHAPTER VI

RESULTS OF THE EXPERIMENT - WALKING PHASE

The main goal of the dissertation study is to explore the perceptual and acoustic factors influencing the plausibility perception. The results of the analysis described in Chapters VI and VII focus first on the perceptual factors allowing investigation of the complex phenomenon of plausibility perception. This investigation is followed by the analysis of acoustic parameters of auralizations in Chapter VII aimed to uncover the acoustic factors that significantly influence judgments of plausibility. Through this investigation, we aim to find the links between acoustic characteristics and subjective ratings.

In order to investigate the influence of perceptual factors on plausibility assessment the statistical analysis specifically focuses on answering the following research question:

- How do loudspeaker position and method of rendering affect plausibility and other attributes' judgment in the walking phase?

- What other factors impacted plausibility and other attributes' judgment in the walking phase?

- How does participants' movement affect plausibility and other attributes' judgment?

- What is the correlation between plausibility and other attributes?

- What was the effect of speed of walking and yaw movement on perceptual evaluation? Was the evaluation affected by participants' behavior or was the participants' behavior indicator of the stimulus characteristics?

Consequently, the main goal of the analysis is to investigate the influence of the main and secondary effects on the ratings obtained during the experiment. The main and secondary effects are defined by the research questions posed above. A secondary goal is to explore correlations between plausibility and other attributes of the questionnaire to gain insight into the complex phenomenon of plausibility perception. Advanced statistical techniques Generalized Linear Mixed Modeling (GLMM) and Linear Mixed Modeling (LMM) were chosen as the main analysis methods.

1 Analysis Method

GLMM and LMM are both advanced statistical techniques used for analyzing data with both fixed and random effects, making them suitable for complex study designs. LMM is specifically tailored for continuous dependent variables and assumes a normal distribution of the response variable. On the other hand, GLMM extends the flexibility of LMM by being suitable for a broader range of response variable types, such as binary or categorical data. The statistical methods allow for the inclusion of random intercepts to account for individual differences in the internal scales of participants and other subjective scores. Furthermore, control parameters that might account for variance in the data are added to the model. Implementing GLMM and LMM allowed for avoiding simply averaging values for the repeated trials implemented in the study design. GLMM was performed in R (version 4.0.2) and RStudio (version 2021.09.1) using the lmer4 package. Post-hoc comparisons were computed using the package emmeans.

In each analysis, an empty model was generated first, which contains only random intercepts for participants. Next, the main contrast conditions within the experiment were added as fixed factors, including all possible interactions. For each generalized linear mixed models analysis, the Akaike information criterion (AIC) was used to select the best candidate to explain the variance. Models were considered different if the difference in AIC was greater than 2, as recommended in the literature (Bozdogan, 1987). For models with an AIC value less than 2, the simpler model was chosen. Selected models were analyzed with the mixed and summary functions, and significant interactions were further analyzed using the emmeand and emmeans functions. For all models, a collinearity check was performed using R's vif function and confirmed that factors have values of less than 5. Post-hoc power analyses were run for each main result using the simr package with 1000 simulations to ensure that the results were sufficiently powered.

The plots provided in Chapters VI and VII feature the most critical significance brackets, emphasizing noteworthy differences between result pairs to enhance clarity. For a comprehensive set of pairwise comparison results, refer to Appendices C and D.

1.1 Fixed and Random Effects

The main fixed effects considered in this analysis were the rendering METHOD (3 levels: R, SRIR, GA) and the LOUDSPEAKER position (4 levels: A, B, C, D) - refer to Chapter V for a detailed explanation of the labels. These two elements were most important in the context of the research questions and were explicitly varied in the conditions of the test. Secondary effects were added to the models and were expected to contribute to the explanation of data variability. These include trial index, order of playback, speed of walking, head movement in the yaw axis, height difference between the participant and loudspeaker, as well as answers to the demographic questionnaire, including the number of years of formal musical training and participation in audio tests or spatial audio tests (refer to Appendix A for the details of the questionnaire). Table 7 provides a summary of the fixed effects.

Speed of walking and head movement on the yaw axis were considered as fixed effects but also as dependent variables. This allowed us to first check if these parameters influenced subjective ratings and also if other factors had an influence on walking speed and yaw movement. The speed of walking was obtained by analysis of the tracking data during each of the trials of the walking phase. Speed was calculated based on the time it took participants to walk from 0.3 to 7.2 m from the starting point. This way, the speeding up at the start of the path and slowing down at the end were not taken into account. The amplitude of head movement along the yaw axis was obtained by analysis of the tracking data for each of the trials in the standing phase.

The random effect of the participant ID was included in each model during analysis. This allowed us to account for differences between the internal scales of participants.

Table 7

Summary of independent variables

Name	Levels	Range	Description
Phase	2	Walking or standing	Phase of the experiment
Loudspeaker position	4	A, B, C, or D	Loudspeakers positions always associated in pairs A-C or B-D during one trial
Playback method	3	Real playback, SRIR, or GA auralizations	Type of rendering method used to present the stimulus
Playback method pair	5	R-R, R-SRIR, R-GA, SRIR-SRIR, or GA-GA	Pair of rendering methods used within one trial
Order of playback	2	1 or 2	Stimulus presented as first or second within one trial
Trial index	4	1-12, 13-24, 25-36, or 37-48	There were 48 trials in the experiment during each phase. To check the influence of time and fatigue, the trials were divided into 4 sections
Height difference	Continuous	0-Inf	Absolute height difference between the level of the participant's ears and the center of the loudspeaker
Speed of walking	Continuous	0-Inf	Speed of walking during stimulus playback
Yaw movement	Continuous	0-Inf	Amplitude of yaw movement during stimulus playback
Years of music training	Continuous	0-Inf	Answer to the question ”How many years of musical training have you received?”
Audio tests	2	0 or 1	Answer to the question ”Have you participated in audio tests before?”
Spatial audio tests	2	0 or 1	Answer to the question ”Have you participated in spatial audio tests before?”

Table 8

Summary of dependent variables

Name	Type	Range/Levels	Description
Plausibility	0-6	Absolute	Trial plausibility score
Blur	0-6	Absolute	Trial blur score
Localization error	0-Inf	Absolute	Ratio between minimum distance to the loudspeaker from the walking path or standing point and distance of answer point to the target loudspeaker (see Section 1.2)
Loudspeaker recognition rate	0-1	Absolute	See Section 1.2
Externalization	0-1	Absolute	See Section 1.2
Reverberation difference	0-6	Relative	Difference of reverberation rating between the two stimuli during one trial
Timbre difference	0-6	Relative	Difference of timbre rating between the two stimuli during one trial
Plausibility difference	0-6	Relative	Difference of plausibility rating between the two stimuli during one trial calculated from the two absolute ratings
Blur difference	0-6	Relative	Difference of blur rating between the two stimuli during one trial calculated from the two absolute ratings
Externalization difference	0-6	Relative	Difference of externalization rating between the two stimuli during one trial calculated from the two absolute ratings
Localization error difference	0-6	Relative	Difference of localization error score between the two stimuli during one trial calculated from the two answers
Speed of walking	Continuous	Absolute	Walking speed during stimulus playback
Amplitude of yaw movement	Continuous	Absolute	Amplitude of yaw movement during stimulus playback

1.2 Summary of Dependent Variables

All dependent variables, except for externalization and loudspeaker recognition rate, were continuous, enabling the implementation of the Linear Mixed Model (LMM) for analysis. The categorical nature of externalization and loudspeaker recognition rate required the application of the Generalized Linear Mixed Model (GLMM) for analysis. Table 8 provides a summary of all dependent variables used in the analysis (refer to Section 6.6.1 in Chapter V for the details of the questionnaire).

The following section presents additional details on specific variables:

Externalization - Participants provided externalization ratings by selecting one of three areas on the diagram: ”outside of the head” (scored as 2), ”close to the head” (scored as 1), and ”inside the head” (scored as 0). Since externalization is a categorical variable, the GLMM analysis method was applied. Due to a significant bias towards answer 2 in the results, the data was converted into binary format. Responses categorized as 0 and 1 were merged and redefined as 0, as both indicated issues with externalization. The response labeled ”outside of the head” was redefined as 1. This adjustment allowed the analysis to focus on predicting the plausibility of receiving a response indicating proper externalization of sound.

Localization Error - During the experiment, participants were asked to indicate the position of the sound source in relation to the room and loudspeakers on a simple graph. The graphic interface allowed participants to mark the sound source position by dragging the circle on the graphical representation of the room and loudspeakers. The geometric position of the marked points was recorded as continuous data however in the graphical interface the area within and around the loudspeakers was divided visually to facilitate the answering process. The areas represented three different distances from the loudspeaker: inside the loudspeaker – 0-0.2 m; around the speaker – 0.2-0.4 m; the area outside the loudspeaker – above 0.4 m (refer to Figure 18 in Chapter V).

To obtain the localization error, the distance between the answer point and the center of the target loudspeaker was divided by the minimum distance from which the participant could hear the loudspeaker while walking on the path. For the standing phase, the localization error was calculated by dividing the distance between the answer point and the center of the target loudspeaker by the distance between the target loudspeaker and the standing point.

The calculation of localization error can be approached in various ways. In our case, the method selected was influenced by the recognition that localization error diminishes in perceived significance as distance increases. For instance, a disparity of 10 centimeters observed from the moon would be imperceptible. In the contrary, as participants moved closer to the sound source, their ability to distinguish positional differences increased. However, this methodology is not without its limitations. Firstly, during the walking phase, participants listened to the loudspeakers at varying distances, making it challenging to establish a consistent minimum audible distance. Secondly, a drawback arises from the difficulty in comparing values between phases, as they were normalized using different criteria. Therefore, in the discussion of results, the constraints of the normalization technique are acknowledged and addressed.

Loudspeaker Recognition Rate - During the experiment, participants were asked to indicate the position of the sound source in relation to the room and loudspeakers on a simple graph. The answers were analyzed to determine which loudspeaker was closest to the point marked on the graph and if this loudspeaker was the same as the one playing the stimulus. A correctly recognized loudspeaker was assigned a value of 1, while an incorrectly recognized loudspeaker was assigned a value of 0.

2 Results analysis

The experiment was conducted in two distinct phases: walking and standing, separated by a three-month interval. The time between phases was determined by the availability of the studio space. As described in Chapter V, the two phases were identical, differing only in participants' behavior. In the first, walking phase, participants walked forth and back along the path listening to the same speech excerpt played twice by two different loudspeakers and rendering methods during the forth and back movements. In the second, standing phase, participants remained stationary at point M07 (refer to Figure 46) and listened to the speech excerpt played back successively on two loudspeakers with two rendering methods. This chapter focuses solely on the analysis of the walking phase, which forms the core of this dissertation and enables the comprehensive examination of data from all 33 participants. Chapter VII presents a comparative analysis of the results from both phases, specifically focusing on 25 participants who took part in both phases.

Table 9

Plausibility in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	88.55	1.00	0.00
Method	12.86	2.00	0.00
Ldspkr	216.80	3.00	0.00
Order	7.81	1.00	0.01
Speed	15.67	1.00	0.00
Method:Ldspkr	84.42	6.00	0.00

2.1 Plausibility

The analysis of the results revealed that the best model for predicting plausibility ratings in the walking phase (on a scale of 0 - ’not at all plausible’ to 6 - ’very plausible’) was: LOUDSPEAKER position * playback METHOD + ORDER of playback + SPEED of walking (refer to Table 9). A significant interaction was observed between LOUDSPEAKER position and rendering METHOD ($\chi^2(2) = 114.78, P < 0.001$), with participants rating GA and SRIR methods slightly less plausible compared to the real loudspeaker for loudspeakers A, C, and D (see Figure 19). Specifically, for loudspeaker A, there was a 0.33-point difference for the estimates of SRIR and R method and 0.36 for the estimates of GA and R method. For speaker C, the difference was 0.56 for SRIR and R method, and 0.77 for GA and R method. Importantly, for loudspeaker B, there were no differences in plausibility ratings between GA and SRIR methods and the real loudspeaker. The lowest plausibility ratings were obtained for speaker D for all rendering methods, with the SRIR method at 3.32 and the GA method even lower at 2.79. Furthermore, real loudspeaker D was rated significantly less plausible than real loudspeaker B, with a difference of 0.33

points. A more subtle difference was observed between the real loudspeakers B and C. This difference is not statistically significant for GA and SRIR methods.

The estimate for the SPEED of walking (see Figure 20) was 1.57 $\pm$ 0.4 points ($\chi^2(1) = 15.67, p < 0.001$), indicating that increasing speed by 0.1 m/s was associated with the increase in plausibility rating by 0.157 points.

The influence of the ORDER of playback was also statistically significant ($\chi^2(1) = 7.81, P < 0.01$), but the difference was minimal, with only a 0.13-point (2.2%) decrease for the stimulus played second.

Refer to Table 29 in Appendix C for model selection analysis. The results of the pairwise comparison are provided in Tables 30-31 in Appendix C.

2.2 Blur

Table 10

Blur in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	65.32	1.00	0.00
Method	8.78	2.00	0.01
Ldspkr	482.93	3.00	0.00
Index_id	24.78	3.00	0.00
Order	11.07	1.00	0.00
Speed	16.31	1.00	0.00
Method:Ldspkr	76.86	6.00	0.00

During the experiment, participants assessed the blurriness of the stimuli on a scale ranging from 0 (very focused) to 6 (very blurry). The blur scale was reversed in comparison to the plausibility scale having the ”best” rating at zero. This was done to avoid applying the same value to different attributes to speed up the evaluation.

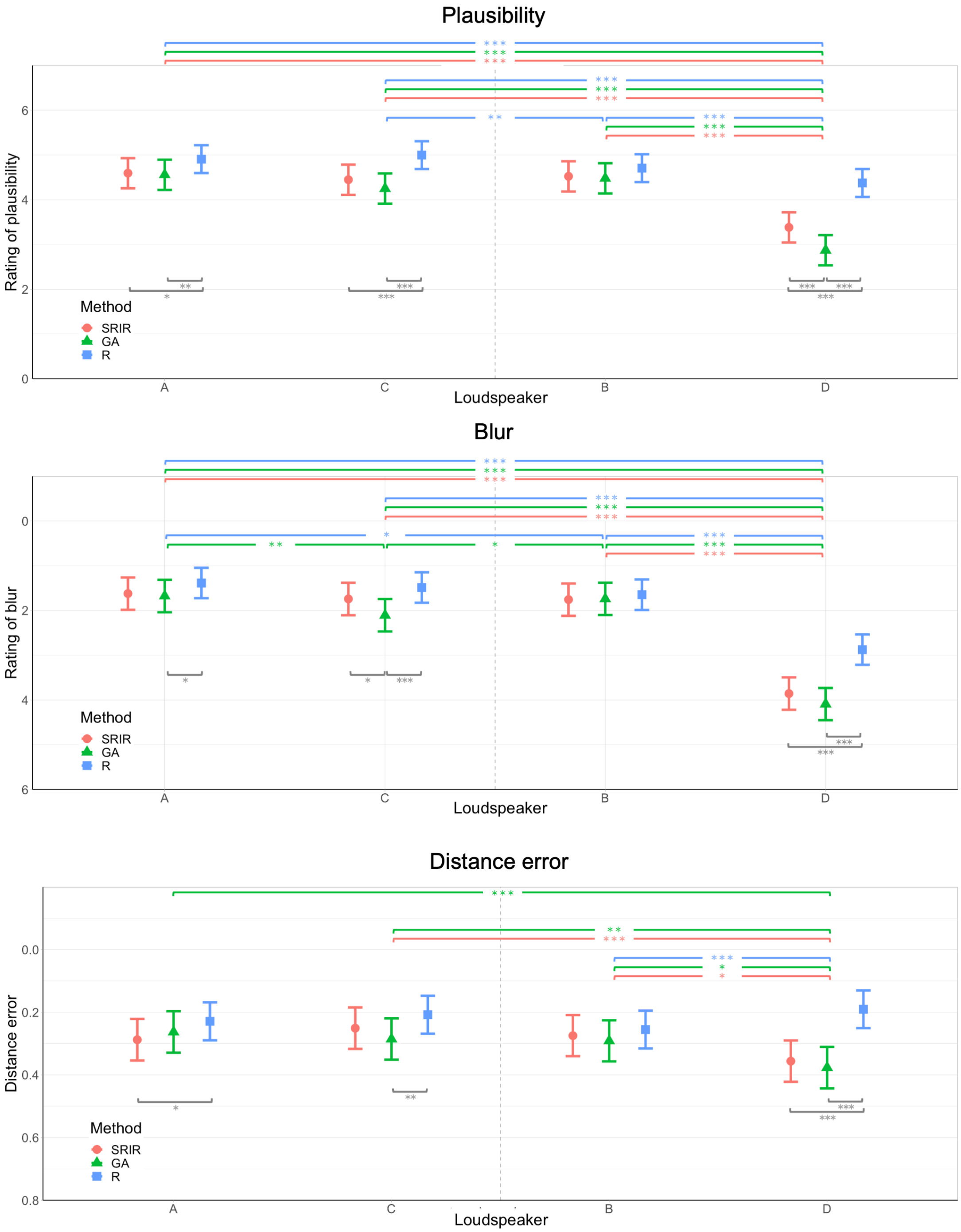


Figure 19: Predicted means and 95% confidence intervals for plausibility (scale: 0 - not at all plausible, 6 - very plausible), blur (scale: 0 - very focused, 6 - very blurry) and localization error ratings in walking phase according to loudspeaker position and rendering method (***$P < 0.05$).

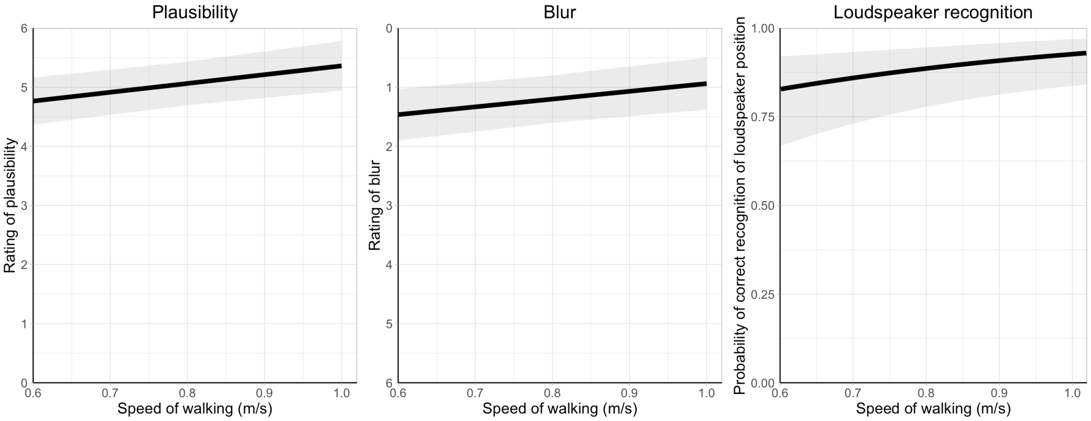


Figure 20: The influence of walking speed on plausibility, blur ratings, and percentage of correctly recognized loudspeaker positions (***P $<$ 0.05)

However, in the analysis, to ensure consistency in comparing different attributes, the y-scale for blur ratings on the plots is reversed to match the plausibility scale.

Results indicated that the best model for predicting blur ratings in the walking phase was: LOUDSPEAKER position * playback METHOD + SPEED of walking + trial INDEX + ORDER of playback. A significant interaction was observed between LOUDSPEAKER position and rendering METHOD ($\chi^2(2) = 76.86, P < 0.001$), where ratings were lower for GA and SRIR methods than for the REAL playback for loudspeakers A, C, and D (see Figure 19). Loudspeaker A rendered with the GA method was rated as slightly more blurry than the REAL loudspeaker (difference of 0.33 points). The difference was more pronounced between GA and R for loudspeaker position C, with a gap of 0.62 points. In contrast with plausibility evaluation where GA and SRIR methods were ‘equally’ different from the REAL loudspeaker, for blur the results of the GA method were more often statistically significantly different from the REAL loudspeaker. Loudspeaker D received the lowest blur ratings for all three rendering methods. However, both GA and SRIR were rated significantly more blurry than the REAL loudspeaker, with differences of 1.03 between SRIR and REAL, and 1.29 between GA and REAL. Noticeably, as for the plausibility results, there were no significant differences between blur ratings for auralizations and the REAL loudspeaker for loudspeaker B.

The estimate for the SPEED of walking (see Figure 20) was 1.57 $\pm$ 0.39 points ($\chi^2(1) = 16.3, p < 0.001$), indicating that increasing speed by 0.1 m/s was associated with more focused sound by 0.157 points.

The influence of the trial INDEX was statistically significant ($\chi^2(3) = 24.78, P < 0.001$), but the difference was minimal. Stimuli were rated as less blurry with subsequent trials, showing a 0.29-point difference between the first (1-13) and last (37-48) sections of trials.

The influence of the ORDER of playback was also statistically significant ($\chi^2(1) = 11.07, P < 0.001$). There was a 0.15 point (2.5%) decrease in blurriness for the second stimulus.

Refer to Table 32 in Appendix C for model selection analysis. Pairwise comparison results are provided in Tables 33-35 in Appendix C.

2.3 Localization Error

A detailed description of the localization error calculation is provided in Section 1.2. The best model for predicting localization error was: LOUDSPEAKER position * playback METHOD + ORDER of playback + trial INDEX. A significant interaction was observed between LOUDSPEAKER position and METHOD of playback ($\chi^2(6) = 37.62, P < 0.001$).

Table 11

Localization error in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	61.05	1.00	0.00
Method	6.27	2.00	0.04
Ldspkr	19.11	3.00	0.00
Index_id	15.21	3.00	0.00
Order	23.54	1.00	0.00
Method:Ldspkr	37.62	6.00	0.00

No statistically significant differences were found between the rendering methods for loudspeaker B, mirroring the findings for plausibility and blur ratings (see Figure 19). However, loudspeaker A exhibited a significant difference between SRIR and R methods, while loudspeaker C showed a difference between GA and R methods. Loudspeaker D recorded the largest localization error for both GA and SRIR methods. In particular, both simulations demonstrated significantly larger localization errors than the REAL loudspeaker: the difference between SRIR and R was 0.186 m, and between GA and R was 0.165 m. Interestingly, the REAL playback of loudspeaker D obtained a statistically significantly lower localization error than loudspeaker B which might be due to the normalization. However, for methods GA and SRIR the effect was opposite. Loudspeaker D obtained a larger localization error than all other loudspeakers for method GA and larger than loudspeakers B and C for method SRIR.

The influence of the ORDER of playback was statistically significant ($\chi^2(1) = 23.54, P < 0.001$), with the localization error being 0.047 m smaller for the stimulus played second.

The influence of the trial INDEX was also statistically significant ($\chi^2(3) =$

$15.21, P = 0.0016$), although the difference was minimal. The localization error was larger for the first section of trials (1-12) than for each of the other trial sections (refer to Table 39).

Refer to Table 36 in Appendix C for model selection analysis. Pairwise comparison results are provided in Tables 37-39 in Appendix C.

Figure 21 illustrates the distribution of localization points for the position question, according to the three playback methods for each of the loudspeakers. The plots also show centroids and ellipses representing 95% confidence intervals. The responses for loudspeakers B and C were very similar across the three playback methods, with centroids and means indicating that the majority of answer points were localized close to the center of the loudspeakers. Loudspeaker A exhibited a slightly larger centroid for the SRIR method, which explains the plots of localization error where the SRIR method has statistically significantly larger localization error than REAL playback (see Figure 19). On the contrary, loudspeaker D exhibited a larger centroid for the SRIR AND GA methods compared to REAL playback. Answer points for the GA and SRIR methods were mostly localized in front and on the left side of the loudspeaker. For all the loudspeakers, some answer points were marked in the place of another loudspeaker. The section below describes the analysis of these errors, referred to as errors of loudspeaker recognition.

2.4 Loudspeaker Recognition Rate

During the experiment, participants were asked to indicate the position of the sound source in relation to the room and loudspeakers on a simple graph. Responses were

Table 12

Loudspeaker recognition rate in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	0.02	1.00	0.89
Order	3.95	1.00	0.05
Method	3.51	2.00	0.17
Ldspkr	24.66	3.00	0.00
Index_id	23.26	3.00	0.00
Speed	6.04	1.00	0.01
Order:Method	1.23	2.00	0.54
Order:Ldspkr	5.98	3.00	0.11
Method:Ldspkr	33.91	6.00	0.00
Order:Method:Ldspkr	15.44	6.00	0.02

analyzed to determine which loudspeaker was closest to the point marked on the graph. Correctly recognized loudspeakers were assigned a value of 1, while incorrectly recognized loudspeakers received a value of 0. There were a total of 334 errors in loudspeaker recognition out of 3146 trials, accounting for 10.6%.

The best model for predicting errors of loudspeaker recognition was: ORDER of playback * playback METHOD * LOUDSPEAKER position + trial INDEX + SPEED of walking. The analysis revealed a significant interaction between ORDER, playback METHOD and LOUDSPEAKER position ($\chi^2(6) = 15.44, P$<0.001) – refer to Table 12 for details. Generally, loudspeakers A, B, and D were better recognized when played during the return walk, whereas loudspeaker C was better recognized when played during the forward walk (see Figure 22). The most notable difference between playback methods was observed for loudspeaker D, where GA and SRIR methods yielded significantly fewer correct answers compared to the REAL loudspeaker when played during the forward walk.

The influence of the trial INDEX was statistically significant ($\chi^2(3) = 23.26, P$<0.001). The estimate of correct answers was 83% for trials 1-12, 87% for trials 13-24, 88% for trials 25-36, and 87% for trials 37-48. There were significantly more errors in loudspeaker recognition in the first section of trials compared to all subsequent sections.

The influence of walking SPEED was also statistically significant ($\chi^2(1) = 6.04, P = 0.006$). The estimate for the speed of walking was 2.4 ± 0.1%, implying that increasing the speed by 0.1 m/s was associated with the increase of the probability of correctly recognizing loudspeaker positions by 0.04.

Figure 23 illustrates the distribution of recognition errors, indicating which loudspeaker was perceived as closest to the marked point. The barplot reveals that for loudspeaker A, the majority of errors indicated loudspeaker C, while for loudspeaker C, the reverse was true, with most errors pointing to loudspeaker A. Lower number of errors was misidentifying loudspeakers A and C as B. Similarly, both loudspeakers B and D had the highest number of errors indicating loudspeaker C.

Refer to Table 43 in Appendix C for model selection analysis. Pairwise comparison results are provided in Tables 44-46 in Appendix C.

2.5 Externalization

Participants provided externalization ratings by selecting one of three areas on the diagram: ”outside of the head” (scored as 2), ”close to the head” (scored as 1), and ”inside the head” (scored as 0). Given that externalization is a categorical variable, analysis based on generalized linear mixed models had to be applied. Responses

Table 13

Externalization in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	17.42	1.00	0.00
Method	14.59	2.00	0.00
Ldspkr	8.75	3.00	0.03
Order	62.78	1.00	0.00
Method:Ldspkr	16.62	6.00	0.01

categorized as 0 and 1 were merged and redefined as 0 because both indicated issues with externalization. The response labeled ”outside of the head” was redefined as 1. This adjustment allowed the analysis to focus on predicting the plausibility of receiving a response indicating ”outside of the head”.

The best model to predict the probability of proper externalization was: playback METHOD * LOUDSPEAKER position + ORDER of playback (see Table 13 and Figure 24). There was a significant interaction between the main effects of the playback METHOD and LOUDSPEAKER position ($\chi^2(6) = 16.62, P$=0.01), as well as the order of playback ($\chi^2(1) = 72.81, P$<0.001). The REAL loudspeaker had a higher probability of proper externalization compared to both GA and SRIR methods for all of the loudspeakers except method GA for loudspeaker A. The maximum difference occurred for loudspeaker B between methods SRIR and REAL - 0.11. Loudspeaker D also obtained a statistically significantly higher probability of proper externalization than loudspeakers A, B, and C for the SRIR and REAL methods.

The influence of the ORDER of playback was statistically significant ($\chi^2(1) = 62.78, P$<0.001), with the predicted probability of proper externalization being 0.12 lower for the stimulus played second.

Refer to Table 40 in Appendix C for model selection analysis. The results of the pairwise comparison are provided in Tables 41-42 in Appendix C.

2.6 Timbre Difference

Table 14

Timbre difference ratings in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	47.17	1.00	0.00
Method_pair	80.17	4.00	0.00
Ldspkr_pair	274.92	1.00	0.00
Index_id	9.26	3.00	0.03

During the experiment, participants rated the timbre difference within each pair of loudspeakers A-C or B-D on a scale of 0 (not different) to 6 (very different). Analysis of the results showed that the best model for predicting timbre difference ratings was: LOUDSPEAKER PAIR + playback METHOD PAIR + trial INDEX (see Table 14 and Figure 25). There was a significant main effect of the LOUDSPEAKER PAIR ($\chi^2(1) = 274.92, P<0.001$). The timbre difference for pair A-C (1.74) was rated statistically significantly lower than for pair B-D (2.79).

During the test, participants listened to pairs of stimuli with different combinations of playback methods. The possible pairs of methods were: R-R, SRIR-R, GA-R, SRIR-SRIR, GA-GA. There was a significant influence of the playback METHOD PAIR on the ratings of timbre difference ($\chi^2(4) = 80.17, P<0.001$). The lowest rating was obtained by a pair of two REAL loudspeakers (R-R, 2.29) and two SRIR auralizations (SRIR-SRIR, 2.26). The highest timbre difference rating was observed in the GA-R pair

(3.03). Pairs SRIR-R and GA-GA were rated similarly, with the mean estimate slightly higher than for pair R-R (2.79, and 2.64, respectively).

The influence of the trial INDEX was also statistically significant ($\chi^2(3) = 9.26, P = 0.026$), but the difference was minimal. The only statistically significant difference in timbre difference ratings was between the second and third sections of trials. The estimate for ratings for each trial section is as follows: trials 1-12 2.79, 13-24 2.92, 25-36 2.67, and 37-48 2.68.

Model selection analysis is provided in Table 14 in Appendix C. Refer to Tables 48-50 in Appendix C for pairwise comparison results.

2.7 Reverberation Difference

Table 15

Reverberation difference ratings in walking phase: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	50.96	1.00	0.00
Method_pair	49.50	4.00	0.00
Ldspkr_pair	788.66	1.00	0.00

Similarly to timbre difference, participants rated the reverberation difference within each pair of loudspeakers A-C or B-D on a scale from 0 (not different) to 6 (very different). The best model for predicting reverberation difference ratings was: LOUDSPEAKER PAIR + playback METHOD PAIR (refer to Table 15 and Figure 25). There was a significant main effect of the LOUDSPEAKER PAIR ($\chi^2(1) = 788.66, P$<0.001). The reverberation difference for pair A-C (1.25) was rated statistically significantly lower than for pair B-D (2.89).

There was a significant influence of the playback METHOD PAIR on the ratings of reverberation difference ($\chi^2(4) = 49.5, P$<0.001). The lowest rating was obtained by a pair of two REAL loudspeakers (R-R, 2.65) and two SRIR auralizations (SRIR-SRIR, 2.76). The highest rating of reverberation difference was observed for pair GA-R and GA-GA (3.23, 3.19 respectively). The SRIR-R pair was rated similarly, with the mean estimate slightly lower than the GA-R pair (3.00).

Refer to Table 51 in Appendix C for model selection analysis. The results of the pairwise comparison are provided in Tables 52-53 in Appendix C.

2.8 Plausibility Difference

Table 16

Plausibility difference: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	48.91	1.00	0.00
Method_pair	37.82	4.00	0.00
Ldspkr_pair	153.33	1.00	0.00

To compare the results of timbre difference evaluation with plausibility ratings, for each trial, the plausibility difference was calculated. The plausibility difference was obtained by taking the absolute value of the difference between the plausibility ratings for each stimuli pair. After that, a separate analysis was performed to allow for a direct comparison of plausibility difference with timbre and reverberation difference results (see Figure 25).

The best model for predicting plausibility difference was: the LOUDSPEAKER PAIR + playback METHOD PAIR (see Table 16 and Figure 25). There was a significant main effect of the LOUDSPEAKER PAIR ($\chi^2(1) = 153.3, P$<0.001). The results for plausibility difference were similar to the timbre difference evaluation. The ratings for pair A-C and B-D were significantly different, with pair A-C obtaining a lower difference (1.05) than pair B-D (1.82).

There was also a significant influence of the playback METHOD PAIR on the plausibility difference ($\chi^2(4) = 37.8, P$<0.001). The influence of the method was also similar to timbre difference ratings. The lowest difference was obtained by a pair of two REAL loudspeakers (R-R, 1.44). The largest difference in plausibility rating was observed on pair GA-R (2.02). Pairs SRIR-SRIR, SRIR-R, and GA-GA obtained similar values with the mean estimate slightly higher than for pair R-R (1.81, 1.82, 1.86 respectively).

Refer to Table 54 in Appendix C for model selection analysis. Pairwise comparison results are provided in Tables 55-56 in Appendix C.

2.9 Speed of Walking

Table 17

Speed of walking: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	4586.39	1.00	0.00
Method	23.09	2.00	0.00
Ldspkr	12.76	3.00	0.01
Index_id	62.93	3.00	0.00
Order	13.17	1.00	0.00

During the experiment, participants walked along a path with a tracking device attached to their headphones. The ideal speed, where participants walked exactly the time the stimulus was playing, was 0.68 m/s. The average speed of walking for all trials in the walking phase was 0.78 m/s. The speed was calculated based on the time it took participants to walk from 0.3 to 7.2 m from the starting point, excluding the initial speeding up and the slowing down at the end. Figure 26 illustrates the distribution of speed among all trials, while Figure 27 presents the distribution of the mean speed for all trials across participants. Both histograms exhibit Gaussian distribution. The tracking data were analyzed to determine which factors affected the speed of walking during trials.

Analysis of the results showed that the best model for predicting the speed of walking was: trial INDEX + METHOD of playback + ORDER of playback + LOUDSPEAKER position (see Table 17). There was a significant main effect of the trial INDEX ($\chi^2(3) = 62.9, P$<0.001). Participants walked statistically significantly faster during the first 12 trials compared to later trials (see Figure 28).

There was also a significant main effect of the METHOD of playback ($\chi^2(2) = 23.1, P$<0.001). Participants walked statistically significantly faster when the REAL loudspeaker was playing compared to GA and SRIR methods (see Figure 28).

Additionally, there was a significant main effect of the ORDER of playback ($\chi^2(1) = 13.2, P$<0.001). Participants walked statistically significantly faster when moving forward by 0.007 m/s.

There was also a significant main effect of LOUDSPEAKER position ($\chi^2(3) = 12.8, P$=0.00519). Participants walked statistically significantly faster when loudspeaker D was playing compared to loudspeakers A and B, but the difference was minimal - 0.008 m/s for both loudspeakers.

Model selection analysis is provided in Table 57 in Appendix C. Refer to Tables 58-61 in Appendix C for pairwise comparison results.

3 Discussion

The next section explores the main findings derived from the analysis of results in the context of the research questions posed in Chapter 5.

3.1 What Are the Main Factors That Influence the Plausibility of an Audio-only Augmented Reality Scene and How Do They Affect Evaluation?

The evaluation of plausibility was primarily shaped by the interaction of two factors: loudspeaker position and method of playback. In the context of plausibility assessment, distinct patterns emerged for two pairs of loudspeakers: A-C and B-D, with significantly different ratings. Pair A-C obtained very similar ratings, with both auralizations being slightly less plausible than the real loudspeaker. While the difference was subtle, it held statistical significance. In contrast, the ratings for loudspeakers B and D demonstrated a substantial divergence. Specifically, loudspeaker D was rated as significantly less plausible than loudspeaker B for all three rendering methods, including real playback. Surprisingly, for loudspeaker B, which shared a similar position with loudspeakers A and C, no statistical difference existed between auralizations and the real loudspeaker, in contrast to the ratings for loudspeakers A and C.

This observation could be linked to the small disparities between the methods of playback being overshadowed by the distinct positions and orientations of loudspeakers B and D. As loudspeaker B was invariably paired with loudspeaker D, it was more difficult to discern the auralizations. Conversely, pair A-C exhibited two loudspeakers with a highly similar position in the room, making it easier to evaluate the differences between different methods of playback. It proved that the context in which sound is presented affects the plausibility evaluation, as was already shown in a previous study by (Neidhardt & Zerlik, 2021). It also means that the difficulty of the task can be controlled by the position of the real sound within the pair.

An additional interesting observation pertained to real loudspeakers B and D. The expected outcome of the plausibility evaluation was that real loudspeakers would always be perceived as highly plausible. In contrast, the results indicated that real loudspeaker D was rated as less plausible than real loudspeaker B. One possible explanation of this phenomenon was the disparity between listeners' expectations regarding how loudspeaker D should sound, given its position in the room, and the actual sound it produced. This disparity might be associated with the inaccuracy of listener expectations and the ability to detect acoustic changes in the room (Shinn-Cunningham & Ram, 2003).

It became evident that loudspeaker D posed the biggest challenge for the auralizations, as both methods obtained low scores of plausibility for loudspeaker D. Section 4 in Chapter VII aims to investigate the possible links between the objective characterization of auralizations and lower perceptual rating scores.

3.2 How Other Attributes of Sound Were Affected by the Method of Playback and Loudspeaker Position?

3.2.1 Blur

Blur ratings exhibited a strong correlation with plausibility ratings, particularly for loudspeakers A, B, and C. Importantly, the most significant deviation between blur and plausibility evaluations occurred for loudspeaker D, where blur ratings for loudspeaker D were lower than plausibility ratings for loudspeaker D in relation to loudspeakers A, B, C. Given that real loudspeaker D also received lower blur ratings than plausibility ratings, it suggests that the blur scores for loudspeaker D were influenced not only by the playback method but also by the distance between the listener and the loudspeaker. This divergence can be attributed to the farther positioning of loudspeaker D from the walking path compared to the other loudspeakers, impacting the ratings by considering both the playback method and the distance.

The main difference observed pertained to the SRIR method, which did not show a statistically significant difference from real playback for loudspeakers A and C, unlike the plausibility ratings. The lower plausibility scores for the SRIR method were driven by factors other than the blurriness of the stimuli. Additionally, there was a statistically significant difference between blur ratings of loudspeakers A-C for GA method. Loudspeaker C was rated lower than loudspeaker A. This difference does not occur for real playback and SRIR method. The reason for lower ratings of loudspeaker C with GA method might be associated with the specific early reflections pattern created by this method.

3.2.2 Localization Error

The localization error estimate values are very similar to the ratings of blur and plausibility. This correlation suggests that blur ratings are mostly associated with the difficulty of source localization. The plot showing the points marked by participants answering the localization question reveals that loudspeaker D has a clearly larger centroid for the two auralizations than for real playback. Answer points are localized mostly in front and on the left side of the loudspeaker, which suggests that the main problem might be related to the perception of distance between the listener and the source (see Figure 21).

3.2.3 Errors of Loudspeaker Recognition

The analysis of participants' answers to the localization question aimed to identify the closest loudspeaker to the indicated point, revealing which loudspeaker participants assumed was playing. However, this method has limitations, importantly in cases where sound localization accuracy was low, and the point was placed near the center between two loudspeakers. Consequently, an answer might be deemed an error even if the participant correctly identified the playing loudspeaker. Despite these limitations, the analysis provides insights into the accuracy of localization.

Loudspeaker recognition errors could stem from two main causes: memory errors due to the high cognitive load of the task or inaccuracies in rendering, leading participants to assume the wrong loudspeaker was playing. In particular, errors in loudspeaker recognition were strongly influenced by the order of playback. Intriguingly, loudspeakers A, B, and D were better recognized when walking back, whereas loudspeaker C was better recognized when walking forward (see Figure 22).

For loudspeakers A and C, confusion occurred primarily between the two, with A being mistaken for C and vice versa, suggesting that these errors were predominantly memory-related (refer to Figure 23). Even though recognition of the loudspeaker seemed easy, the task proved to be more difficult in practice. Informal discussions with the participants of pilot tests and the actual study revealed that sometimes they could not remember what was the order of the loudspeakers' playback when finishing the trial. The secondary cause of the recognition errors for pair A-C was related to the accuracy of rendering as both loudspeakers were sometimes confused with loudspeaker B.

Conversely, loudspeakers B and D were often confused with loudspeaker C, indicating in this case rendering inaccuracies as the primary cause. This hypothesis is supported by the observation that errors for loudspeaker D were almost exclusive to the two auralization methods and were not present during real playback.

Loudspeaker A exhibited a better recognition rate when participants were walking back the path, while loudspeaker C was better recognized during forward movement. A plausible hypothesis for this trend suggests that correctly locating a loudspeaker is easier when approaching it, as opposed to when moving away. This difficulty in localization accuracy when moving away is consistent with findings indicating that accuracy is better in the frontal direction than in the rear (Blauert, 1985). Additionally, when participants were finishing the trial, they heard loudspeaker A at a minimum distance, which facilitated memorizing its location.

Conversely, loudspeaker D demonstrated a different pattern than loudspeaker C, with better recognition during backward movement than during forward movement. This difference can be attributed to participants initially hearing loudspeaker D directly on their left side when starting to walk back, facilitating easier localization. Conversely, when moving forward, both auralization methods posed greater challenges for localization, probably due to the imperfect distance rendering as suggested by Figure 21. This observation aligns with the concurrent ratings of blur, supporting the notion that increased ambiguity in spatial perception may be linked to difficulties in accurately localizing sound sources.

3.2.4 Externalization

Overall, the externalization of both auralizations was rated similarly to the real loudspeaker. As expected, the ratings for real loudspeakers are slightly higher than for both auralization methods but the difference is minimal (refer to Figure 24). The difference between externalization of auralizations might stem from the implementation of non-individualized HRTFs although previous studies were not consistent in the observations of the influence of individualized HRTFs on externalization. The study by Begault et al. (2001) showed no difference between individualized and non-individualized HRTFs while Werner et al. (2016) found consistent improvement of externalization evaluation when using individualized HRTFs. It has to be noted that none of the studies were evaluating externalization in 6DoF. Usually, previous research on externalization was studying mainly 3DoF (implementing only head movements). In our study, participants were asked to keep their heads straight while walking, however, they could still experience the change in the angle between listener and source together with modifications of the Direct-to-Reverberant Ratio (DRR) while walking. For loudspeaker D, the change of the angle between source and listener was much smaller than for other loudspeakers however the DRR was the lowest, giving more importance to the reverberant sound which is an important cue for externalization.

Externalization ratings were also affected by the distance between the loudspeaker and the listener - loudspeaker D positioned farthest away from the path was rated as mostly externalized for real loudspeaker and method SRIR, but again the difference is small. In contrast, loudspeakers A, B, and C were perceived from a short distance when crossing their position on the path which could give the impression of ”near-internalization”. This finding is consistent with previous studies on externalization evaluation, in which externalization was found to be highly correlated with distance perception (Best et al., 2020). The discrepancy in ratings between loudspeaker D and others was not observed with the GA method, possibly due to its rendering quality, which decreased externalization even at greater distances.

Another important cue that could influence externalization assessment was the source azimuth. Previous research showed that externalization increases with sound laterality (Leclère et al., 2019). Consequently, in the listening test, the two contradictory effects affected the judgment - when the source was in the closest position to the listener which could decrease externalization, it was also fully lateralized which increased externalization.

3.2.5 Timbre, Reverberation, and Plausibility Difference

The results of timbre and reverberation difference ratings are very similar. There was no interaction between the pair of loudspeakers and the method of playback. As expected, pair A-C received lower difference ratings than pair B-D (see Figure 25). Interestingly, for pair A-C the difference in timbre was lower than for reverberation while for pair B-D the difference for both timbre and reverberation was on a similar level.

The method of playback had a significant influence on attribute evaluation. The minimum difference rating was obtained by pairs of two real loudspeakers. This difference can be understood as a baseline for the expected perceptual difference coming from the position divergence between the two loudspeakers. The other results should be interpreted in relation to this baseline. A pair of methods SRIR-SRIR obtained ratings equivalent to R-R which indicates that method SRIR succeeded in maintaining a good (at least realistic) homogeneity of the timbre and of the reverberation. It may be linked to the principle of the method based on convolution with a single SRIR (see Chapter V). However, the pair of methods SRIR-R obtained significantly higher ratings of timbre and reverberation difference which means that there was still some overall difference in the rendering of the timbre and of the reverberation between R and SRIR methods. For the GA method, there was a difference in the overall rendering of timbre and reverberation (GA-R) and also a more heterogeneous rendering across loudspeaker positions (GA-GA).

Ratings of timbre and reverberation seem to be correlated with plausibility difference. The only difference is that in contrast to timbre and reverberation difference, - SRIR-SRIR method plausibility difference estimates are higher than for the R-R method. This means that some other factor influenced plausibility ratings for the SRIR-SRIR method and caused the two loudspeakers to differ in the perception of plausibility. Overall, these results suggest that timbre and reverberation perception played a role in the plausibility judgments.

3.3 Secondary Effects

3.3.1 Trial Index

The trial index effect was statistically significant however the influence was minimal. It affected blur, localization error, errors of loudspeaker recognition, and timbre difference. The influence on blur and localization error was very similar. The stimuli were rated as less blurry in the subsequent trials and the localization error was reduced in the subsequent trials. The effect of the trial index was seen also in the number of errors in loudspeaker recognition. There are 4-5% more errors in the first 12 trials than in the rest of the trials. This leads to the conclusion that during the first 12 trials, participants were still learning the task.

3.3.2 Speed

The speed of walking data analysis aimed to validate which factors affected the speed of walking and to explore the possible correlation between the movement speed and the perception of the sound events. The histogram of mean speed among participants reveals that mean speed follows the normal distribution. The speed of walking was affected by the trial index, method of playback, order of playback, and loudspeaker position. Participants walked faster during the first 12 trials than during the other trials. Two hypotheses can explain this effect. Firstly, there was a period of adaptation to the task during which participants were adjusting their speed to be able to listen to the entire stimuli while walking (which was part of the task). Secondly, it is possible that after the first trials, participants found out that slowing down helps them to assess the stimuli more easily. The second hypothesis is supported by the localization error and blur ratings which are objective measures of stimuli. For both attributes, the ratings became consistent after the first 12 trials. The stimuli were rated as less blurry and localization error was smaller after 12 first trials which shows that slowing down allowed for better judgment.

The speed of walking was also affected by the method of playback. Participants walked faster when listening to the real loudspeakers and slowed down while listening to the auralizations. Real loudspeaker was easier to evaluate allowing participants to speed up. This hypothesis is supported by the results of blur and localization error ratings. Real loudspeaker was rated as less blurry and with lower localization error than auralizations which means that the real loudspeaker was easier to localize. The task required less effort for the real loudspeaker than for auralizations. As the subjective ratings of plausibility were related to the objective evaluation of localization error and blur it leads to the conclusion that the easier the task the higher were plausibility ratings.

Interestingly, participants walked marginally faster when moving forward compared to walking backward along the path. This observation could be explained by participants assessing their speed accurately when reaching the path simultaneously with stimulus completion. This realization allowed them to adjust their speed while walking back to ensure they listened to the entire stimulus during that segment.

3.3.3 Order of Playback

The analysis examined the sequence in which the stimuli were played and found that it had a distinct impact on various attributes such as plausibility, externalization, localization error, and blur. However, the influence of the playback order varied for each attribute. The stimulus played first was perceived as more plausible (by 2.2%) and more externalized (by 7%). However, it also received higher ratings for blur (by 2.5%) and had larger localization errors. These results suggest that memory errors played an important role in the perception of the stimuli. The second stimulus, which was more memorable, had a stronger impact on the participants' memories. In contrast, the first stimulus was not as well-remembered, leading to occasional confusion about the source of the sound and resulting in noticeable localization errors. Paradoxically, the less accurate memory of the first stimulus seemed to enhance its perceived plausibility and externalization, creating a contrast between the two stimuli. The order of playback was also involved in interaction with loudspeaker position and rendering method for loudspeaker recognition errors. For loudspeakers A, B, and D there were fewer errors for the second stimulus. Only loudspeaker C revealed different behavior probably because of the ease of remembering loudspeaker A position when it was played as second and thus facilitating the memory of the position of both loudspeakers within the pair.

4 Summary

The plausibility evaluation was mainly shaped by the interaction of two factors: loudspeaker position and method of playback. In particular, distinct patterns emerged for two pairs of loudspeakers: A-C and B-D, with significantly different ratings. Pair A-C obtained very similar ratings, with both auralizations being slightly less plausible than the real loudspeaker, while pair B-D demonstrated a substantial divergence. This divergence could be linked to the small disparities between the playback methods being overshadowed by the distinct positions and orientations of loudspeakers B and D. Moreover, loudspeaker D exhibited lowest ratings for both auralization methods indicating the limits of its rendering accuracy. Further investigation aims to explore the links between objective auralization parameters and lower perceptual rating scores.

Additionally, unexpected findings revealed that real loudspeaker D was rated as less plausible than real loudspeaker B, possibly due to listeners' expectations regarding how loudspeaker D should sound compared to its actual output. This observation is very important as it proves that plausibility should be considered as a continuous percept, not binary.

Blur ratings exhibited a strong correlation with plausibility ratings, particularly for loudspeakers A, B, and C, while localization error estimates aligned closely with blur and plausibility ratings. The method of playback had a significant influence on attribute evaluation, with real loudspeakers generally rated as less blurry and with a localization error lower than that of auralizations. Moreover, the order of playback impacted various attributes such as plausibility, externalization, localization error, and blur.

The analysis of loudspeaker recognition errors revealed intriguing patterns influenced by various factors. Errors in recognizing the correct loudspeaker were influenced by both cognitive processes and rendering inaccuracies. Importantly, the order of playback significantly affected the frequency of recognition errors, with fewer errors observed for the second stimulus compared to the first. This trend was consistent across most loudspeakers, suggesting a memory-related effect where the second stimulus was more memorable, leading to fewer errors in identifying the loudspeaker.

Confusion between loudspeakers A and C indicated memory-related errors, while loudspeakers B and D were often mistaken for loudspeaker C, pointing to rendering inaccuracies as the primary cause. This distinction suggests that the participants' errors were influenced by both the memory-related cognitive load and the fidelity of the rendering method. Additionally, differences in recognition rates between moving forward and backward hinted at the impact of directional movement on localization accuracy.

The evaluation of externalization in the experiment indicated that both auralizations were rated similarly to the real loudspeaker, with only minimal differences observed. Despite the slight discrepancy, the overall ratings suggested that participants perceived the auralizations as correctly externalized, reflecting a successful rendering of spatial cues. However, the distance of the loudspeaker from the listener influenced the perception of externalization, with loudspeaker D, positioned farthest from the path, receiving slightly higher ratings compared to other loudspeakers. This finding was consistent with previous research indicating a strong correlation between externalization and distance perception. The results suggested that the implementation of non-individualized Head-Related Transfer Functions (HRTFs) in the auralizations might have contributed to the minor differences observed in externalization ratings. However, other hypotheses including the imperfect decoding of the room effect from HOA to binaural through virtual loudspeakers need to be taken into account.

The analysis of walking speed during the experiment revealed several significant findings. Participants tended to walk faster during the initial 12 trials compared to subsequent trials, suggesting an adaptation period or a learning curve for the task. This speed adjustment likely aimed to synchronize participants' movement with the duration of the stimuli. In particular, the speed of walking was influenced by the method of playback, with participants walking faster when listening to real loudspeakers compared to auralizations. This difference in walking speed reflected the perceived ease of evaluating the stimuli, with real loudspeakers requiring less effort for localization compared to auralizations. This important finding suggests that plausibility evaluation could be implemented by observing behavioral responses to different stimuli.

The following chapter focuses on the comparison of standing and walking phases, exploring the influence of movement on the perception of sound.

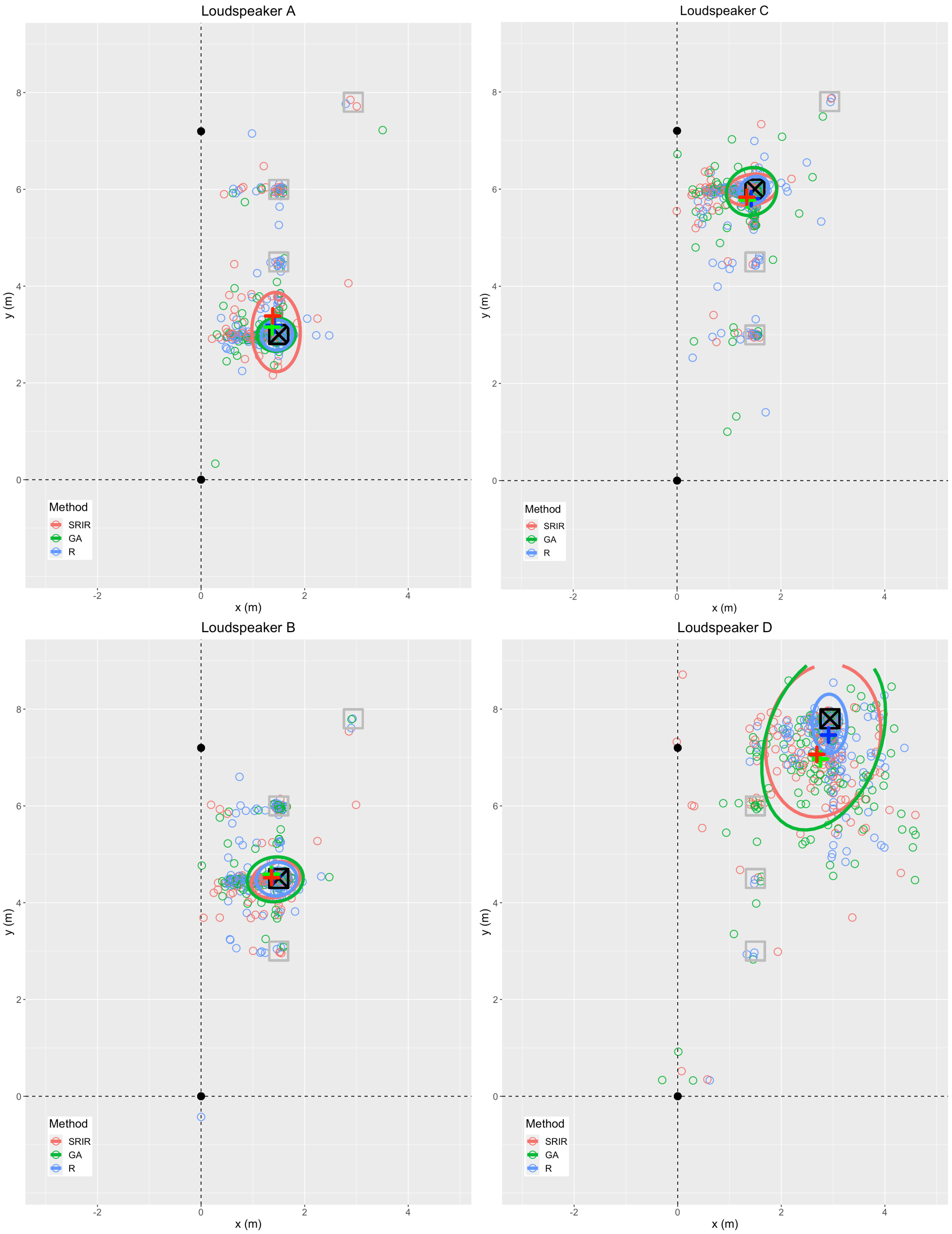


Figure 21: Stimulus position reported by participants during the walking phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Positions [0,0] and [0,7.2] mark the starting and ending points of the walking path.

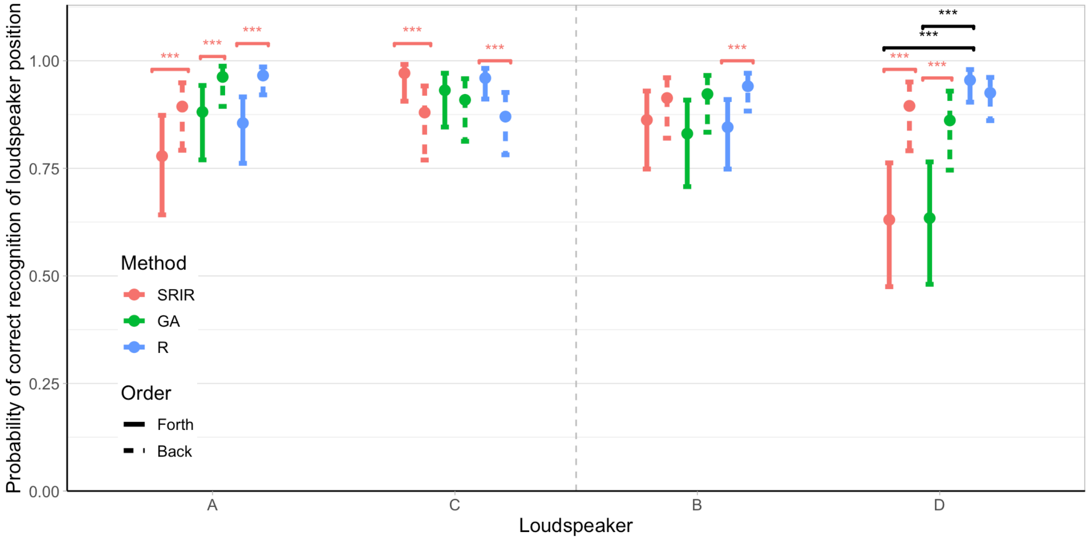


Figure 22: Predicted means and 95% confidence intervals for probability of correct recognition of loudspeaker position in walking phase. Black brackets indicate statistically significant differences between methods while red brackets indicate statistically significant differences between the two orders of playback.

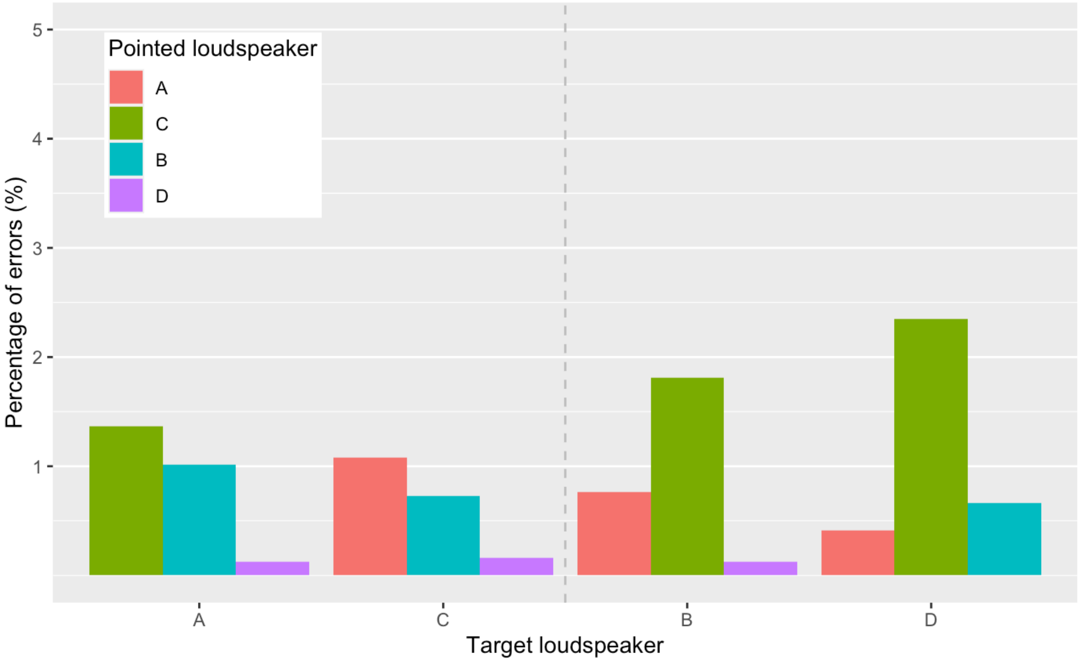


Figure 23: Distribution of loudspeaker recognition errors in walking phase. The pointed loudspeaker indicates which loudspeaker was the closest to the position marked by the participant. For better clarity, the vertical scale was limited to the 0-5% range.

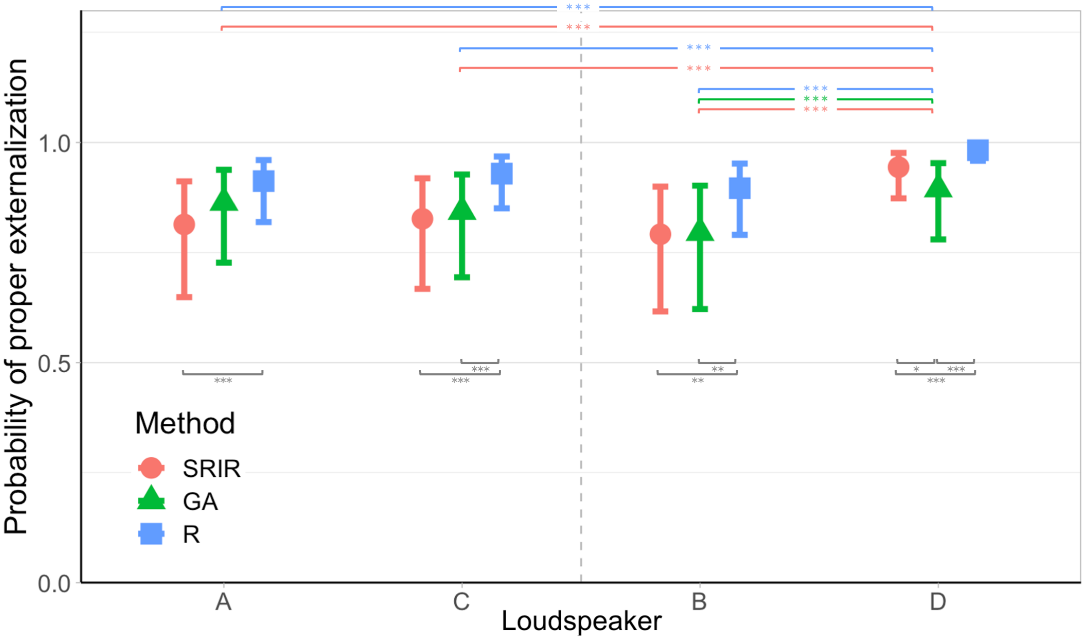


Figure 24: Predicted values and 95% confidence intervals for externalization ratings in walking phase according to the rendering method and loudspeaker position (***P < 0.05)

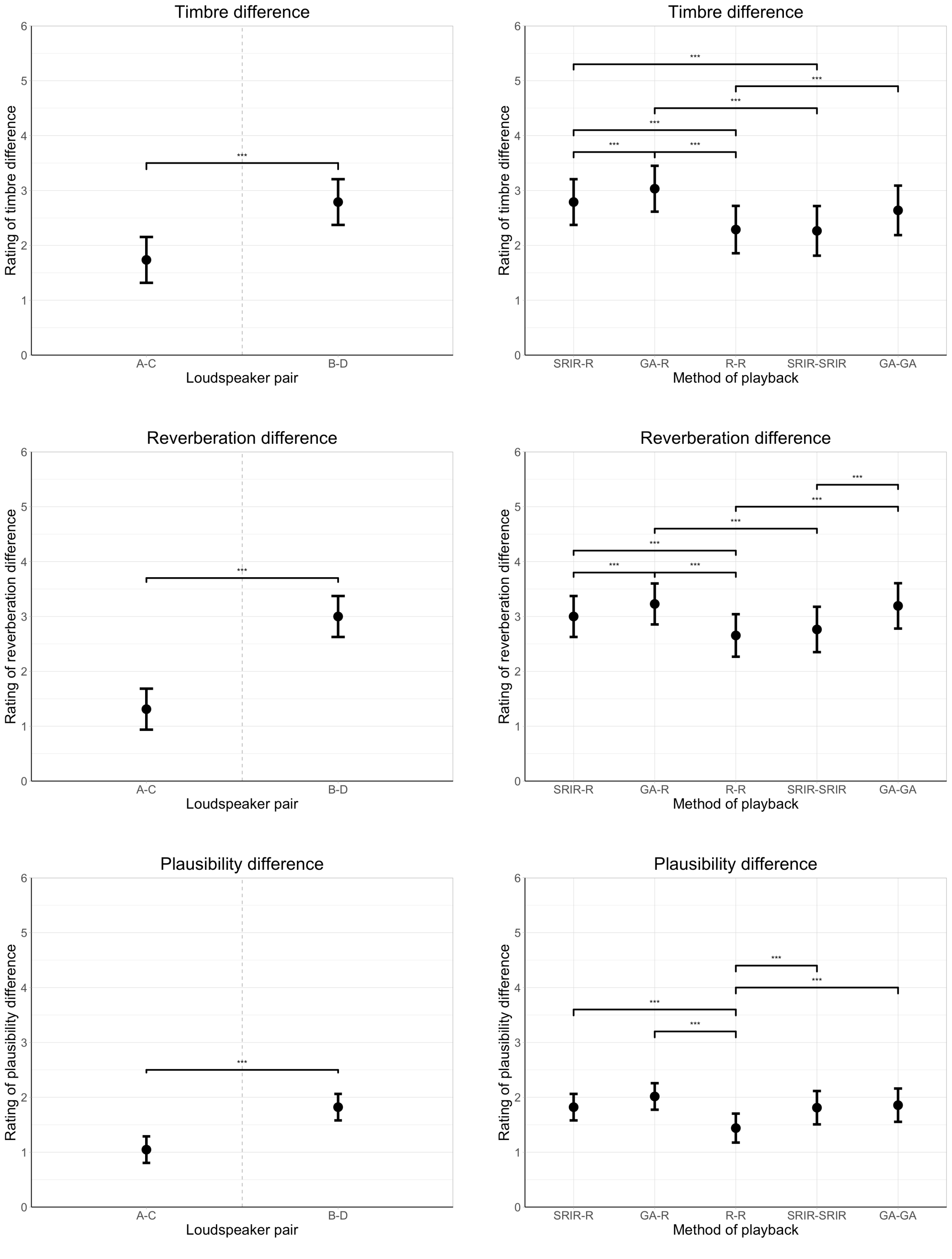


Figure 25: Predicted means and 95% confidence intervals for timbre, reverberation, and plausibility difference ratings according to loudspeaker pair and rendering method pair(***P < 0.05)

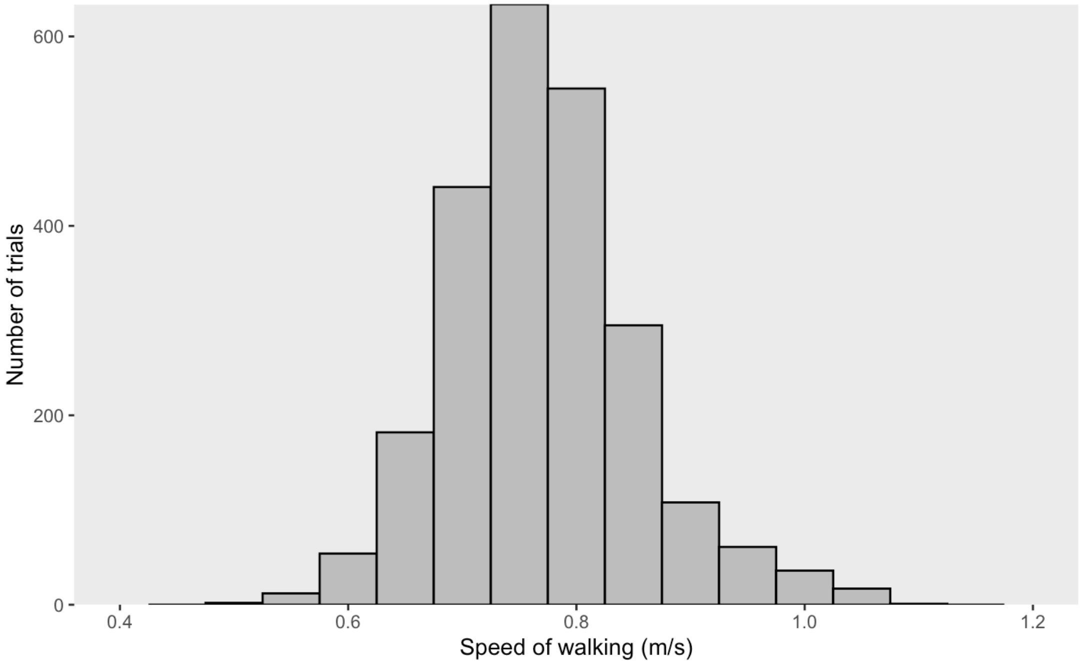


Figure 26: Histogram of speed of walking data

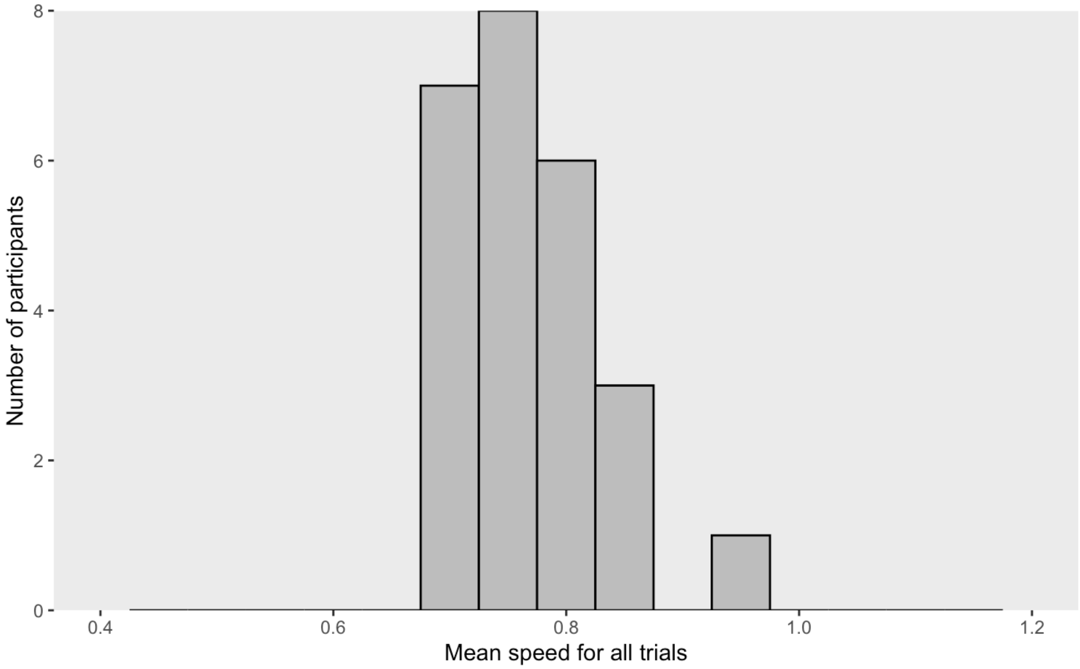


Figure 27: Histogram of mean of speed for all the trials

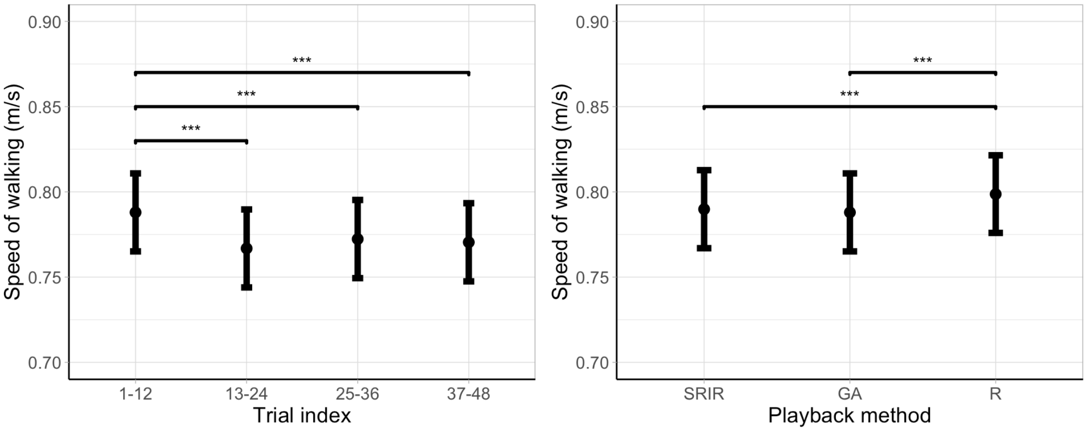


Figure 28: Influence of trial index and playback method on the walking speed of participants

CHAPTER VII

RESULTS OF THE EXPERIMENT - WALKING VS STANDING PHASE

This chapter focuses on the analysis of results from both phases of the experiment: walking and standing. The reason behind initially presenting the results only from the walking phase in the previous chapter, stems from the larger number of participants in this phase, totaling 33. By incorporating data from all responses, we ensure a more robust statistical analysis. However, not all participants were able to participate in the subsequent standing phase, resulting in a reduced sample size of 25 listeners who took part in both phases.

As a consequence, this chapter primarily focuses on examining the differences between the two phases. Nevertheless, we will juxtapose these findings with the analysis conducted in the preceding chapter to verify the consistency of results, particularly concerning the walking phase, considering the smaller sample size in this analysis.

1 Comparison of Walking and Standing Phases

The data collected during both the standing and walking phases of the experiment was combined and analyzed to assess the impact of the experimental phase on subjective evaluation. This analysis encompassed data from 25 participants who took part in both phases, resulting in the examination of a total of 4798 ratings. Following the methodology described in Section **??**, the analysis was conducted to explore the influence of various factors on attribute evaluation. These factors included the experimental phase (walking or standing), acoustic rendering method, localization of virtual and real loudspeakers, trial index, order of playback, absolute height difference between participants and loudspeakers, amplitude of yaw movement, rendering method pair, as well as inter-subject and intra-subject variability.

1.1 Plausibility

Table 18

Plausibility: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	446.29	1.00	0.00
Order	9.45	1.00	0.00
Height_abs	4.43	1.00	0.04
Phase	0.37	1.00	0.54
Method	20.22	2.00	0.00
Ldspkr	148.37	3.00	0.00
Method_type	14.06	4.00	0.01
Phase:Method	0.19	2.00	0.91
Phase:Ldspkr	1.99	3.00	0.57
Method:Ldspkr	68.34	6.00	0.00
Phase:Method:Ldspkr	21.49	6.00	0.00

Analysis of the results showed that the best model for predicting plausibility ratings (scale 0 - not at all plausible, 6 - very plausible) was: PHASE * LOUDSPEAKER position * playback METHOD + rendering METHOD PAIR + ORDER of playback + HEIGHT difference between listener and loudspeaker (see Figure 29). There was a significant interaction between PHASE, LOUDSPEAKER position, and rendering METHOD ($\chi^2(6) = 21.50, P < 0.01$) as indicated in Table 18. The significant difference between the two phases occurred for loudspeaker D simulation SRIR where the standing phase was rated as more plausible (3.85 in the walking phase, 4.40 in the standing phase). The opposite situation occurred for REAL playback on loudspeaker D where the walking phase was rated as more plausible (4.87 in the walking phase, 4.50 in the standing phase), see Figures 29 and 30.

Results revealed a statistically significant influence of the rendering METHOD PAIR ($\chi^2(4) = 14.06, P < 0.01$). The rendering method pair GA-GA received higher ratings than all other pairs of methods besides methods R-R, as illustrated in Figure 31.

The influence of playback ORDER was statistically significant ($\chi^2(1) = 9.41, P < 0.01$) but the difference is minimal: only 0.09 points less for stimulus which was played as second.

Interestingly, there was a significant influence of the HEIGHT difference between the participant and the loudspeaker ($\chi^2(1) = 4.42, P < 0.05$). The estimate for height difference was -0.028 ± 0.01 point (p = 0.0368). This implies that for each additional 10 cm of height difference between the participant and the loudspeaker, the plausibility rating was lower by 0.28 points (see Figure 42). The results are very similar to the outcome of the analysis on the smaller group of participants in the previous chapter.

Model selection analysis is provided in Table 62 in Appendix D. Refer to Tables 63-66 in Appendix D for pairwise comparison results.

1.2 Blur

During the experiment, participants rated the blurriness of the stimuli on the scale: 0 - very focused, 6 - very blurry. To facilitate the comparison of different attributes, the y scale for blur ratings on the plots was reversed. Results indicated that the best model

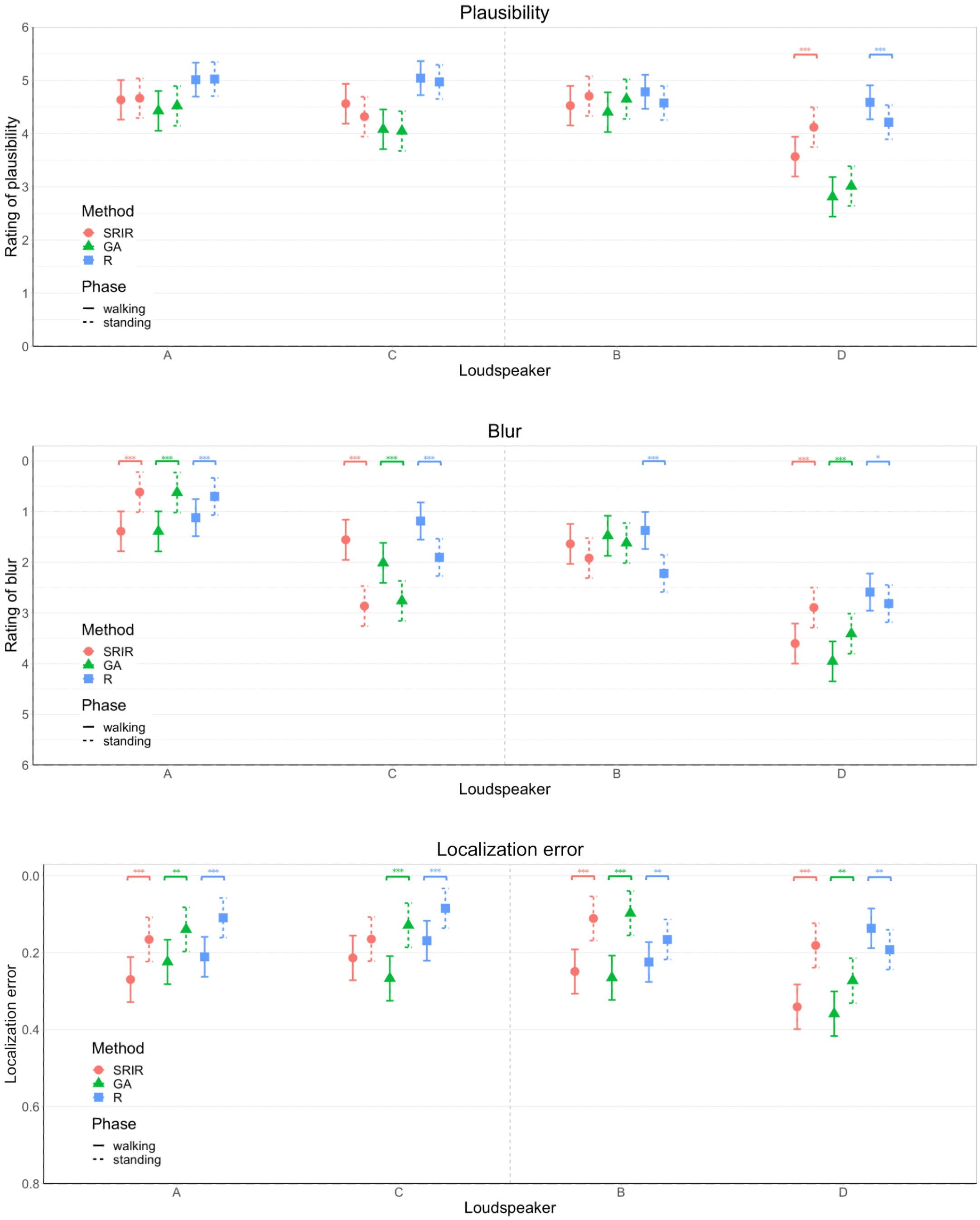


Figure 29: Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05) according to the rendering method, loudspeaker position, and phase. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry).

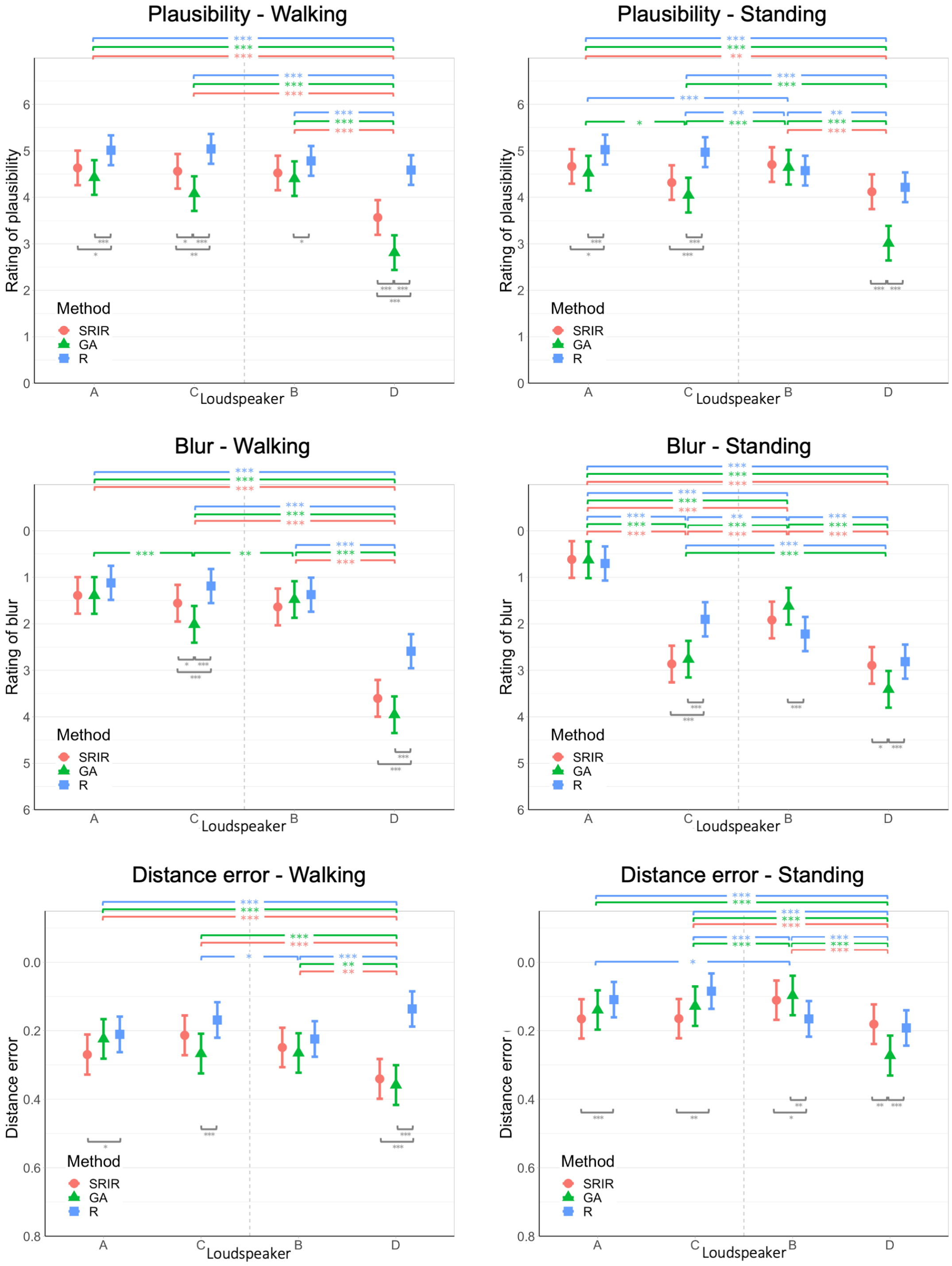


Figure 30: Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05) according to the rendering method, loudspeaker position and phase. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry).

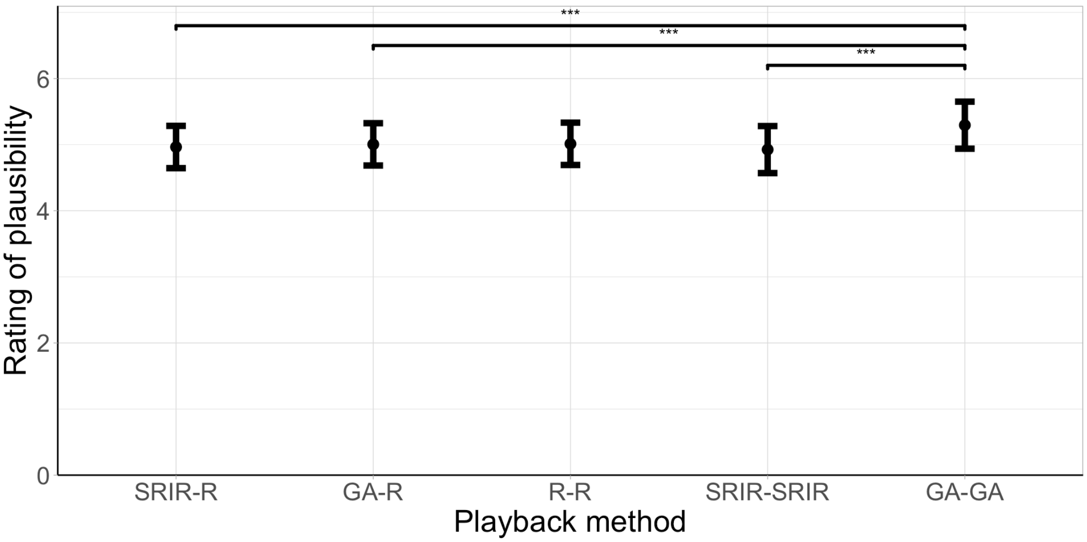


Figure 31: Predicted means and 95% confidence intervals for plausibility ratings according to the rendering method pair (***P < 0.05).

for predicting blur ratings is: PHASE * LOUDSPEAKER position * playback METHOD + trial INDEX + HEIGHT difference between loudspeaker and participant. There was a significant interaction between PHASE, LOUDSPEAKER pair, and playback METHOD ($\chi^2(6) = 40.62, P < 0.001$) as indicated in Table 19. In general, there are larger evaluation differences between phases for blur in comparison to plausibility ratings (see Figure 29 and 30). Loudspeaker A was rated as significantly more focused in the standing phase than in the walking phase (biggest difference for auralization SRIR and GA - 0.77). Loudspeaker C was rated as significantly more blurry in the standing phase than in the walking phase (biggest difference for simulation SRIR - 1.3). Loudspeaker B was also rated as significantly more blurry in the standing phase than in the walking phase (maximum difference for REAL loudspeaker - 0.85). At last, loudspeaker D was rated as more focused in the standing phase than in the walking phase for auralization

Table 19

Blur: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	19.54	1.00	0.00
Phase	26.06	1.00	0.00
Method	6.51	2.00	0.04
Ldspkr	379.83	3.00	0.00
Index_id	23.60	3.00	0.00
Height_abs	11.54	1.00	0.00
Phase:Method	5.56	2.00	0.06
Phase:Ldspkr	63.24	3.00	0.00
Method:Ldspkr	66.65	6.00	0.00
Phase:Method:Ldspkr	40.62	6.00	0.00

SRIR and GA (difference 0.71 and 0.55 respectively) and slightly more blurry for a REAL loudspeaker (difference 0.23).

The influence of trial INDEX was statistically significant ($\chi^2(3) = 23.6, P$<0.001) but the difference was minimal. The stimuli were rated as less blurry with the subsequent trials: there was a 0.24 point difference between the first (1-12) and last (37-48) section of trials (see Figure 44).

The estimate for the HEIGHT difference between the participant and loudspeaker (refer to Figure 42) was 0.048 ± 0.014 point ($\chi^2(1) = 11.54, p$<0.001). This implies that for each additional 10 cm of height difference between the participant and the loudspeaker the rating of blur increases by 0.48 point.

Model selection analysis is provided in Table 67 in Appendix D. Refer to Tables 68-72 in Appendix D for pairwise comparison results.

Table 20

Localization error: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	16.58	1.00	0.00
Method	6.34	2.00	0.04
Ldspkr	27.03	3.00	0.00
Phase	10.24	1.00	0.00
Order	27.89	1.00	0.00
Index_id	15.58	3.00	0.00
Height_abs	23.00	1.00	0.00
Method:Ldspkr	68.33	6.00	0.00
Method:Phase	0.35	2.00	0.84
Ldspkr:Phase	7.12	3.00	0.07
Method:Ldspkr:Phase	41.31	6.00	0.00

1.3 Localization Error

The best model for predicting localization error (refer to Section 1.2 in Chapter VI for more details on localization error calculation) for both phases was: PHASE * playback METHOD * LOUDSPEAKER position + ORDER of playback + trial INDEX + HEIGHT difference. There was a significant interaction between PHASE, METHOD of playback, and LOUDSPEAKER position ($\chi^2(6) = 41.31, P$<0.001), see Table 20. The statistically significant differences between the two phases of the experiment appeared for all of the loudspeakers (see Figure 29). In general, localization error was larger in the walking phase than in standing for all of the loudspeakers and methods besides REAL loudspeaker D where the localization error was smaller in the walking phase.

Besides that a main effect of the ORDER of playback ($\chi^2(1) = 27.89, P$<0.001) was observed. The localization error was 0.03 lower for the stimuli that were played when coming back.

There was also a main effect of HEIGHT difference between participant and loudspeaker ($\chi^2(1) = 23.00, P$<0.001), see Figure 42. The estimate for height difference was 0.1 ± 0.02 (P$<$0.001). This implies that for each additional 10 cm of HEIGHT difference between the participant and the loudspeaker, the localization error increases by 0.1.

The main effect of the trial INDEX was also observed ($\chi^2(1) = 15.58, P$=0.0014). The estimate for localization error was 0.21 for trials 1-12, 17.8 cm for trials 13-24, 0.18 for trials 25-36, and 0.185 for trials 37-48 (refer to Figure 44). The position error was statistically significantly larger for the first 12 trials than for other groups of trials.

Model selection analysis is provided in Table 1 in Appendix D. Refer to Tables 78-81 in Appendix D for pairwise comparison results.

1.4 Loudspeaker Recognition Rate

Table 21

Loudspeaker recognition rate: Analysis of Deviance Table (Type III Wald chi-square tests)

Chisq	Df	Pr(>Chisq)	
(Intercept)	91.70	1.00	0.00
Phase	35.05	1.00	0.00
S	9.37	3.00	0.02
Order	17.22	1.00	0.00
Phase:S	84.42	3.00	0.00

During the experiment, participants were asked to indicate the position of the sound source in relation to the room and loudspeakers on the overhead view diagram of the room. The answers were analyzed to determine which loudspeaker was closest to the point marked on the graph. In total, there were 264 errors of loudspeaker recognition

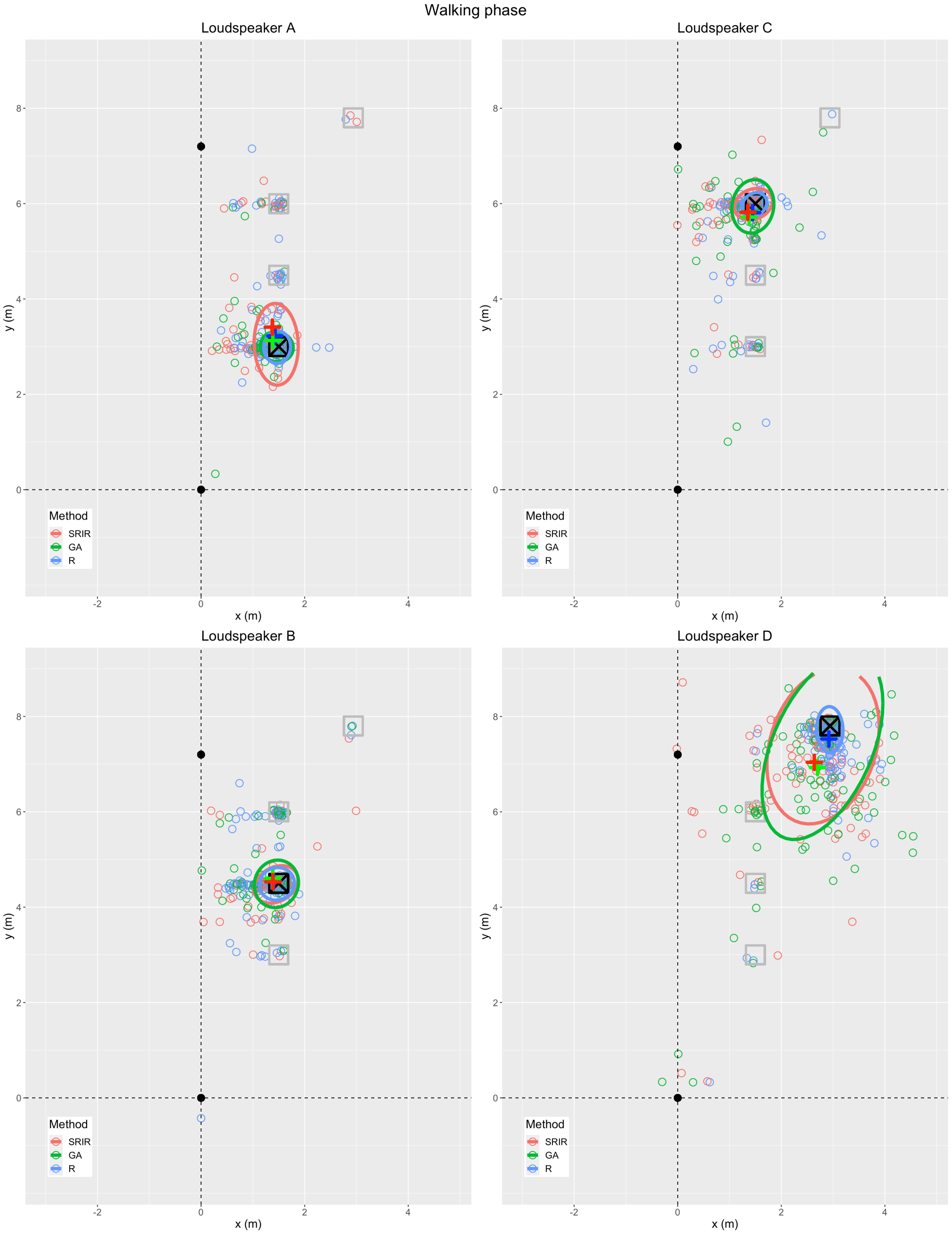


Figure 32: Stimulus position reported by participants during the walking phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Positions [0,0] and [0,7.2] mark the starting and ending points of the walking path.

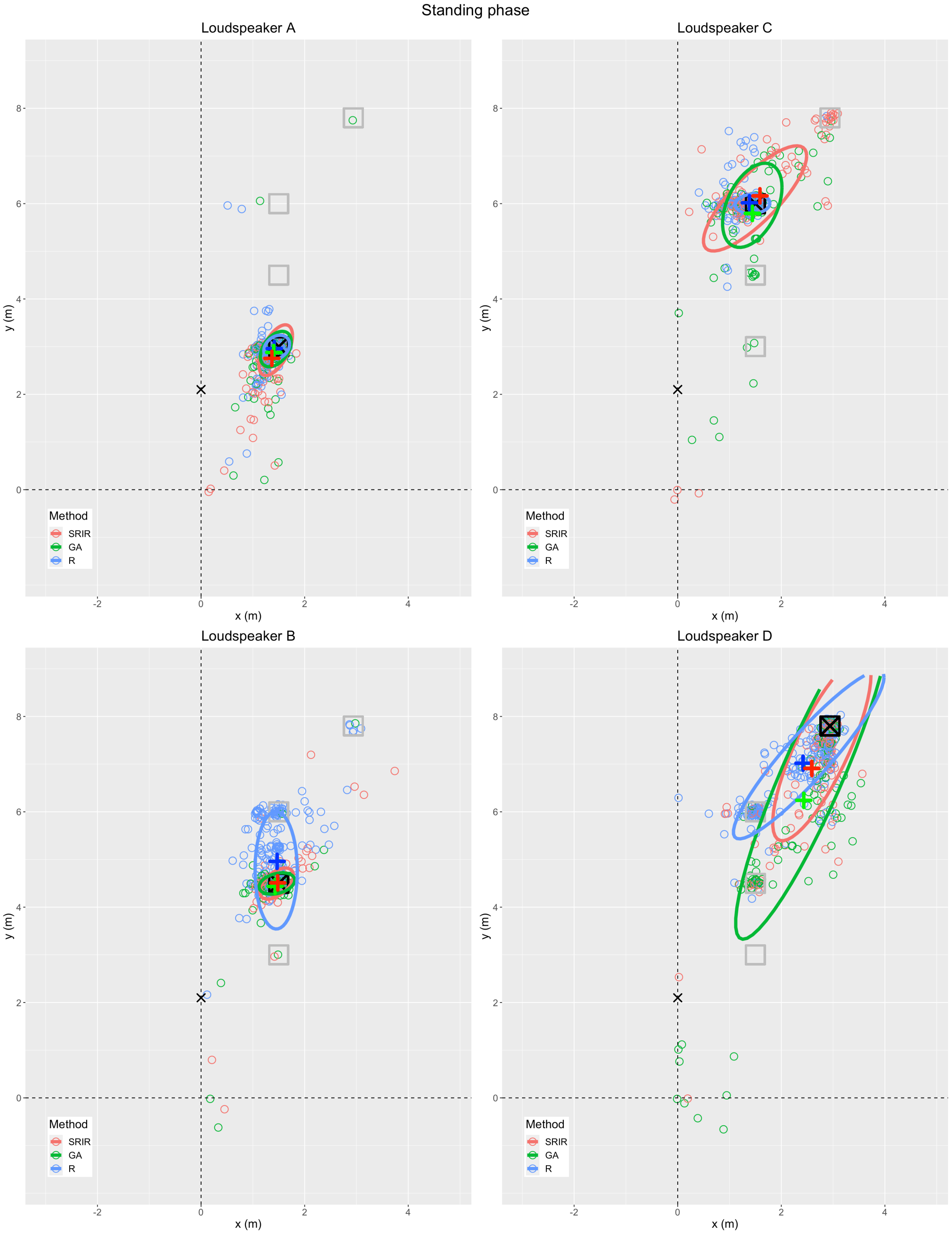


Figure 33: Stimulus position reported by participants during the standing phase for all loudspeakers with centroids (marked with a cross) and ellipses representing 95% confidence intervals around data points. Position [0,2.1] marks the standing point during the experiment.

out of 2400 trials (11%) in the walking phase and 372 errors of loudspeaker recognition out of 2400 trials (15.5%) in the standing phase. The best model for predicting the probability of correct recognition of loudspeaker position for both phases was: PHASE * LOUDSPEAKER position + ORDER of playback. There was a statistically significant interaction between PHASE and LOUDSPEAKER position ($\chi^2(3) = 84.42, P<0.001$), refer to Figure 34 and Table 21. The statistically significant difference between the two phases was observed for loudspeakers A, B, and D (see Figure 34). Loudspeaker A had a much lower probability of correct answers in the walking phase. Conversely, for loudspeakers B and D, there were more incorrect answers in the standing phase than in walking. During the walking phase, all loudspeakers had almost the same probability of correct recognition (except the loudspeaker D had a lower recognition probability than C). During the standing phase, all loudspeakers obtained statistically significantly different recognition probabilities. The biggest difference was observed between loudspeakers B and D. Loudspeaker A had the highest probability of correct recognition while loudspeaker D had the lowest.

The influence of the ORDER of playback was statistically significant ($\chi^2(1) = 17.22, P<0.001$). There was a 0.01 points higher probability of correct recognition for stimulus which was played as second.

Model selection analysis is provided in Table 82 in Appendix D. Refer to Tables 83-84 in Appendix D for pairwise comparison results.

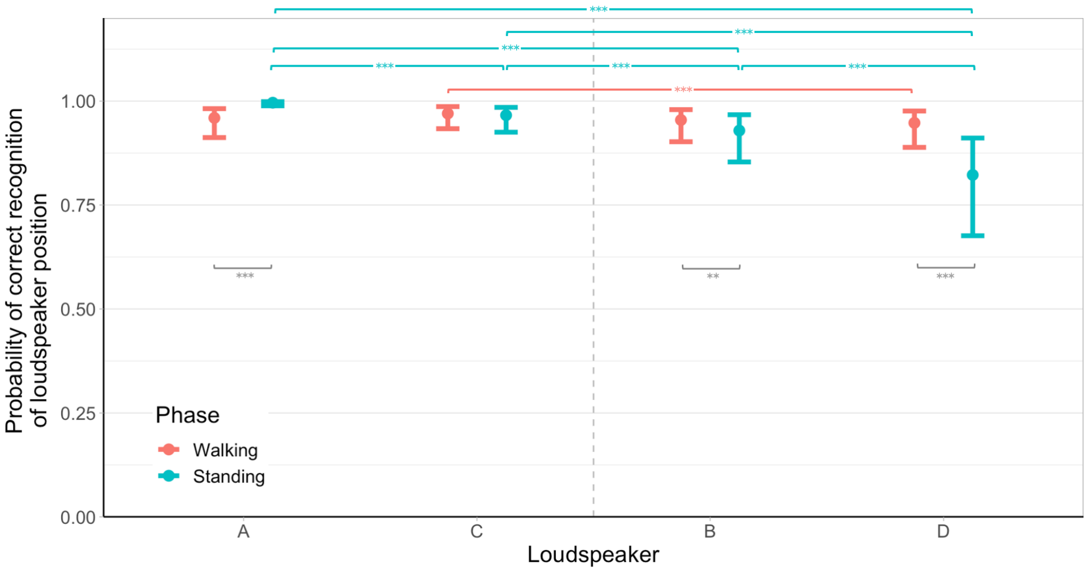


Figure 34: Predicted means and 95% confidence intervals for percentage of correctly recognized loudspeaker positions in both phases according to the loudspeaker position and phase.

1.5 Externalization

Participants provided externalization ratings by selecting one of three areas on the diagram: ”outside of the head” (scored as 2), ”close to the head” (scored as 1), and ”inside the head” (scored as 0). Given that externalization is a categorical variable, analysis based on general linear mixed models (GLMM) had to be applied. Responses categorized as 0 and 1 were merged and redefined as 0 because both indicated issues with externalization. The response labeled ”outside of the head” was redefined as 1. This adjustment allowed the analysis to focus on predicting the plausibility of receiving a response indicating ”outside of the head”.

The best model for predicting proper externalization probability was: PHASE * LOUDSPEAKER position + playback METHOD + ORDER of playback + HEIGHT

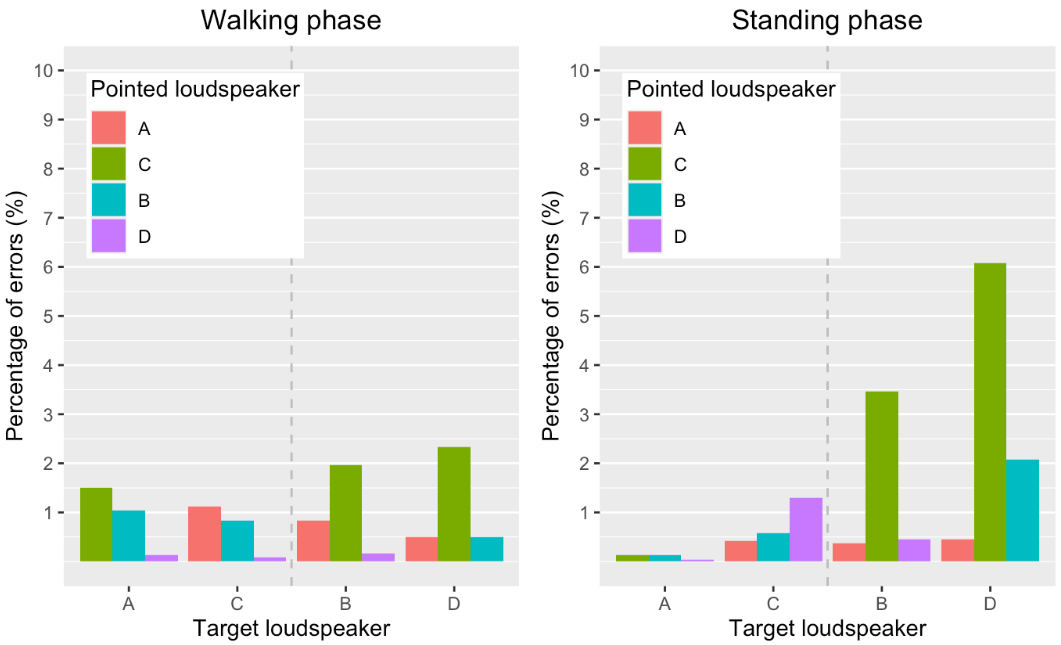


Figure 35: Distribution of loudspeaker recognition errors according to the target loudspeaker and the phase. The pointed loudspeaker indicates which loudspeaker was the closest to the position marked by the participant. For better clarity, the vertical scale was limited to the 0-10% range.

difference between loudspeaker and participant. There was a significant interaction between PHASE and LOUDSPEAKER position ($\chi^2(3) = 111.13, P < 0.001$), see Figure 36 and Table 22. There was a statistically significant difference in proper externalization probability between the two phases for all loudspeakers although in opposite directions. Loudspeaker A was less externalized in standing than in the walking phase in contrast to loudspeakers B, C, and D which were more externalized in the standing phase. Moreover, loudspeaker D obtained a significantly higher probability of proper externalization than any other loudspeaker in the walking phase, while loudspeaker A obtained lower values of probability than any other loudspeaker in the standing phase.

Table 22

Externalization: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	37.86	1.00	0.00
Phase	9.03	1.00	0.00
Ldspkr	61.65	3.00	0.00
Method	110.37	2.00	0.00
Order	53.83	1.00	0.00
Height_abs	7.42	1.00	0.01
Phase:Ldspkr	111.13	3.00	0.00

Moreover, there was a significant main effect of the playback METHOD ($\chi^2(2) = 110.37, P < 0.001$), see Figure 36. In general, REAL playback obtained statistically significantly higher probability values than methods GA and SRIR.

In addition to that, a main effect of the ORDER of playback ($\chi^2(1) = 53.83, P$<0.001) was observed. The stimuli that were played as second obtained 0.07 point lower probability of receiving an externalization score of value 1. This result is coherent with the outcome of the analysis from Chapter VI where stimuli played second obtained 0.15 lower probability of proper externalization when analyzing data from the walking phase only.

There was also a main effect of HEIGHT difference between the participant and loudspeaker ($\chi^2(1) = 7.42, P = 0.0065$), see Figure 43. The externalization rating estimate for height difference was -0.08 ± 0.03 point (p = 0.0056). This implies that each additional 0.1 m of absolute height difference between the participant and loudspeaker decreases the probability of obtaining an externalization score of 1 by 0.08.

Model selection analysis is provided in Table 73 in Appendix D. Refer to Tables 74-76 in Appendix D for pairwise comparison results.

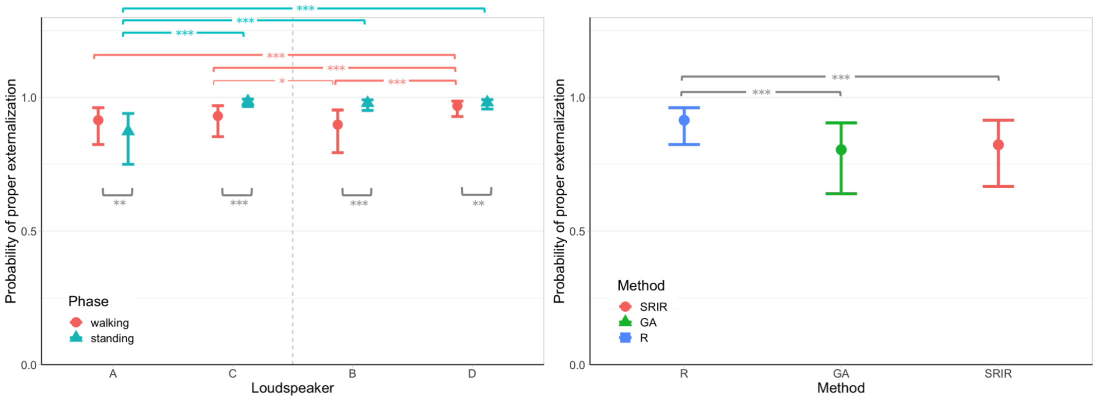


Figure 36: Predicted probability and 95% confidence intervals for an answer ”outside of the head” for externalization question according to the loudspeaker position and phase (plot on the left) and rendering method (plot on the right) (***P < 0.05)

1.6 Timbre Difference

Table 23

Timbre difference: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	59.81	1.00	0.00
Method_type	128.98	4.00	0.00
Ldspkr_type	158.04	1.00	0.00
Phase	40.31	1.00	0.00
Height_abs	11.43	1.00	0.00
Ldspkr_type:Phase	56.89	1.00	0.00

During the experiment, participants rated the timbre difference within each pair of speakers A-C or B-D (scale: 0 - not different, 6 - very different). Analysis of the results showed that the best model for predicting timbre difference ratings was: rendering METHOD PAIR + PHASE * LOUDSPEAKER PAIR + HEIGHT difference between listener and loudspeaker. There was a significant influence of the playback METHOD PAIR on the ratings of timbre difference ($\chi^2(4) = 129.0, P$<0.001). The lowest rating was obtained by a pair of two REAL speakers (R-R, 2.01) and two SRIR auralizations (SRIR-SRIR, 2.10).

The highest rating of timbre difference was observed on pair GA-R (2.83). Pairs SRIR-R and GA-GA were rated similarly with the mean estimate slightly lower than for pair GA-R (2.50, 2.56 respectively).

There was a significant interaction of PHASE and LOUDSPEAKER PAIR (($\chi^2(1) = 56.9, P$<0.001), see Table 23. The effect of loudspeaker position was especially critical in a walking phase where the timbre difference for pair A-C was significantly lower than for pair B-D (see Figure 37). The difference between pairs in the standing phase was minimal. In general, the timbre difference was larger in standing phase than in walking. During the test, participants also listened to pairs of stimuli with different combinations of playback methods. The possible pairs of methods were: R-R, SRIR-R, GA-R, SRIR-SRIR, GA-GA.

There was also a main effect of HEIGHT difference between the participant and loudspeaker ($\chi^2(1) = 11.4, P$<0.001), see Figure 43. The estimate for height difference was 0.064 ± 0.019 cm (p = 0.001). This implies that for each additional 10 cm of height difference between participants, the timbre difference increased by 0.64 points.

Model selection analysis is provided in Table 85 in Appendix D. Refer to Tables 86-88 in Appendix D for pairwise comparison results.

1.7 Reverberation Difference

Similarly to timbre difference, participants rated the reverberation difference within each pair of speakers A-C or B-D (scale: 0 - not different, 6 - very different). The best model for predicting reverberation difference ratings was: PHASE * LOUDSPEAKER PAIR + rendering METHOD PAIR. There was a significant interaction of PHASE *

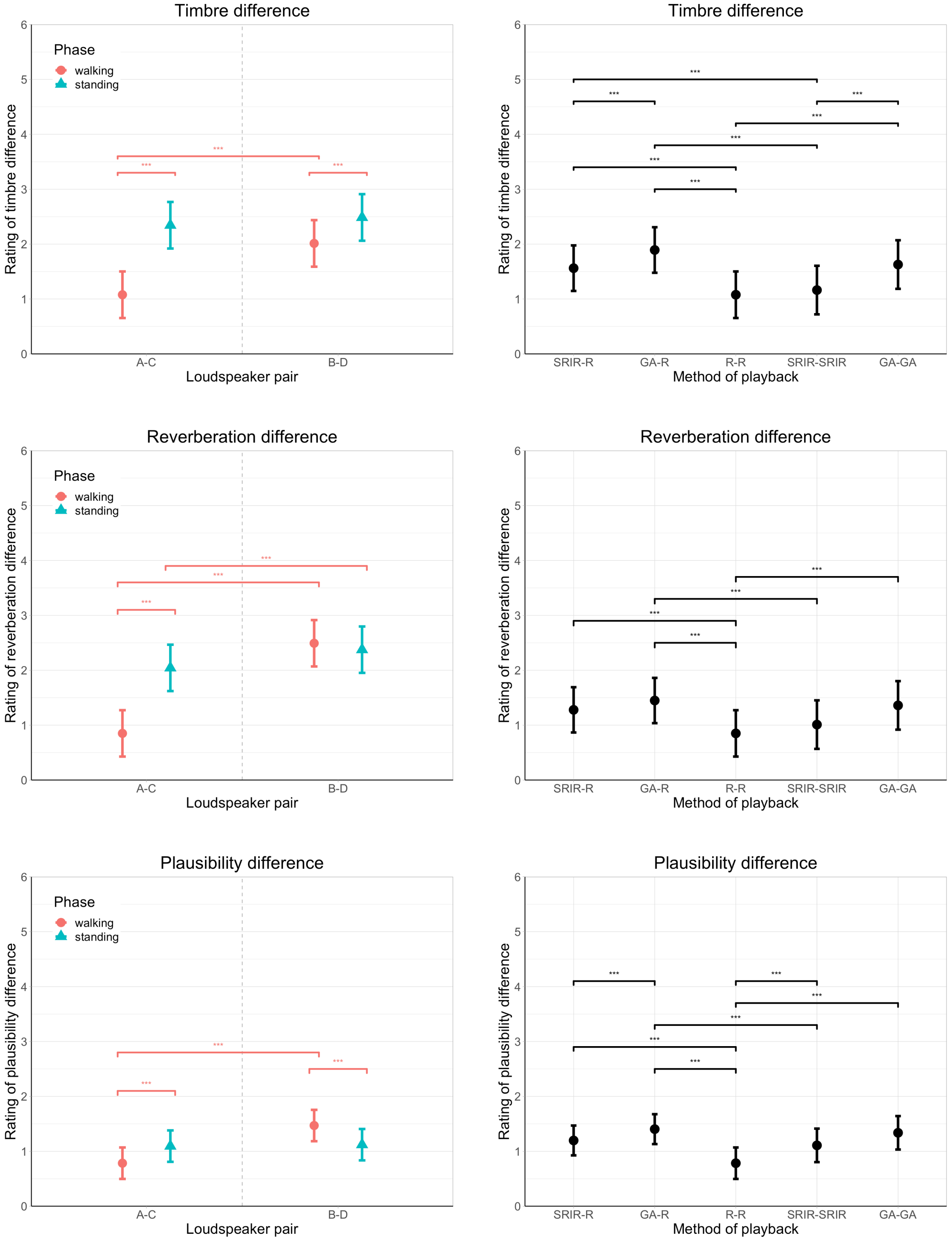


Figure 37: Predicted means and 95% confidence intervals for timbre, reverberation, and plausibility difference ratings according to the phase and loudspeaker pair (plots on the left) and the rendering method pair (plots on the right) (***P < 0.05)

Table 24

Reverberation difference: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	36.37	1.00	0.00
Method_type	62.26	4.00	0.00
Ldspkr_type	456.81	1.00	0.00
Phase	241.37	1.00	0.00
Ldspkr_type:Phase	145.44	1.00	0.00

LOUDSPEAKER PAIR (($\chi^2(1) = 145.44, P$<0.001), see Table 24. The effect of loudspeaker position was especially critical in the walking phase where the reverberation difference for pair A-C (1.28) was significantly lower than for pair B-D (2.92), see Figure 37. In the standing phase, the effect of the loudspeaker pair was weaker however similarly to the walking phase reverberation difference was statistically significantly lower for pair A-C (2.47) than for pair B-D (2.80). There was no significant difference between phases for pair B-D.

There was a significant influence of the playback METHOD PAIR on the ratings of reverberation difference ($\chi^2(4) = 62.26, P$<0.001). The lowest rating was obtained by a pair of two REAL speakers (R-R, 2.65) and two SRIR auralizations (SRIR-SRIR, 2.76). The highest rating of reverberation difference was observed on pair GA-R and GA-GA (3.23, 3.19 respectively). Pair SRIR-R was rated similarly with the mean estimate slightly lower than for pair GA-R (3.00).

Model selection analysis is provided in Table 89 in Appendix D. Refer to Tables 90-92 in Appendix D for pairwise comparison results.

Table 25

Plausibility difference: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	74.02	1.00	0.00
Method_type	69.01	4.00	0.00
Ldspkr_type	90.17	1.00	0.00
Phase	18.56	1.00	0.00
Index_id	8.50	3.00	0.04
Ldspkr_type:Phase	41.69	1.00	0.00

1.8 Plausibility Difference

To compare the results of timbre difference evaluation with plausibility ratings, for each trial, the plausibility difference was calculated between the two loudspeakers of a pair. The plausibility difference was obtained by taking the absolute value of the difference between the plausibility ratings for each stimuli pair. After that, a separate analysis was performed to allow for a direct comparison of plausibility difference with timbre and reverberation difference results (see Figure 37). The best model for predicting plausibility difference was: playback METHOD PAIR + PHASE * LOUDSPEAKER PAIR. There was a significant influence of the playback METHOD PAIR on the plausibility difference ($\chi^2(4) = 69.0, P$<0.001). The lowest difference was obtained by a pair of two REAL speakers (R-R, 1.47). The largest difference in plausibility rating was observed in pairs GA-R (2.09) and GA-GA (2.02). Pairs SRIR-R and SRIR-SRIR obtained similar values with the mean estimate slightly higher than for pair R-R (1.88, 1.79 respectively).

There was also a significant influence of the interaction of PHASE and LOUDSPEAKER PAIR on the plausibility difference (($\chi^2(1) = 41.7, P$<0.001), see Table 25. The ratings for pair A-C and B-D in the walking phase were significantly different, with pair A-C obtaining a lower difference (1.20) than pair B-D (1.88). In the standing phase, results for pairs of loudspeakers are very similar. Pair A-C obtained 1.51 and B-D 1.54.

Model selection analysis was provided in Table 93 in Appendix D. Refer to Tables 94-96 in Appendix D for pairwise comparison results.

1.9 Amplitude of Yaw Movement - Standing Phase

Table 26

Amplitude of yaw movement: Analysis of Deviance Table (Type III Wald chi-square tests)

	Chisq	Df	Pr(>Chisq)
(Intercept)	86.70	1.00	0.00
Method	4.07	2.00	0.13
Ldspkr	14.98	3.00	0.00
Index_id	37.12	3.00	0.00

During the standing phase of the experiment, participants were instructed to maintain a straight head position but were allowed to make small movements in case they were required to see all the loudspeakers. The tracking data from the second standing phase of the experiment was analyzed to determine the influence of different factors on the amplitude of the yaw movement of participants. For each trial in the standing phase, the movement of the participant's head on the yaw axis was analyzed to find the amplitude. The analysis of yaw amplitude data indicated heteroscedasticity, which violates the assumption of constant variance. To address this issue, a log transformation to the data was applied. The log transformation is known to stabilize the variance. The effectiveness of the log transformation was verified through diagnostic plots, which showed a notable improvement in the homogeneity of variances across the range of predictor variables.

The best model for predicting the amplitude of yaw movement was: the LOUDSPEAKER position + trial INDEX + playback METHOD. There was a significant main effect of LOUDSPEAKER ($\chi^2(3) = 14.98, P$<0.001), see Table 26. The amplitude of yaw movement was highest for loudspeaker A and lower for all other loudspeakers (see Figure 38).

There were also significant main effects of trial INDEX ($\chi^2(3) = 37.12, P$<0.001) and playback METHOD ($\chi^2(3) = 4.07, P$=0.0018). The amplitude of yaw movement was larger for the last set of trials (37-48) than all other sets of trials (see Figure 45). The amplitude of yaw movement was highest for the GA auralization method and lowest for REAL playback (see Figure 39).

Model selection analysis is provided in Table 97 in Appendix D. Refer to Tables 98-100 in Appendix D for pairwise comparison results.

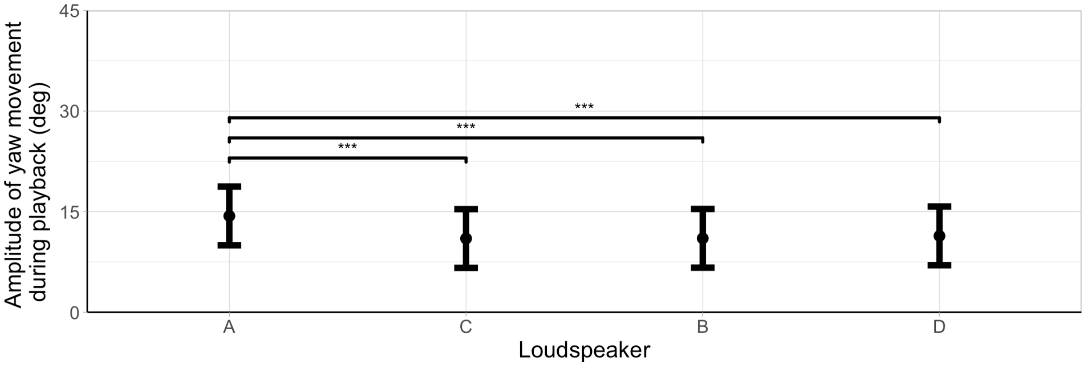


Figure 38: Influence of loudspeaker position on amplitude of yaw movement of participants

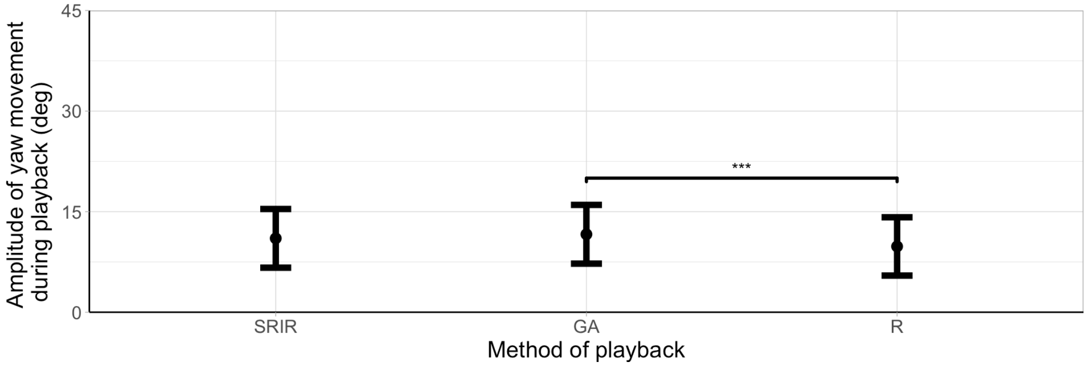


Figure 39: Influence of playback method on the amplitude of yaw movement of participants

1.10 Influence of Yaw Movement on Evaluation in Standing Phase

Results obtained during the standing phase of the experiment were analyzed to check the influence of the amplitude of yaw movement on subjective evaluation. The attributes evaluation was analyzed to investigate the influence of main factors: acoustic rendering method, localization of virtual and real loudspeakers, trial index, order of playback, height difference between participant and loudspeakers, and amplitude of yaw movement.

Influence of Yaw Movement on Plausibility Analysis of the results showed that the best model for predicting plausibility ratings (scale 0 - not at all plausible, 6 - very plausible) was: LOUDSPEAKER position * playback METHOD + amplitude of YAW MOVEMENT. There was a significant main effect of the amplitude of YAW MOVEMENT on plausibility ratings ($\chi^2(1) = 11.89, P$<0.001). The estimate for YAW MOVEMENT amplitude is -0.008 ± 0.002 point (p < 0.001). This implies that each additional 50 degrees of YAW MOVEMENT was associated with the decrease of the plausibility rating by 0.4 points (see Figure 40).

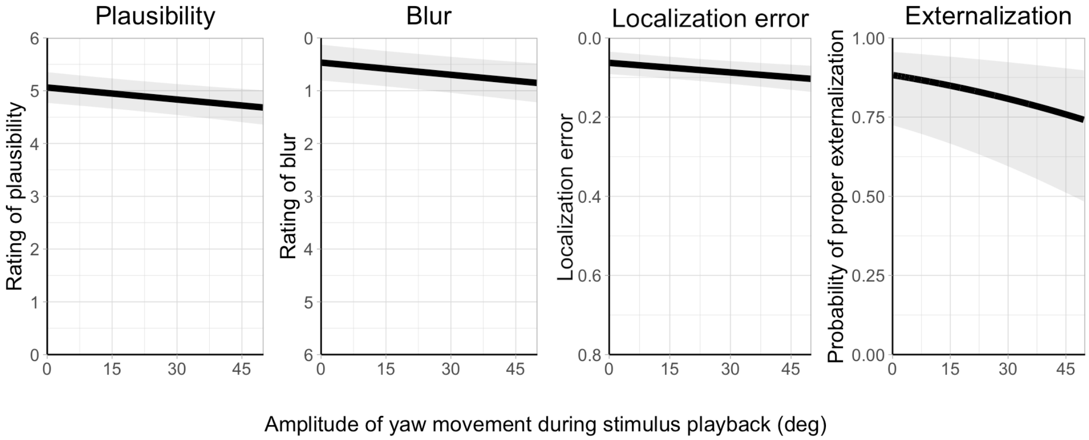


Figure 40: Influence of amplitude of yaw movement of participants on plausibility, blur, externalization ratings, and localization error in standing phase.

1.10.1 Influence of Yaw Movement on Blur

Analysis of the results showed that the best model for predicting blur ratings (scale 0 - very focused, 6 - very blurry) was: LOUDSPEAKER position * playback METHOD + amplitude of YAW MOVEMENT. There was a significant main effect of amplitude of YAW MOVEMENT on blur ratings ($\chi^2(1) = 12.0, P$<0.001). The estimate for YAW MOVEMENT amplitude is -0.008 ± 0.002 point (p < 0.001). This implies that each additional 50 degrees of YAW MOVEMENT was associated with the increase of blur rating by 0.4 points (see Figure 40).

1.10.2 Influence of Yaw Movement on Localization Error

Analysis of the results showed that the best model for predicting localization error (see section 1.3) was: LOUDSPEAKER position * playback METHOD + amplitude of YAW MOVEMENT + ORDER of playback. There was a significant main effect of amplitude of YAW MOVEMENT on localization error ($\chi^2(1) = 10.06, P = 0.0015$). The estimate for YAW MOVEMENT amplitude is 0.0008 ± 0.0002 m (P$<$0.001). This implies that each additional 50 degrees of YAW MOVEMENT was associated with the increase of localization error by 0.04 m (see Figure 40).

1.10.3 Influence of Yaw Movement on Externalization

Analysis of the results showed that the best model for externalization (see section 1.3) was: LOUDSPEAKER position * playback METHOD + amplitude of YAW MOVEMENT. There was a significant main effect of amplitude of YAW MOVEMENT on externalization ratings ($\chi^2(1) = 9.89, P = 0.0017$). The estimate for YAW MOVEMENT amplitude was -0.019 ± 0.006 m (P=0.0017), see Figure 40).

1.10.4 Influence of Yaw Movement on Loudspeaker Recognition Error

Analysis of the results showed that the best model for externalization (see section 1.3) was: LOUDSPEAKER position * playback METHOD + amplitude of YAW MOVEMENT. There was a significant main effect of amplitude of YAW MOVEMENT on loudspeaker recognition error ($\chi^2(1) = 6.99, P = 0.0082$). The estimate for YAW MOVEMENT amplitude was -0.012 ± 0.004 m (P=0.012). For the angles below 50 degrees the effect was minimal.

1.11 Influence of Height Difference

This section examines the impact of the absolute height differential between the listener and the loudspeaker on various attributes, including plausibility. Figure 41 illustrates the distribution of height differentials among participants, revealing that the majority of participants exhibit differences below 10 cm, while several participants noted differences between 15-25 cm. Analysis indicates that height difference significantly affected evaluations of plausibility, blur, localization error (refer to Figure 42), as well as externalization and timbre (see Figure 43). Height did not appear as a statistically significant effect in the previous Chapter, which exclusively covered results from the walking phase. However, the inclusion of a larger sample size encompassing both walking and standing phases in this analysis enabled us to discern the significant impact of height difference.

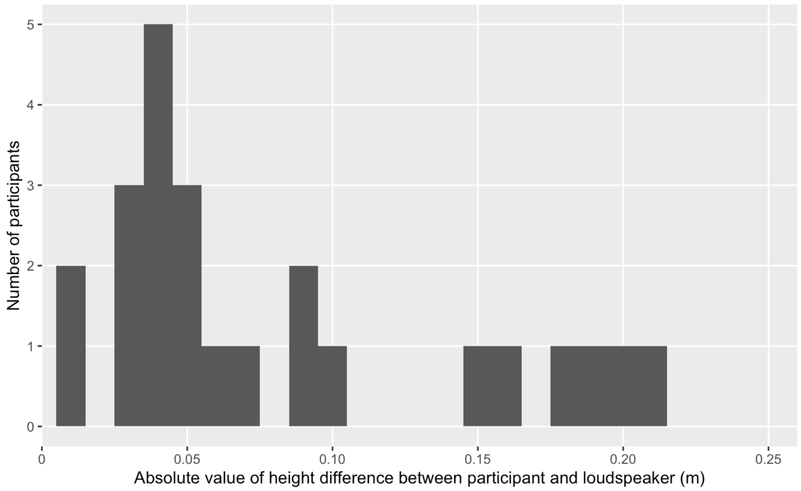


Figure 41: Histogram of the absolute value of height difference between participant and loudspeaker

Figure 42 illustrates the relationship between the HEIGHT difference between the participant and the loudspeaker (x-axis) and the predicted plausibility, blur, and localization error score (y-axis). The curve shows a decreasing trend, indicating that as the HEIGHT difference increases, the plausibility, blur, and localization error ratings tend to decrease.

For plausibility, the estimate for the HEIGHT difference was -0.028 ± 0.01 point (p

= 0.0368). This implies that for each additional 10 cm of height difference between the participant and the loudspeaker, the plausibility rating was lower by 0.28 points.

The blur rating estimate for HEIGHT difference was 0.048 ± 0.014 point ($\chi^2(1) = 11.54, p$<0.001). This implies that for each additional 0.1 m of height difference between the participant and the loudspeaker, the rating of blur increased by 0.48 points.

The localization error estimate for the HEIGHT difference was 14.4 ± 5.5 cm (p = 0.01141). This implies that for each additional 0.1 m of height difference between participants, position error increased by 14.4 cm (see Figure 42).

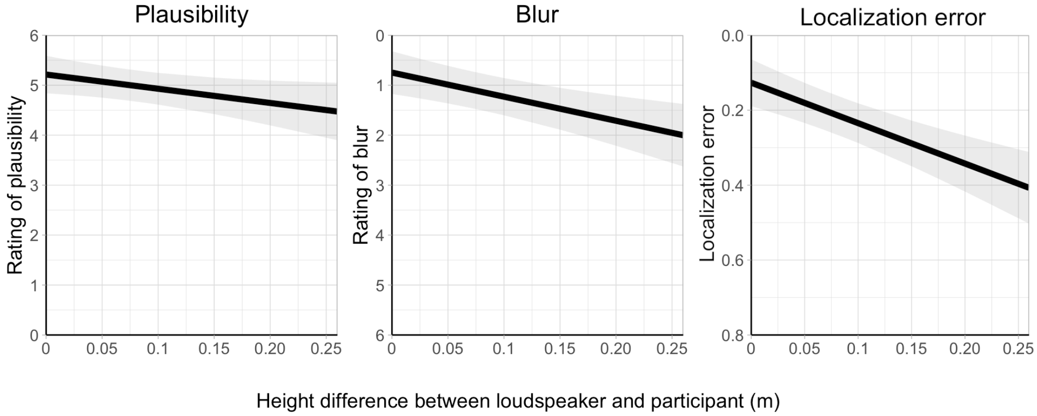


Figure 42: Influence of height difference between participant and loudspeaker on plausibility and blur ratings and localization error.

The externalization rating estimate for the HEIGHT difference was -0.08 ± 0.03 point (p = 0.0056). This implies that each additional 0.1 m of height difference between the participant and the loudspeaker decreased the probability of obtaining the answer ”outside of the head” for the externalization question by 0.08.

The timbre difference rating estimate for HEIGHT difference was 0.064 ± 0.019 points (p < 0.001). This implies that each additional 0.1 m of height difference between the participant and the loudspeaker increased the difference in timbre rating by 0.64

points. The timbre difference y scale was reversed to facilitate the comparison (see Figure 43).

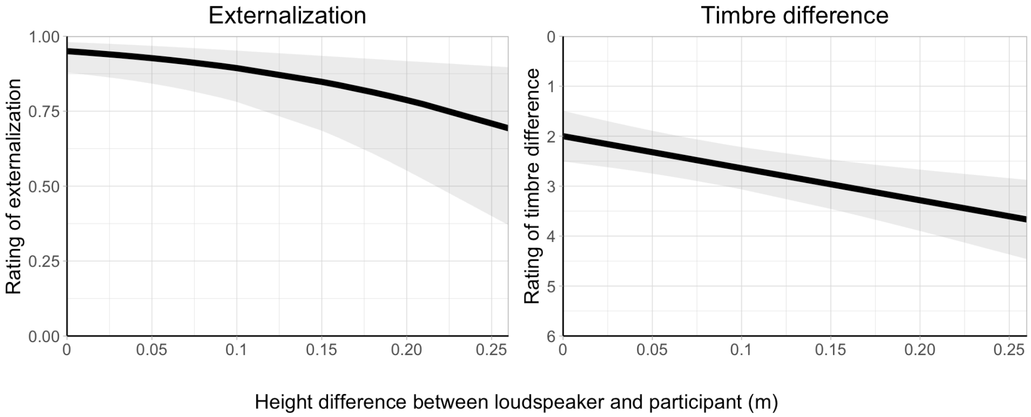


Figure 43: Influence of height difference between participant and loudspeaker on externalization and timbre difference ratings.

1.12 Influence of Trial Index

The analysis of the trial index aimed to explore the influence of the trial index and, consequently, participants' fatigue on perceptual evaluation. Analysis of ratings of blur and localization error revealed the statistically significant effect of the trial INDEX (see Figure 44). The influence of trial INDEX on blur rating was statistically significant ($\chi^2(3) = 23.6, P$<0.001) but the difference was minimal. The stimuli were rated as less blurry with the subsequent trials: there was a 0.24 point difference between the first (1-12) and last (37-48) section of trials.

The main effect of the trial INDEX on localization error was observed ($\chi^2(1) = 15.58, P$=0.0014). The estimate for localization error was 21% for trials 1-12, 17.8% for trials 13-24, 18% for trials 25-36, and 18.5% for trials 37-48. The position error was statistically significantly larger for the first 12 trials than for other groups of trials.

Trial INDEX had also a statistically significant influence on the amplitude of yaw movement ($\chi^2(3) = 37.12, P$<0.001). The amplitude of yaw movement was larger for the last set of trials (37-48) than all other sets of trials (see Figure 45).

1.13 Correlation between Plausibility and Other Attributes

One of the most important goals of the analysis is to examine the correlation between plausibility and other attributes assessed during the listening tests. Since certain variables were evaluated in pairs by soliciting a difference between the two stimuli (e.g., timbre difference), the disparity between ratings was computed for all attributes that received independent ratings for each stimulus within the pair (e.g., plausibility). This approach enabled the assessment of correlations among all attributes in a unified analysis.

To predict the correlation between plausibility and other attributes, we constructed a model encompassing all attribute differences. Additionally, individual models were created for each attribute difference to determine which explains the most significant amount of variance in plausibility difference. The results presented in Table 27 align with expectations, demonstrating that the best-fitting model is the one containing all attribute differences. This means that all of the attributes contributed to the perception of plausibility. Following this comprehensive model, the next most influential is the blur difference model.

Table 28 provides results of the Anova analysis of the comprehensive model. The findings indicate that each attribute independently influences plausibility and is statistically significant. The most substantial amount of variance is explained by blur,

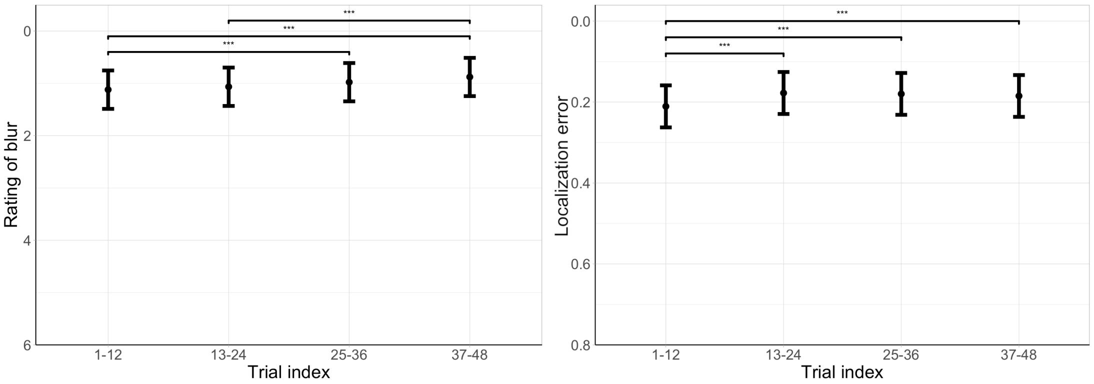


Figure 44: Predicted means and 95% confidence intervals for blur ratings, and localization error. The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry).

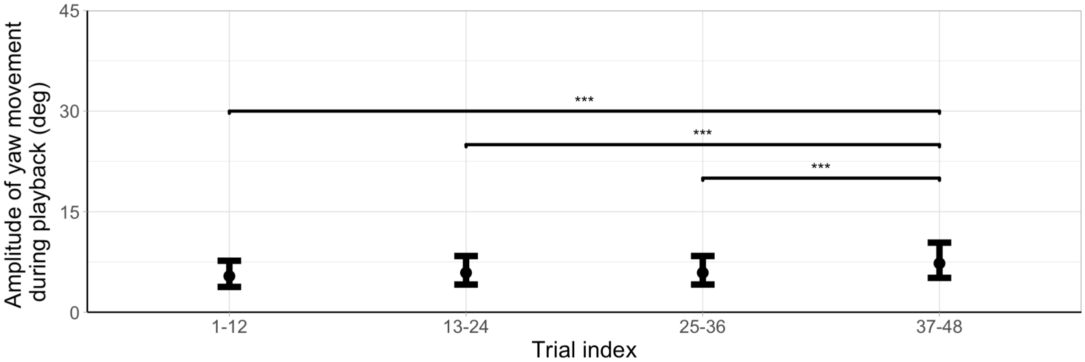


Figure 45: Influence of trial index on the amplitude of yaw movement of participants

followed by externalization ratings. Particularly, there is a variation in the order of importance between model selection and ANOVA results for the model containing all attributes. In model selection, localization error follows blur, while in Anova results, externalization comes after blur. This discrepancy likely arises from the correlation between localization error and blur, where both factors explain a shared portion of the variance. In the results of the ANOVA for comprehensive model, blur accounts for the largest amount of variance, and the results for localization error specifically indicate the variance explained solely by localization error.

Table 27

Model selection for plausibility difference

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
all	8	6723.80	0.00	1.00	-3353.87
blur difference	4	7288.19	564.39	0.00	-3640.09
localization error difference	4	7718.25	994.44	0.00	-3855.12
externalization difference	4	7823.23	1099.43	0.00	-3907.61
timbre difference	4	7830.16	1106.36	0.00	-3911.07
reverberation difference	4	7844.59	1120.79	0.00	-3918.29
null	3	8131.34	1407.54	0.00	-4062.67

Table 28

Analysis of Deviance Table (Type III Wald chi-square tests) for difference of plausibility ratings

	Chisq	Df	Pr(>Chisq)
(Intercept)	0.76	1.00	0.38
Blur Difference	477.14	1.00	0.00
Externalization Difference	142.30	1.00	0.00
Localization Error Difference	62.09	1.00	0.00
Timbre Difference	40.72	1.00	0.00
Reverberation Difference	10.66	1.00	0.00

2 Discussion

2.1 Influence of Participant's Movement on Subjective Evaluation

One of the important goals of this work is to validate the influence of participants' movement in 6 DoF on the evaluation of plausibility and other sound attributes. The comparison between the walking and standing phase of the experiment provides insight into this interaction.

2.1.1 Plausibility

The subjective ratings of plausibility yielded interesting insights regarding the impact of movement on plausibility evaluations, as depicted in Figures 29 and 30. During the walking phase, the statistically significant difference among the real loudspeakers was primarily observed between D and the remaining three, possibly due to the similar position of loudspeakers A, B, and C along the path. Participants approached, passed by, and moved away from these loudspeakers in a comparable manner. However, in the standing phase, significant differences emerged between A, B, and C, with A and C perceived as more plausible, particularly A, likely because of its proximity to the participant's standing position.

For the SRIR auralization, the plausibility evaluation pattern of loudspeakers in the walking phase mirrored the real rendering (with A, B, and C receiving similar scores and differing significantly from D), although with slightly lower scores overall. Conversely, during the standing phase, this pattern shifted, with A and B maintaining high scores while C and D converged with lower scores, suggesting a correlation with the distance from the participant's position. For the GA method, this trend was similar although with scores generally lower compared to the real scenario.

When comparing the results between the two phases, a statistically significant difference emerged when analyzing the data for loudspeaker D. Importantly, real loudspeaker D received lower plausibility ratings during the standing phase in comparison to walking. At the same time, auralization SRIR was rated as more plausible in the standing phase. It became apparent that participants were more proficient at identifying inconsistencies in the rendering of auralization SRIR while in motion. Conversely, when stationary, distinguishing between real loudspeaker D and auralization SRIR became challenging. Interestingly, auralization GA obtained statistically significantly lower plausibility ratings from both methods of playback in both phases.

The results indicate that there was no difference in plausibility evaluation between both phases for loudspeakers A and C (the only difference is that loudspeaker C rendered with GA method was significantly lower than A in standing phase). In contrast, assessment of loudspeakers B and D differed significantly between the two phases. First of all, loudspeaker B in walking phase was rated similarly to A and C, while loudspeaker D obtained lower ratings than other loudspeakers for all rendering methods. In contrast, during standing phase real loudspeaker B was rated lower than real loudspeakers A and C, and there was no difference between rendering methods. This effect must have been caused by the context in which this loudspeaker was presented as the ratings of blur and localization error indicate that there was some issue with localization for real loudspeaker B. Following section will focus on finding the link between this ratings and acoustic parameters. For method SRIR loudspeaker D was rated as plausible as real playback - this indicated that limited amount of cues were beneficial for SRIR method realism in this case. On the other hand, method GA was much less plausible than real playback similarly to walking phase.

These findings suggest that for loudspeaker D movement allowed participants to detect discrepancies in auralization SRIR and authenticity of real loudspeaker playback. Additionally, in walking phase, the parallax effect for loudspeakers A, B, and C was strong facilitating the plausibility evaluation. In standing phase, the results are more dependent on the relative comparison of the two sources.

2.1.2 Blur

The disparity between the two phases became more prevalent when analyzing the results of blur evaluation, as illustrated in Figures 29 and 30. In the standing phase, the stimuli were rated as increasingly blurry as the distance from the standing point increased. Consequently, for loudspeaker A, the ratings were highest, while for loudspeaker D, positioned farthest from the standing point, the ratings were lowest. During the walking phase, the ratings for loudspeakers A, B, and C were similar. This similarity could be explained by the fact that these loudspeakers were heard from similar distances, in contrast to loudspeaker D, which was positioned further away from the path.

When examining the differences between methods of playback, generally, the real loudspeaker consistently received lower blur ratings compared to other playback methods. The only exception occurred for loudspeaker B in the standing phase where both auralization methods were evaluated as less blurry than the real loudspeaker. This discrepancy might be attributed to the inaccuracy of sound localization revealed in the number of errors of loudspeaker recognition for loudspeaker B (see Section 1.4). In summary, the results highlighted that the difference between the two phases for evaluation of blur was primarily affected by two factors: the method of playback (auralizations being slightly more blurry than the real loudspeaker) and distance (greater distance between the listener and the loudspeaker resulting in higher blur ratings).

2.1.3 Localization Error

The results regarding localization error demonstrated a significant influence of the phase on the localization answers (refer to Figures 29 and 30). However, as discussed in Section 1.2 in Chapter VI the results of the analysis make it challenging to compare the two phases because of the constraints of the normalization method.

In walking phase the distance error is relatively small for all loudspeakers (the smallest error of loudspeaker D is an ”artifact” of the normalization).

Specifically for loudspeaker D during the walking phase, a substantial disparity emerged between auralizations and real playback, with the latter obtaining considerably smaller localization errors, as illustrated in Figure 30. This difference between methods of playback was much lower in the standing phase. The analysis indicated that the localization accuracy of auralizations for loudspeaker D posed more challenges during the walking phase than in the standing phase, which exhibited the interaction between motion and localization error evaluation.

In general, real playback consistently demonstrated smaller localization errors compared to auralizations, confirming this trend. However, there was one exception during the standing phase for loudspeaker B, where real playback received larger localization error ratings than auralizations. These findings aligned with the results from the blur evaluation. Diagrams depicting participants' responses to the localization question during the standing phase (see Figures 32 and 33) revealed that answers for loudspeaker B with real playback were skewed towards loudspeaker C, which means that participants frequently misidentified the loudspeaker that was playing (see the following section). This observation could be attributed to the similar angles at which loudspeakers were positioned when viewed from the standing point (refer to Figure 46). Consequently, the evaluation was based mostly on distance perception.

2.1.4 Loudspeaker Recognition Rate

Statistically significant differences in loudspeaker recognition rate between the two phases are evident in the data (refer to Figure 34).

The differences occur for:

- Loudspeaker A: Exhibited a lower probability of correct recognition during the walking phase compared to the standing phase. This result could be explained by the fact that during the standing phase, loudspeaker A was very close to the listener, which helped to avoid ambiguities in its location. In the walking phase, the most significant concentration of loudspeaker A errors was linked to loudspeaker C, as illustrated in Figure 35. This suggested a potential correlation between these errors and memory-related issues, given the consistent pairing of

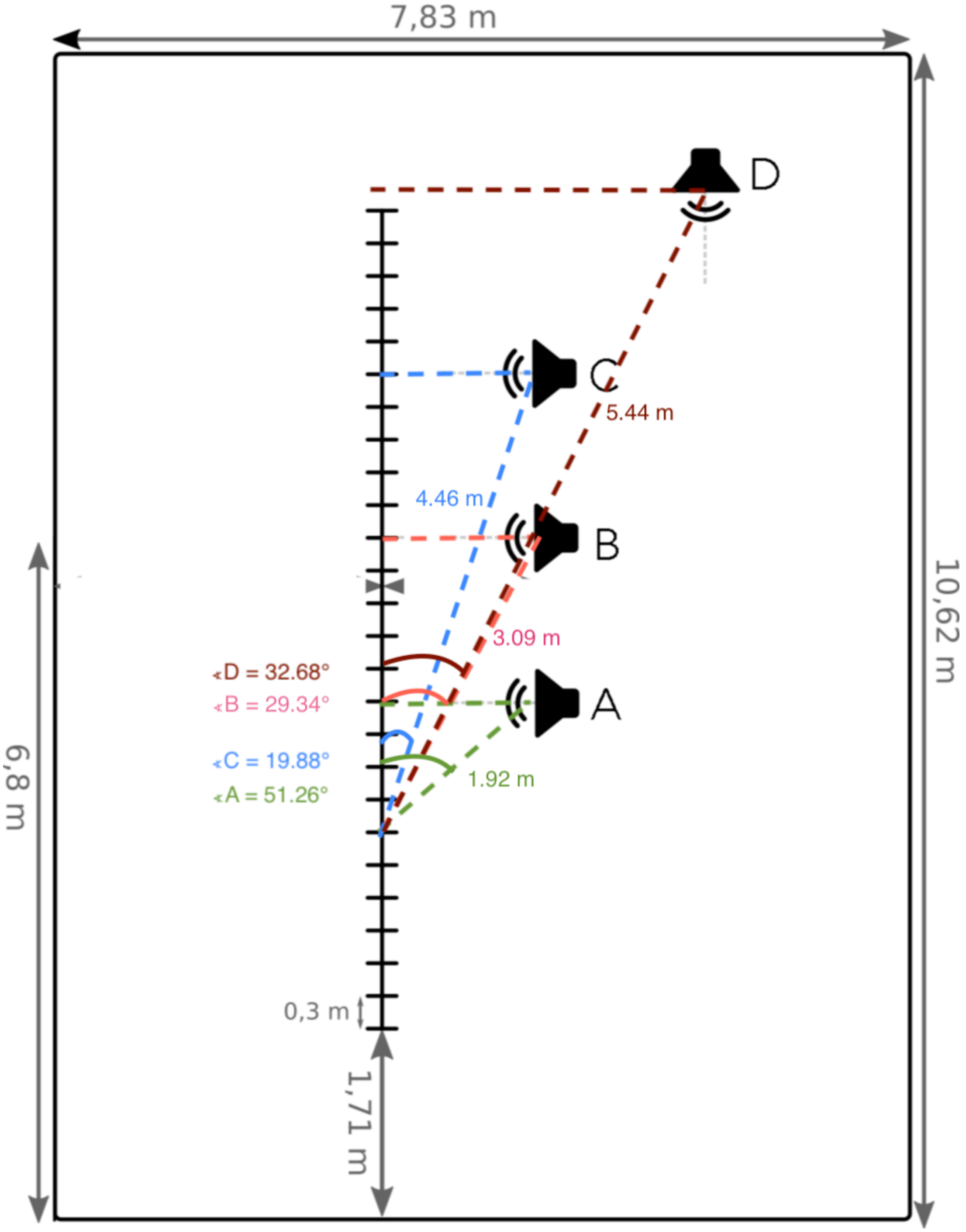


Figure 46: Room setup depicting angles and distances from the listening point to each of the loudspeakers during the standing phase

loudspeakers A and C. Here, memory issues affected remembering the order of playback.

- Loudspeakers B and D: A noteworthy discrepancy was observed with more errors occurring during the standing phase rather than the walking phase. The analysis highlighted that these errors were usually misidentifying loudspeaker C as the source of sound (refer to Figure 35). This effect might have been caused by the similar angle at which loudspeakers were positioned from the standing point. Because it was difficult to rely on angle detection, participants had to focus mostly on distance perception, which could lead to ambiguities in loudspeaker recognition (see Figure 46). This analysis was further explored in the section focused on the objective analysis of the auralizations.

In conclusion, motion facilitated correct recognition of the loudspeaker playing, which was proven by a lower number of errors of loudspeaker recognition in the walking phase thanks to the visual and auditory parallax effect.

2.1.5 Externalization

The results of the externalization evaluation were significantly influenced by the interaction between phase and the position of the loudspeakers, as depicted in Figure 36. In the standing phase, loudspeaker A appeared less externalized compared to all other loudspeakers. This observation could be linked to the proximity of loudspeaker A to the standing point, with 1.92 meters separating them. Consequently, it received the lowest externalization ratings. In contrast, the other loudspeakers, situated further from the standing point (e.g., loudspeaker B at 3.09 m, loudspeaker C at 4.46 m, and loudspeaker D at 5.44 m), obtained higher levels of externalization (refer to Figure 46).

During the walking phase, loudspeakers A, B, and C all received similar externalization ratings since they were perceived from approximately the same distance along the path. In contrast, loudspeaker D was rated as more externalized due to its placement further away from the walking path.

In summary, the primary factor influencing externalization evaluations was the distance between the listener and the loudspeaker, aligning with previous research on sound externalization. Many externalization evaluation methods, including this experiment, relied on the source's distance from the head, which served as a simplified representation of the listener's perception of distance (Durlach et al., 1992a). It remains uncertain whether sound externalization and distance perception are linked or distinct percepts (Best et al., 2020).

Additionally, the externalization could also be affected by the angular position of the sound source. Previous research indicated that the externalization increases for sources located on the side (Best et al., 2020). The sources in the walking phase were more lateral, while in the standing phase were mostly situated in front. The smaller distance between the listener and the source in the walking phase was contradicted by the lateral position of the source which enhanced externalization.

Furthermore, the impact of the method of playback on externalization was minor, as illustrated in Figure 36, suggesting that both auralization methods successfully achieved a high level of externalization. The only exception was auralization GA for loudspeaker D, which received lower externalization ratings compared to method SRIR and the real loudspeaker. This outcome is similar to the findings in the plausibility ratings, indicating a potential correlation between externalization and plausibility evaluations.

2.1.6 Timbre

The results of timbre evaluation demonstrated a significant influence of the phase on the ratings (refer to Figure 37). Interestingly, a significant interaction was found between the phase and the pair of loudspeakers. During the walking phase, the effect of the loudspeaker pair became much more evident. Specifically, loudspeaker pair A-C received much lower timbre difference ratings when compared to pair B-D. This discrepancy could be attributed to the distance of the loudspeakers from the walking path and their orientation: pair A-C had the same distance and direction, while pair B-D had different distances and directions.

In the standing phase, both pairs received similar ratings, but overall, the timbre difference was greater in this phase than during walking. This observation could also be linked back to distance and direction. While there was a relatively larger difference in distance between pair A-C and B-D, the additional difference in loudspeakers' orientation for pair B-D balanced out the disparities for both pairs during the standing phase. The overall lower difference in the walking phase comes from the fact that the four loudspeakers had a very similar timbre. In the walking phase participants heard the source from many perspectives which gave a more complete impression of the source directivity. In contrast, in the standing phase participant heard the source from only one perspective which was greatly influenced by the relative position of the listener and source.

In summary, the evaluation of timbre difference was primarily influenced by the distance and orientation of the loudspeakers. The movement of participants facilitated the perception of various perspectives on the sound source, resulting in a more complete impression of its directivity. As the directivity pattern was largely consistent between the four loudspeakers, it led to smaller ratings of timbre difference in the walking phase.

2.1.7 Reverberation

The results of reverberation difference were very similar to the timbre difference evaluation for pair A-C, as illustrated in Figure 37. The difference was much lower in the walking phase compared to standing. In contrast, for pair B-D, the ratings for both phases were very similar. The results could again be linked to the distance and position of the loudspeakers. During the walking phase, loudspeakers A and C were heard from nearly identical distances, while in the standing phase, the distance difference between them increased significantly. On the other hand, for loudspeakers B and D, the difference evaluation was affected mostly by their position in the room. The presence of loudspeaker D in a room corner, along with its distinct orientation, led to a perceivable difference from loudspeaker B in the reverberation, similarly for the two phases.

In conclusion, reverberation evaluation was mostly affected by the distance and position of the loudspeakers. These results were as anticipated, as alterations in distance modify the ratio between direct and reverberated sound. Also, the position of the loudspeaker closer to the walls created a very different pattern of early reflections, which could significantly impact the perception of reverberation.

2.1.8 Conclusions

The results indicated that the interaction between participants' motion and evaluation existed for all attributes. Motion had mainly a twofold influence on the evaluation:

- Facilitating the Distinction between Real and Virtual Sound

  The main effect of motion appeared to be its ability to improve sensitivity to disparities between virtual and real sound or to uncover differences that might remain unnoticed when stationary. The evidence of that was the most problematic loudspeaker - D. During the walking phase, it received lower ratings for auralizations and higher for real loudspeakers than in the standing phase. This suggested that participants' motion played a substantial role in identifying inconsistencies in the auralizations and assessing the realism of real playback. Importantly, the same effect was observed for localization error and blur.

- Distance as a Factor

  It is important to remember that the experimental design linked the motion with the participant's proximity to the sound sources. In several cases, the difference between the walking and standing phase was, in fact, the difference in distance from which the loudspeakers were heard. The distance affected mostly the perceptual attributes of sound including blur, localization error, and externalization, and did not influence the subjective judgment of plausibility.

In addition to these aspects, motion helped with solving ambiguities of loudspeaker position - there were fewer loudspeaker recognition errors during the walking phase compared to standing.

These results confirmed that during self-motion, listeners were capable of processing additional auditory cues and assessing whether these cues aligned with their expectations regarding changes in sound within a given environment. Incorporating the translation of the listener introduced several layers of detail to the auditory experience: altering the listener's position within the virtual room affected sound intensity and direct-to-reverberant ratio (DDR), both crucial cues for assessing distance. Besides that, the early reflection pattern was position-dependent. Surprisingly, maintaining a consistent spatio-temporal pattern of early reflections (as in the SRIR method) did not significantly degrade rendering quality. In fact, in certain aspects, participants rated this method as superior to the GA method, which simulated changing patterns of reflections within the room.

2.2 What Is the Influence of Secondary Effects on Sound Evaluation?

2.2.1 Height Difference between Participant and Loudspeaker

The height difference between participants' ear level and loudspeaker center had a significant impact on the subjective ratings. In particular, the larger the absolute difference between heights, the lower plausibility, blur, externalization, and higher timbre difference ratings, and localization error values were obtained (refer to Figures 42 and 43). This effect could be explained by the implementation of generalized HRTFs into the study. The difference in height between participants and the loudspeaker was actually compensated. This meant that for participants whose height was lower or higher than the loudspeakers, virtual sources were rendered accordingly using the corresponding HRTF direction. However, the use of non-individualized HRTFs has been shown to degrade localization in the vertical plane and increase the number of confusions along the cone of confusion (Wenzel et al., 1993). Confusion about localization affects the perception of blur, externalization, localization error, and, in consequence, plausibility. This effect was consistent with the results of the study by Mendonça et al. (2012), which showed that using non-individualized HRTF was degrading the sense of presence. Higher timbre difference ratings associated with a higher value of absolute height difference could also be explained by the use of non-individualized HRTFs in the study. The perception of elevation is based on spectral cues coming from the reflections off the pinnae and torso. The differences in spectral cues between the HRTF set used in the study and individual HRTFs of the listener resulted in the coloration of the sound (Merimaa, 2010; Takanen et al., 2012b). Consequently, there was a larger timbre difference between real and virtualized loudspeakers for participants with different heights than the loudspeaker where spectral cues had to be implemented.

2.2.2 Trial Index

The analysis of the trial index aimed to explore the influence of the trial index and, consequently, participants' fatigue on perceptual evaluation. The examination of results from both phases reaffirmed the conclusions drawn from the walking phase analysis. The trial index had a significant impact on the ratings of blur, and localization error as shown in Figure 44. Across these attributes, a consistent pattern emerged: the initial 12 trials received higher ratings for the blurriness, and exhibited larger localization errors. This trend suggests a learning curve during the initial trials. The trial index had also an influence on the behavior of participants. Participants' faster walking pace during the first 12 trials, as discussed in Chapter VI, Section 3.3.2, likely contributed to the increased difficulty in sound evaluation during this trial section. Moreover, the trial index had also a statistically significant influence on the amplitude of yaw movement. However, the effect appeared for the last set of trials which had a larger amplitude of yaw movement than all other sets of trials (see Figure 45). It might indicate that with increasing fatigue, participants made the task easier by allowing for more head movements.

2.2.3 Order of Playback

The analysis considered the order of playback, revealing its distinct impact on plausibility, externalization, localization error, and loudspeaker recognition errors. However, the influence varied among these attributes. The stimulus played first was evaluated as more plausible and externalized; however, it had larger localization errors and more loudspeaker recognition errors. The results suggest that memory errors played a significant role. The second stimulus, being more memorable, left a fresher impression on the participants' memories. In contrast, the first stimulus was not as well-remembered, occasionally causing confusion about which loudspeaker was playing and resulting in notable localization errors. Paradoxically, the less accurate memory of the first stimulus appeared to enhance its perceived plausibility and externalization, establishing a distinct contrast between the two stimuli.

2.2.4 Rendering Method Pair

The assessment of plausibility was influenced by the pairings of the rendering methods (see Figure 31). Participants assigned different ratings to a stimulus based on whether its counterpart in the pair was real or virtual. Surprisingly, the GA-GA rendering method pair received higher ratings than all other rendering methods for plausibility. This suggests that auralizations were rated more favorably when there was no real counterpart in the pair, even outperforming two real stimuli. This aligns with prior studies where stimuli presented without any real reference were consistently rated as nearly fully plausible while adding a real reference in the subsequent phase of the study degraded the plausibility scores significantly ((Neidhardt & Zerlik, 2021)). Interestingly, despite the brief time intervals between trials in this test which allowed the real reference to remain fresh in participants' memory, the ratings for virtual pairs still outperformed those with a real counterpart.

2.3 What Behavioral Measures Tell About the Interaction of Movement and Perceptual Evaluation?

2.3.1 Yaw Movement in Standing Phase

During the standing phase of the experiment, participants were instructed to maintain a straight head position but were allowed to make small movements in case they were required to see all the loudspeakers. The analysis of the tracking data from this phase revealed subtle yet statistically significant interactions between yaw amplitude and other factors. As expected, loudspeaker A positioned closest to the standing point required the largest head movements to include the loudspeaker in the field of view, as illustrated in Figure 38. The extent of yaw movement was also influenced by the method of playback. The greatest amplitude was observed for the GA method, while the smallest was observed for real playback, as shown in Figure 39. This observation can be related to the fact that larger head movements facilitate the detection of auralization. Previous studies have indicated that head movements facilitate sound source localization and improve localization accuracy (Wallach, 1940; Kato et al., 2003; Honda et al., 2013), as well as assist in resolving front-back confusions (F. Wightman et al., 1994; Perrett & Noble, 1997)) and have been employed to facilitate the evaluation of tracking delays (Yairi et al., 2007)). The GA method appeared to introduce more perceptual ambiguities that were more easily resolved with head movement. In contrast, real playback was easier to evaluate and participants did not require head movement to aid in evaluation. The statistically significant interaction between yaw movement and the evaluation of sound attributes reinforces this hypothesis, as depicted in Figure 40. The amplitude of the yaw movement had a negative impact on the assessment of plausibility, blur, localization error, and externalization. In general, ratings for plausibility, blur, localization error, and externalization were lower for auralizations than for real playback, aligning with the observation that participants moved their heads less during real playback.

Figure 45 illustrates the difference in yaw amplitude between trial sections. The last section of trials revealed higher yaw movement amplitude than all other sections of trials. It might indicate that with increasing fatigue, participants made the task easier by allowing for more head movements.

2.3.2 Conclusions

The analysis of tracking data from both phases reveals a consistent pattern in participants' behavior. In the walking phase, it was observed that participants moved at a faster pace during real playback. Consequently, higher scores for sound attributes, including plausibility, blur, localization error, and loudspeaker recognition rate, were associated with increased walking speed (as discussed in Section 3.3.2). Similarly, in the standing phase, there was a reduction in the amplitude of yaw movement during real playback. This also led to higher ratings for plausibility, blur, localization error, and loudspeaker recognition rate. Participants adjusted their walking speed and minimized head movements when the task was comparatively easier, as seen during real playback. Conversely, during auralization playback, they adopted a slower pace and employed more head movements to cope with the more challenging task.

These findings suggest a meaningful correlation between the stimulus and the behavior of the participants. Importantly, in future studies, an analysis of listener behavior may yield insight into the nature of the stimulus they are perceiving. This approach can be very beneficial for studies where a large number of variables need to be tested. By collecting the behavioral data, the test duration might be greatly reduced. However, additional research is necessary in this area.

2.4 Is the Proposed Methodology an Effective Method for Evaluating Plausibility and Other Sound Attributes in 6DoF AR Environments?

This work presents a novel experimental design for plausibility evaluation with self-translation of the user. The design of the study focused on resembling a real-life scenario of AAR where real sounds are present in the scene but do not allow for a direct comparison, as they may originate from different types of sources (e.g. different voices) and might not be heard from the same perspective. In the experimental design, the stimuli were played in pairs by real and virtual loudspeakers with the same acoustical characteristics, but with varying positions. The following sections discuss the effectiveness of the proposed approach.

2.4.1 Tuning Internal Reference

The results from the experiment underscore the subjectivity of sound plausibility, which is greatly influenced by the specific context in which the sound is presented. As previous studies have noted, plausibility evaluation revolves around the agreement between perceived sound event and internal reference shaped by previous sound experiences (Lindau & Weinzierl, 2012; Pike et al., 2014). Consequently, in the context of AAR, the internal reference can be tuned by the real sound environment around the listener.

A previous study by (Neidhardt & Zerlik, 2021) already proved that including real counterparts in the plausibility evaluation task influenced the perception of plausibility. The methodology of this study provided more insight into the effect of tuning internal reference. Using two pairs of loudspeakers positioned differently, enabled the comparison of how plausibility is perceived when stimuli are presented with similar or different real counterparts. The results of the test proved that in the case of similarly positioned loudspeakers, distinguishing the auralizations was much easier than in the pair with different positioning. It shows that by manipulating the real reference within the testing pair, the difficulty of the test can be adjusted. This adaptability can prove especially valuable when deploying this test in real-life assessments of AAR systems, where the desired rendering accuracy may vary according to the specific use case. Depending on the objectives of the experiment, the participant's internal reference can be adjusted by altering the loudspeaker's position within the pair.

2.4.2 Evaluation Method

Previous studies on plausibility evaluation implemented two approaches to the questionnaire task. The first approach uses a Yes/No paradigm with the results' analysis based on the Signal Detection Theory (Lindau et al., 2007; Lindau & Weinzierl, 2012; Neidhardt & Knoop, 2017; Wirler et al., 2020; Neidhardt & Zerlik, 2021). The second approach implements categorical scales (Neidhardt et al., 2018; Enge et al., 2020; Neidhardt & Kamandi, 2022) where participants are asked to rate the plausibility on a scale between 1-7 or 1-4.

The results of this study revealed that using the scales might allow to detection of more subtleties of the plausibility evaluation in comparison to the Yes/No Paradigm. The results indicated that there was a ”grey area” of plausibility perception where the sound was not fully plausible but at the same time not fully implausible. In this experiment, there were instances when even real sounds were not rated as entirely plausible. The

Yes/No Paradigm, by forcing a binary choice between real and virtual, might overlook valuable information regarding plausibility perception.

One limitation of using categorical scales is that participants may refrain from using extreme values. However, this effect is mitigated by the selected analytical approach, where individual participants' scale responses were considered in variability analysis as a random effect.

2.4.3 Walking Predetermined Path

When designing an experiment in 6 DoF, careful consideration of participant movement within the room is essential. The nature of this movement has been demonstrated to have a significant impact on perceptual sound evaluation, as evidenced in prior studies by Hendrickx et al. (2017); Enge et al. (2020), and is further affirmed by the results of the current study. Furthermore, the act of head rotation has proven beneficial in resolving front-back confusion (Blauert & Butler, 1985; F. Wightman et al., 1994) and plays an important role in assessing system latency (Yairi et al., 2007). Previous studies on plausibility evaluation included different approaches to participants' movement. In the study by Neidhardt and Kamandi (2022), participants were allowed to freely rotate their heads while walking a predetermined path. In another study, participants could freely explore the space as many times as they needed (Neidhardt & Knoop, 2017).

However, allowing participants to explore space arbitrarily has two main disadvantages. First, it can result in each participant adopting a distinct hearing perspective, potentially influencing the evaluation. Without comprehensive movement analysis, distinguishing whether differences among participants stem from their varied movements or other factors becomes challenging. Secondly, with free exploration the time of each trial increases substantially thus it limits the amount of trials possible to complete during one session. On the other hand, with constraints on the participant's movement, the evaluation is affected by the type of movement determined by the investigators. However, if movement is taken into account during analysis it may lead to meaningful conclusions.

In the dissertation experiment, participants were asked to walk a predetermined path without rotating their heads. The aim was to listen to the full stimulus while walking. Additional visual cues were added to help adjust the speed of walking. This approach ensured that all participants experienced very similar hearing conditions. Small disparities in the movement behavior were accounted for by analysis of the tracking data which led to interesting conclusions (see Section 2.3).

2.4.4 Auralization Methods

It is important to emphasize that the primary objective of the experiment was not to evaluate the auralization methods; rather, these methods were chosen to assess plausibility and evaluate the methodological approach. Both auralization methods shared a common implementation for simulating direct sound, yet differ in the simulation of early reflections and late reverberation. Both methods effectively achieved the planned goals — they were sufficiently imperfect to discern differences in perception between virtual and real sounds. This is depicted in Figure 30, where the differences in plausibility scores for loudspeakers A, B, and C are subtle but statistically significant. At the same time, both methods remained within the realm of realism, presenting a challenging task even for experts during the listening sessions. Interestingly, there were instances where the real loudspeaker received lower ratings than its virtual counterparts, as exemplified in Figure 34, where real loudspeaker B exhibited the highest rate of position errors. Additionally, in Figure 30, real loudspeaker B is rated as more blurry and receives higher localization error scores compared to both auralization methods.

Overall, both simulations yielded relatively high plausibility ratings, especially for loudspeakers A, B, and C, given their simplifications.

2.5 Which Attributes Contribute to the Plausibility Perception?

Results of the analysis looking at the correlation between plausibility difference and other attribute differences reveal that all of the attributes contribute independently to the explanation of variance in plausibility. Blur ratings explained the largest portion of variance followed by localization error which was partially correlated with blur, and then externalization, timbre, and reverberation. The blur and localization error ratings were driven mainly by the ambiguity between the perceived location of the sound source and the visual anchor. It seems that this issue was crucial for the evaluation of plausibility. The next most important factor was externalization.

The most important conclusion is that plausibility is a complex subjective percept. As it relates to the comparison of sound to the inner reference, it encompasses different aspects of sound. All of the factors evaluated in this study contributed to the plausibility perception. The fact that the highest correlation with plausibility occurred for blur and localization error proves that congruency between visual anchors and sound is crucial for plausibility. The next most important aspect of sound was externalization. Externalization is one of the most important aspects of binaural sound which can be easily disrupted (Best et al., 2020). It is also very sensitive to room divergence as proven by several studies (Werner et al., 2016; Klein et al., 2021). The least important attribute was reverberation. This result is in line with previous studies which showed that listeners are not very sensitive to changes in reverberation (Shinn-Cunningham & Ram, 2003).

3 Objective Analysis

The objective evaluation of both auralization methods and reference measurements of KU100 along the path was conducted to investigate the acoustic differences between them. This analysis aims to characterize auralization methods and real loudspeakers measurements to later find a link between objective characteristics of different rendering methods and subjective ratings of participants. The goal of the analysis is to answer the question of which acoustical factors have the biggest impact on sound perception in the AR context, particularly for plausibility perception.

3.1 Energy of Time Segments for Each Measurement Point

In the initial phase of the analysis, the time distribution of energy along the path was examined. The energy was calculated from BRIRs for each loudspeaker and point along the walking path. BRIRs were calculated from the measurements done with the KU 100 dummy head and from the output of the auralization scripts implemented in Max/MSP. The KU 100 measurements were carried out without headphones that were worn by the participants during listening tests. That is why to ensure comparability of the auralization BRIRs with measurements, the headphone compensation filter was disabled during the recording process. The A-weighting filter was applied to the signal before energy calculation.

Figure 47 illustrates the mean energy of the left and right channels derived from binaural impulse responses (BRIR), divided into three time segments: 0-5 ms (direct sound), 5-80 ms (early reflections), and >80 ms (late reverberation). The duration of these time segments was aligned with the implementation in both auralization methods (see Chapter V). Noteworthy, the plots for loudspeakers A, B, and C exhibit striking similarities. The evaluation of direct sound energy of both auralization methods has a smoother contour than the real loudspeaker which means that energy increases at different speeds when approaching the loudspeaker. This difference might come from the inaccuracy of the directivity model implemented in auralizations. Across all, there is an approximate 2 dB disparity in the early reflection level between method GA and R, while method SRIR closely mirrors the real loudspeaker. Additionally, around 3 dB difference is observed in late reverberation level between the real loudspeaker and auralizations. The analysis of loudspeaker D reveals more differences. The direct sound energy for both auralizations is higher (1-4 dB difference) than for real playback. Early reflections energy fluctuates along the path, initially with similar energy levels for method GA and R, and lower for method SRIR. Toward the end, method GA surpasses methods SRIR and R (around 2 dB).

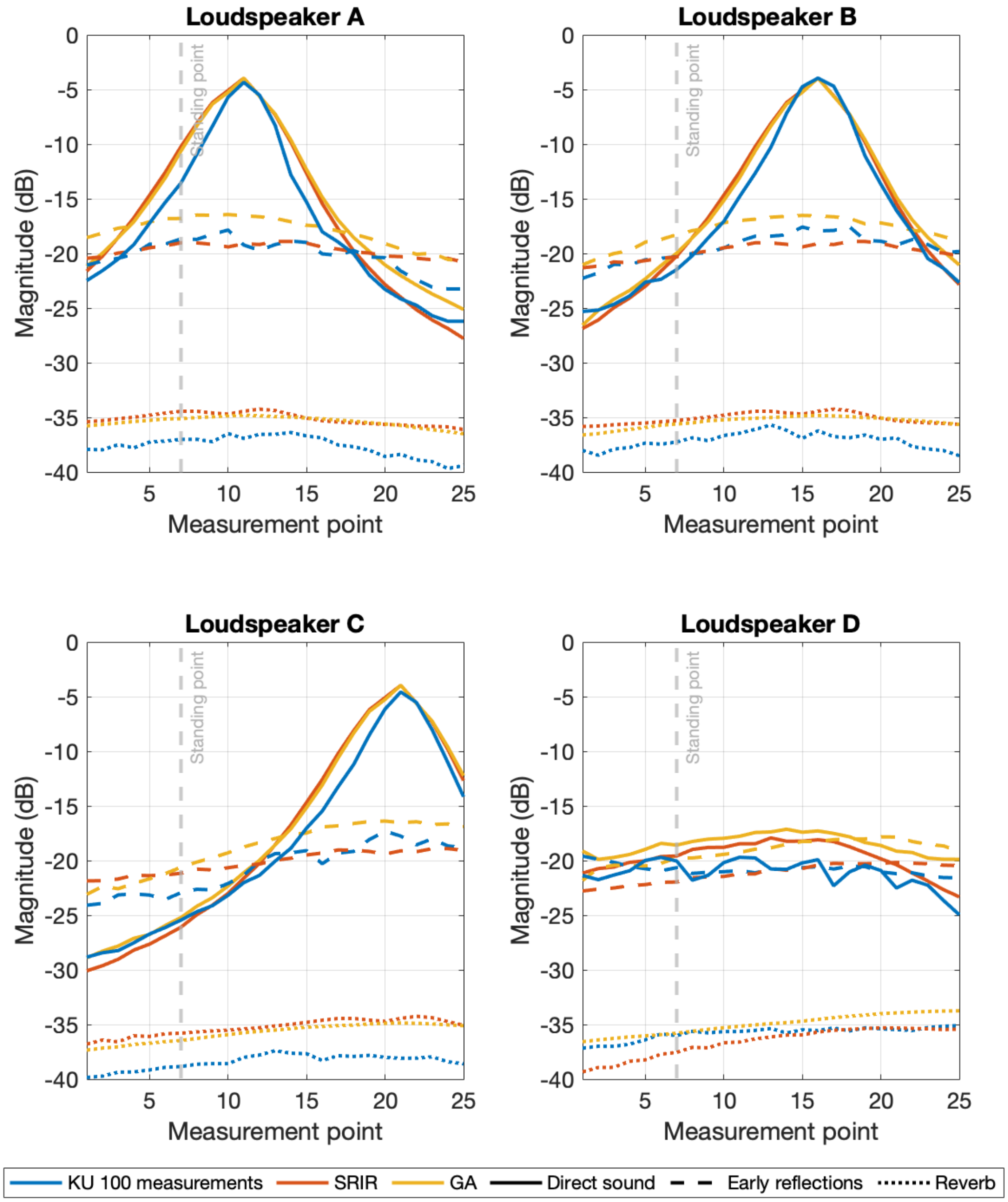


Figure 47: Energy of time segments for each measurement point for all loudspeakers (0-5 ms, direct sound, 5-80ms - early reflections, >80ms - late reverberation). The vertical line denotes the point at which participants were positioned during the standing phase.

3.2 Total Energy of the Signal

The total energy of a signal encompasses the cumulative energy from its initial onset. Differences in total signal energy reflect variations in intensity, a crucial cue for distance perception. Figure 48 shows the total energy of the signal for each measurement point along the path. For loudspeakers A, B, and C the total energy is driven mainly by direct sound energy, in contrast to loudspeaker D where reverberated sound plays a much more important role. Across loudspeakers A, B, and C, the disparity between auralizations and real playback remains consistent, around 2 dB before and after reaching the closest proximity to the loudspeaker. However, loudspeaker D exhibits different characteristics. The energy for the GA method was consistently higher from real playback, peaking at approximately 4 dB. Initially, the energy levels between method SRIR and real playback are comparable but diverge after reaching the standing point, with method SRIR reaching a peak difference of roughly 2 dB.

The black lines on the plots indicate the theoretical values of energy evolution according to the distance law. Because of the directivity pattern, the energy evolution of loudspeakers A, B, and C is ”sharper” than the ideal distance law. Furthermore, the directivity of loudspeaker D results in the distance law no longer being adhered to after point 15. Consequently, despite geometrically approaching the source, the sound level begins to decrease, contrary to the predictions of the distance law.

3.3 Direct-to-Reverberant Ratio

Figure 49 shows the analysis of the Direct-to-reverberant ratio (DRR) for each measurement point. DRR represents the ratio of direct sound (between 0 and 5 ms) and

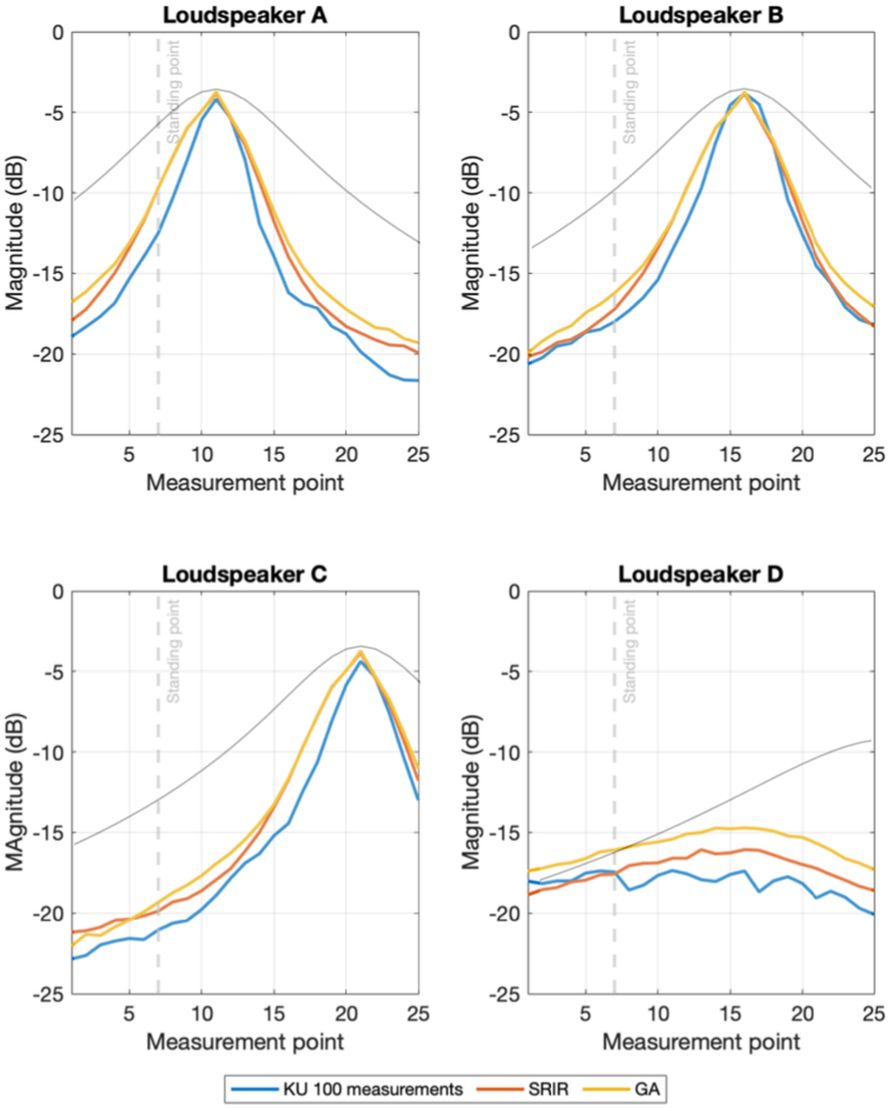


Figure 48: Total energy of the signal for each measurement point for all loudspeakers. The vertical line denotes the point at which participants were positioned during the standing phase. The black line indicates the theoretical values of the energy evolution according to the distance law.

reverberated sound energy (which arrives later than 5 ms). For BRIRs a combined DRR value was calculated using the ratio of the summed left and right direct signal energies, to the summed left and right reverberant signal energies, as shown below:

$$DRR(dB) = 10 * log(\frac{Energy_{right}(0-5ms) + Energy_{left}(0-5ms)}{Energy_{right}(5ms-end) + Energy_{left}(5ms-end)})$$

DRR values for loudspeakers A, B, and C are very similar between simulation SRIR and R. Simulation GA has a lower amplitude of change along the path. DRR is higher at the beginning and end of the path and lower when approaching the loudspeaker. Furthermore, for loudspeaker D the DRR values are highest for method GA and lowest for real playback for the majority of points along the path. It indicates that the energy of the direct sound of loudspeaker D was highest for the GA method.

3.4 C_{50}

Figure 50 shows the analysis of C_{50} for each measurement point. C_{50} is related to the attribute clarity or intelligibility of speech and represents the ratio of the early sound energy (between 0 and 50 ms) and the late sound energy (that arrives later than 50 ms).

$$C_{50}(dB) = 10 * log(\frac{Energy(0-50ms)}{Energy(50ms-end)})$$

C_{50} values for loudspeakers A, B, and C are very similar between simulation SRIR and R. Simulation GA has a lower amplitude of change along the path: C_{50} is higher at the beginning, and end of the path and lower when approaching the loudspeaker. For

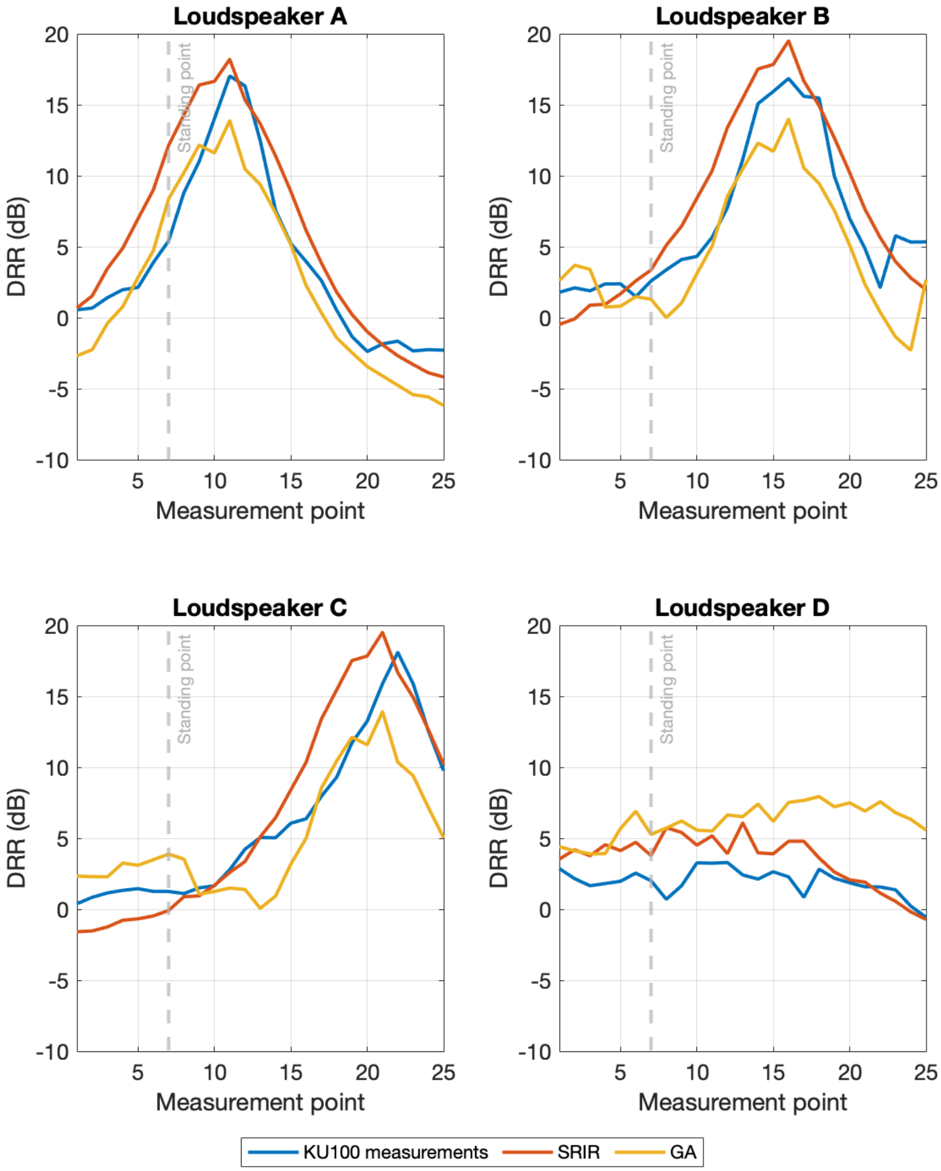


Figure 49: DRR for each measurement point for all loudspeakers (sum of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase.

loudspeaker D, GA has consistently higher values of C_{50} than rendering methods R and SRIR.

3.5 C_{80}

Figure 51 shows the analysis of C_{80} for each measurement point. C_{80} is related to the music clarity and represents the ratio of the early sound energy (between 0 and 80 ms) and the late sound energy (that arrives later than 80 ms).

$$C_{80}(dB) = 10 * log(\frac{Energy(0 - 80ms)}{Energy(80ms - end)})$$

The differences between rendering methods for loudspeakers A, B, and C are minimal. However, there is a more pronounced distinction between methods for loudspeaker D - both auralization methods yield higher C_{80} values compared to real playback, by approximately 3 dB.

3.6 $[1 - IACC_{E_3}]$ - Apparent Source Width

Early interaural cross-correlation coefficient ($IACC_E$) represents the difference between the left and right binaural signals during the first 80 ms of IR. $IACC_E$ near value 1 means that the source is exactly in front or behind the listener. $IACC_E$ of value 0 means that there is no correlation between signals. $IACC_{E_3}$ is derived by averaging the $IACC_E$ across three-octave bands with mid-frequencies of 500 Hz, 1 kHz, and 2 kHz. This metric gives preference to frequency ranges where wavelengths are similar to or smaller than the acoustical distance between the two sides of a head. ASW (Apparent

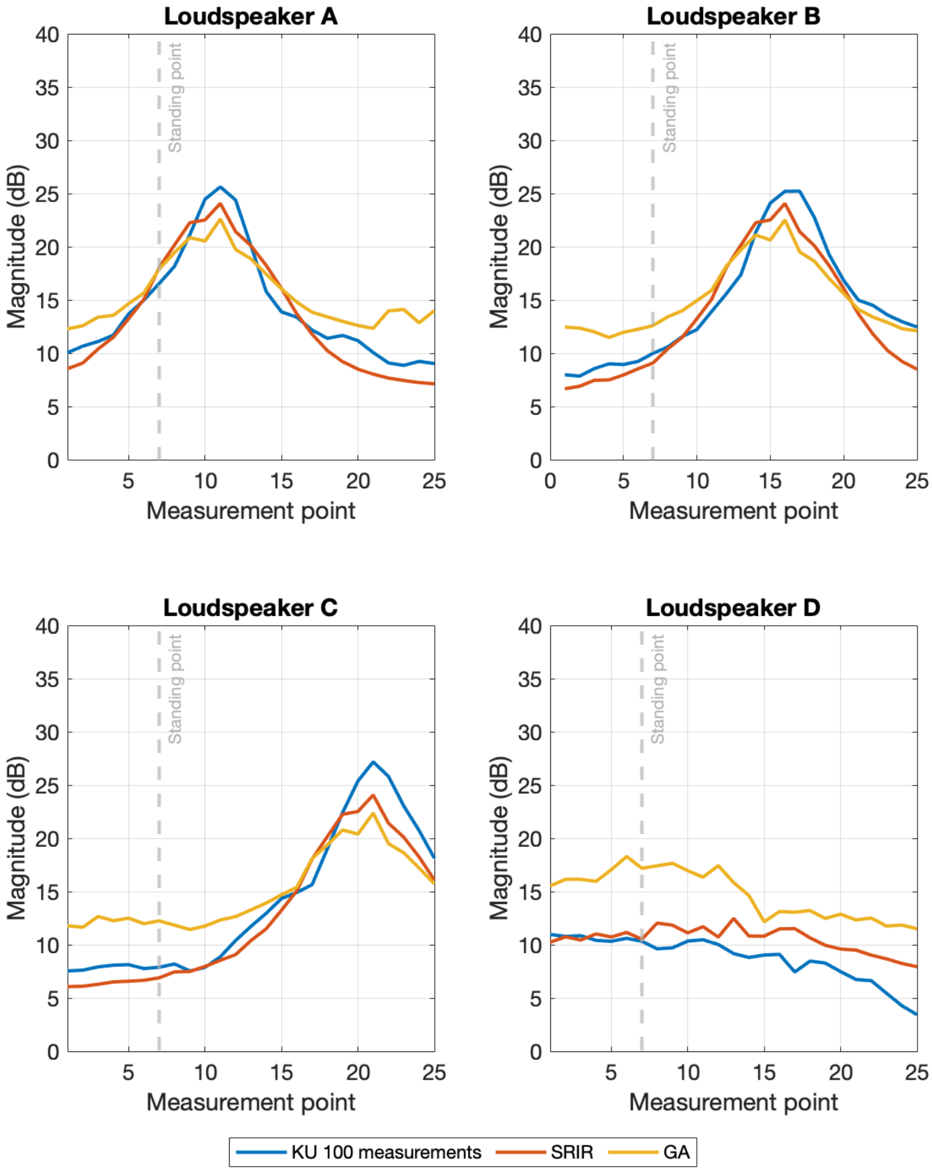


Figure 50: C_{50} for each measurement point for all loudspeakers (mean of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase.

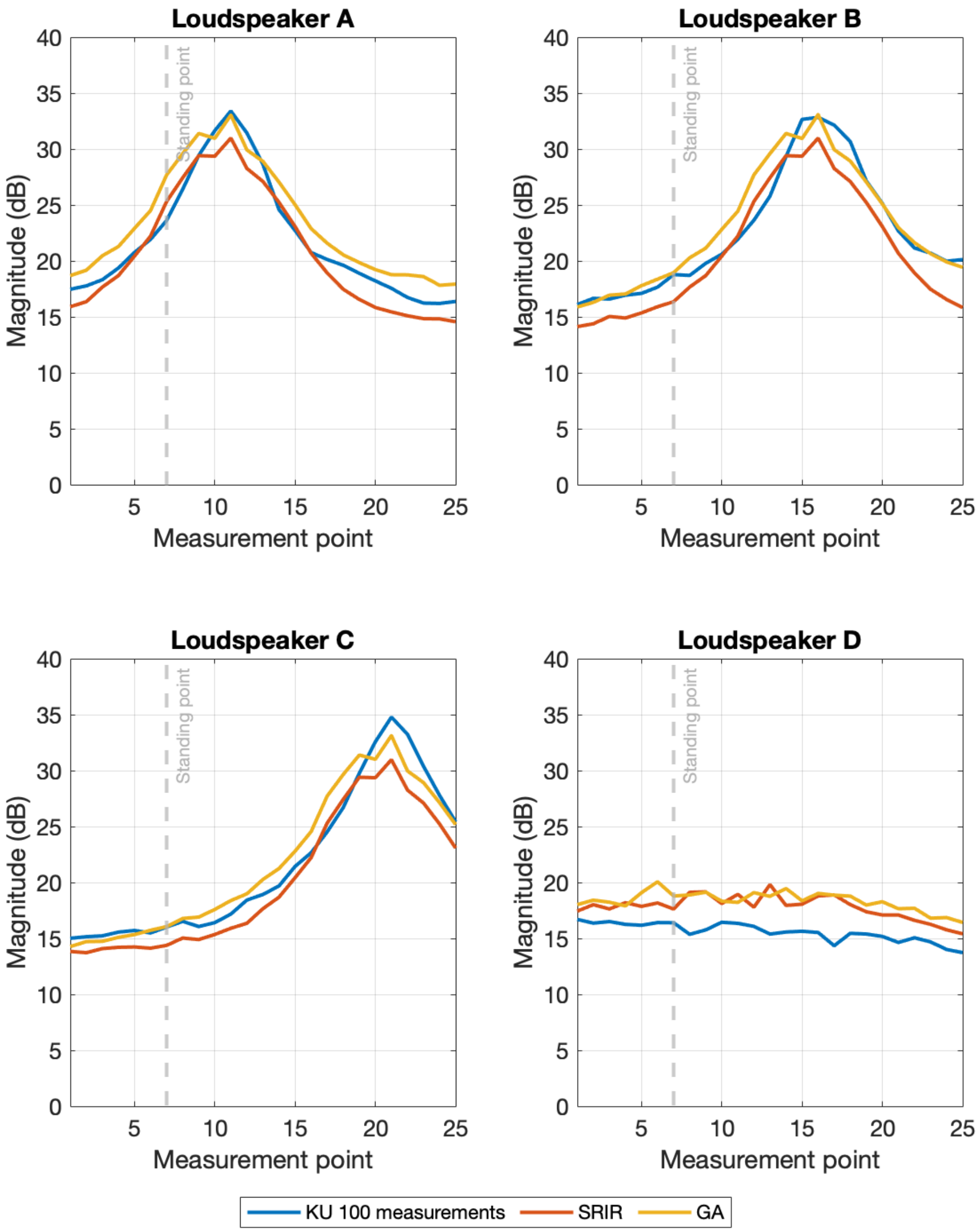


Figure 51: C_{80} for each measurement point for all loudspeakers (mean of L and R channels). The vertical line denotes the point at which participants were positioned during the standing phase.

Source Width) which is a subjective measure of the perception of sound source width has been shown to be directly correlated with $[1 - IACC_{E_3}]$ (Beranek, 1995; Okano et al., 1998). The larger values of $[1 - IACC_{E_3}]$ are correlated with the perception of the wider sound source. The results of $[1 - IACC_{E_3}]$ calculation for different methods of playback and loudspeaker D are shown in Figure 52. Plots indicate that values of $[1 - IACC_{E_3}]$ are larger for methods R and SRIR than for method GA.

3.7 Diffuseness

The analysis of diffuseness was done on SRIR measurements performed with Eigenmike along the path (refer to Figure 53). The SRIR of each point and loudspeaker was convolved with the speech stimulus implemented in the study. Then the direction and magnitude of the intensity vector were estimated based on 4 first components of the HOA stream.

For loudspeakers A, B, and C the diffuseness reaches its minimum when at the minimum distance to the loudspeaker. However, for loudspeaker D diffuseness is almost constant and higher than all of the other loudspeakers. Means of diffuseness along the path are similar for loudspeakers A, B, and C and lower than the mean value of loudspeaker D. It indicates that loudspeaker D was in general more diffuse than other loudspeakers.

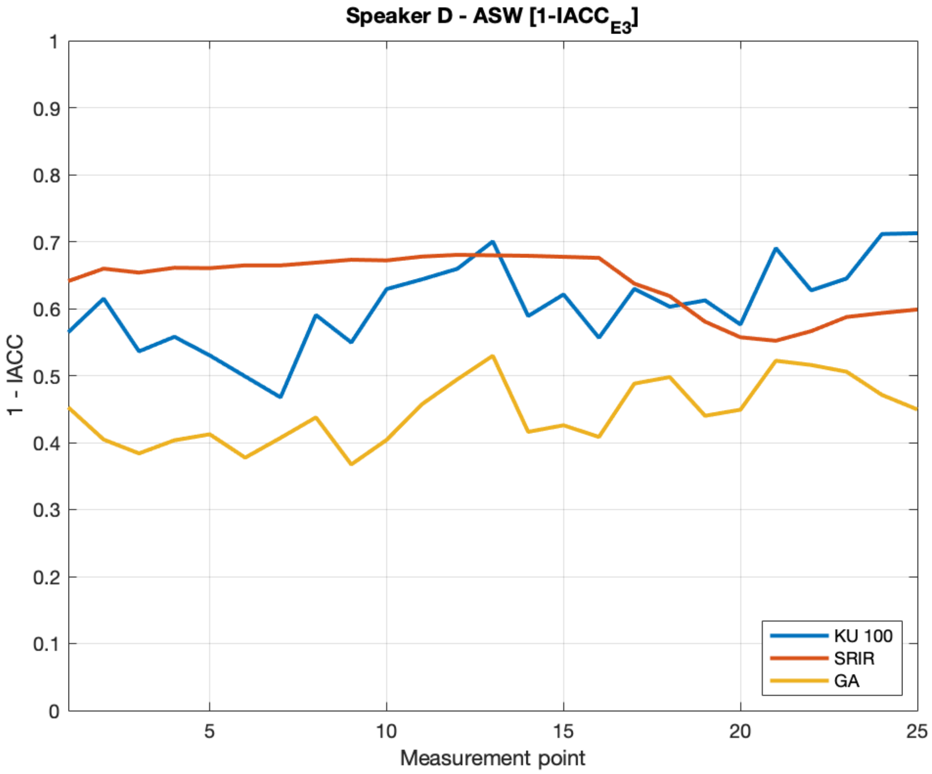


Figure 52: ASW for each measurement point for loudspeaker D

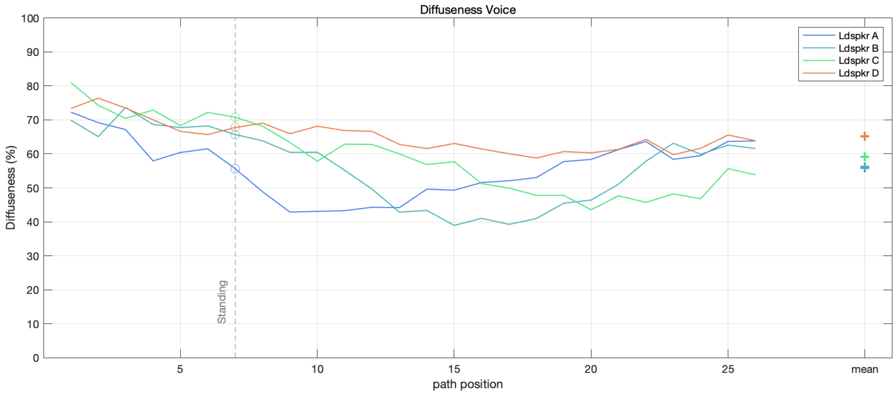


Figure 53: Diffuseness calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The circles indicate values at the standing point. Means along the path are represented on the right.

3.8 Direction of Arrival

The intensity vector indicates the ”theoretical” global direction of arrival, which can be calculated for each position along the path (refer to Figure 54. The intersection of these vectors provides an estimation of image stability while walking.

For all loudspeakers, the intersections are generally close to the target. However, the estimation of direction is significantly inaccurate for the initial points of the path. Specifically, it appears too frontal for loudspeakers A, B, and particularly for C, while it is excessively lateral for loudspeaker D. The thick line represents the direction from the standing point perspective. Importantly, for loudspeakers B and D, these lines intersect loudspeaker C for a portion of the points, aligning with the observed tendency in the localization error responses during the standing phase, where loudspeakers B and D were often misperceived as C.

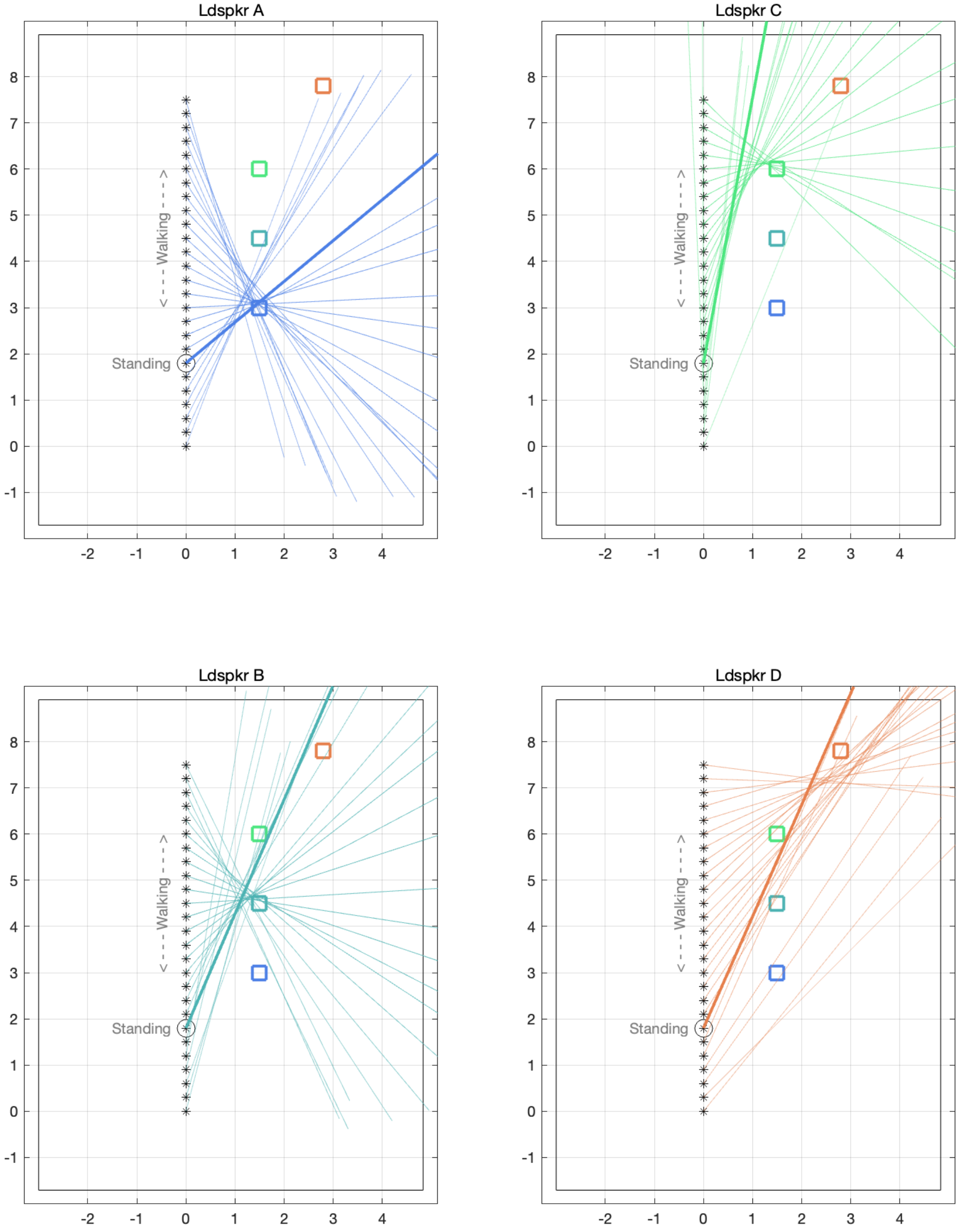


Figure 54: Direction of arrival calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The thick lines indicate the estimation of direction at the standing point.

3.9 Elevation

Figure 55 indicates the elevation of the intensity vector calculated from SRIRs measured for each of the points along the path.

For loudspeakers A and B, the elevation remains close to 0° for points directly in front of the loudspeaker on the path, while it decreases for points farther away, likely due to floor reflections. Loudspeaker D shows fluctuations around -10°, while loudspeaker C reaches a minimum elevation of around -20° in the middle of the path.

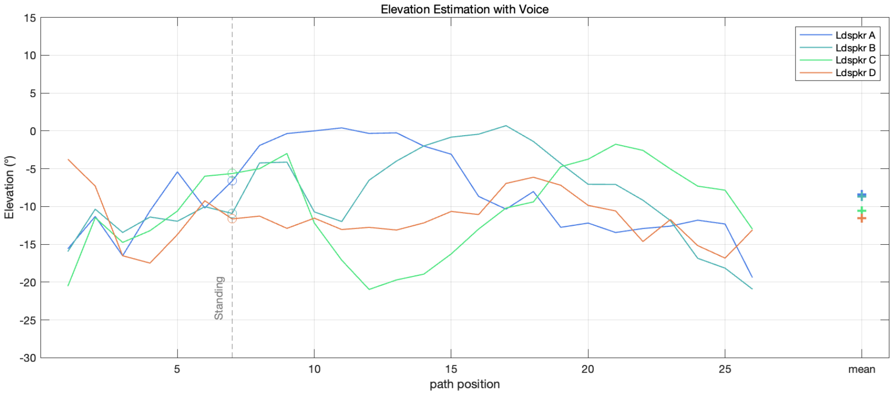


Figure 55: Source elevation calculated from SRIR measured with Eigenmike for each point of the path and loudspeaker. The circles indicate values at the standing point. Means along the path are represented on the right.

4 Correlation Between Perceptual and Acoustic Factors

The objective evaluation of acoustic factors included both auralization methods and reference measurements of KU 100 along the path, aiming to discern the acoustic disparities between them. The primary objective of this section is to establish a correlation between the acoustic characteristics of different auralization methods and the subjective ratings provided by participants, ultimately identifying the key acoustic parameters influencing the assessment of plausibility. However, it has to be noted that investigating the correlation between acoustic factors and perceptual ratings was not the core goal of the experiment. Therefore, this analysis can only provide viable hypotheses and discuss possible areas for future research.

The examination of the correlation between plausibility difference and other attributes in Section 2.5 revealed that blur and localization error were the primary factors explaining the variance in plausibility difference. Hence, it is accurate to consider the evaluation of these attributes as representing the core results of the experiment. That is why, the analysis focuses on exploring the correlation between plots depicting ratings for plausibility, blur, and localization error in both phases, as illustrated in Figure 56, along with the acoustic parameters analyzed in Section 3.

4.1 Energy of Time Segments

The analysis of the energy evolution of time segments along the path 47 revealed that auralizations and measurements of loudspeakers A, B, and C are very similar. There are no significant objective differences in level for different segments of IR between the three loudspeakers. This finding confirms that the difference in perceptual ratings in the standing phase between loudspeakers A-B and C-B (loudspeakers A and C obtained statistically significantly different ratings between auralizations and real playback, in contrast to loudspeaker B which did not have any statistically significant difference between playback methods) comes from the fact that loudspeaker B was

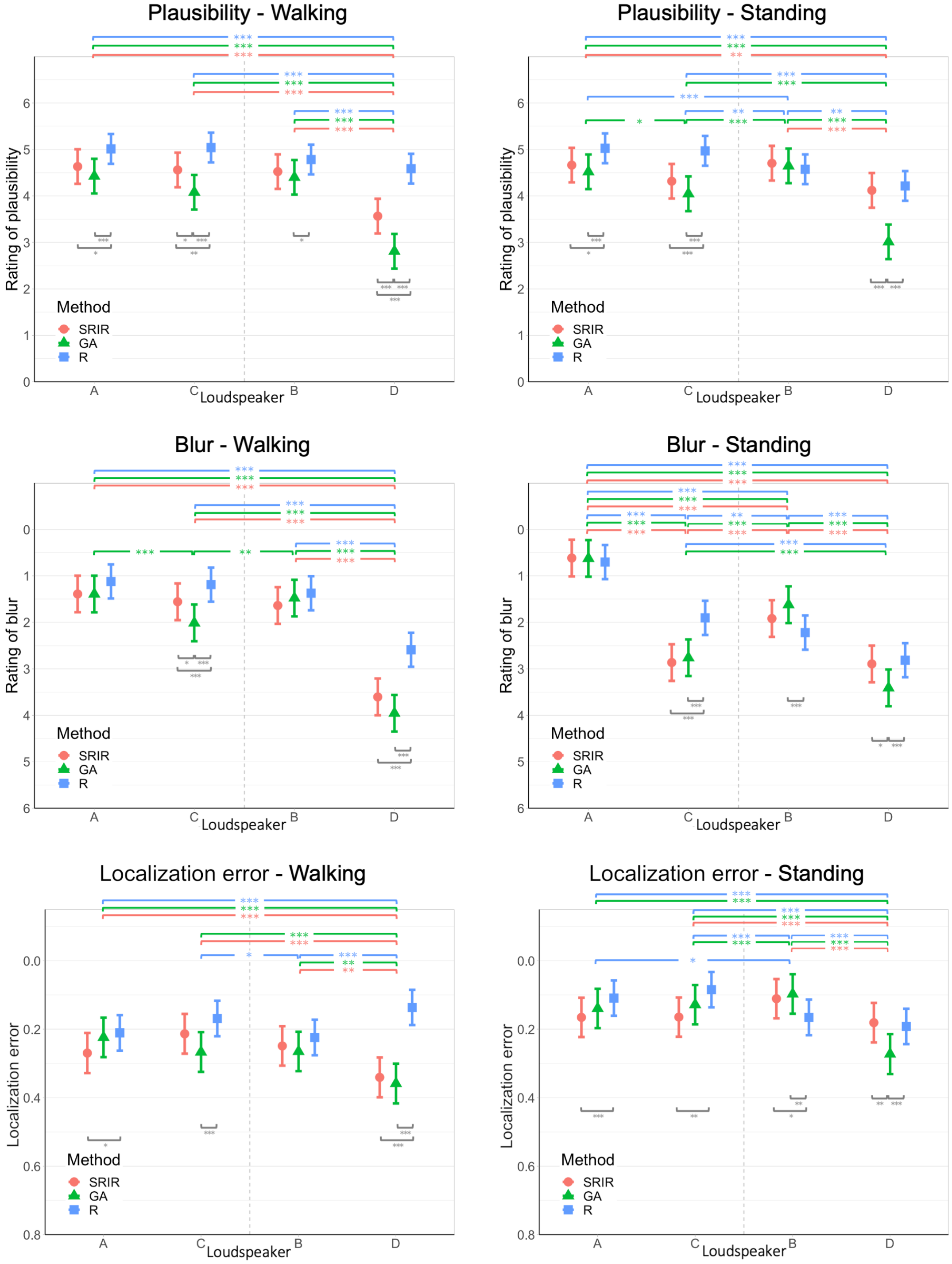


Figure 56: Predicted means and 95% confidence intervals for plausibility, blur, and localization error ratings (***P < 0.05). The y-scale for blur was reversed to facilitate the comparison (0 - very focused, 6 - very blurry).

always presented with loudspeaker D and not from the differences in level of different time segments (refer to Figure 56).

4.2 Plausibility and Blur

The analysis focused on the development of a hypothesis of what acoustic differences between virtual and real loudspeakers could lead to differences between the ratings of plausibility for loudspeaker D in both phases (see Figure 56). In particular, the GA method was evaluated with lower plausibility ratings in both phases, while the SRIR method was as plausible as the real loudspeaker in the standing phase and obtained lower values in the walking phase, although still higher than the GA method.

As the ratings of plausibility and blur for loudspeaker D are very similar, it can be assumed that in this case plausibility ratings were affected by the perceived ”blurriness” of sound. As the first step, the possible hypothesis of the cause for the perception of blur had to be established.

According to the literature (Berg & Rumsey, 2003), blur perception could be associated with:

- Clarity (C50, C80)
- Apparent Source Width (ASW)
- Localizability

The following section discusses each of the hypotheses.

4.2.1 Clarity (C50, C80)

C50 and C80 parameters refer to the clarity of speech and music. These acoustic descriptors compare the sound energy in the early reflection segment with the late reflection segment (after 50 ms for speech clarity and 80 ms for music clarity). Higher values indicate better clarity, as it indicates that the direct sound is more prominent compared to early reflections. This means that speech sounds are more easily distinguished from background noise and reverberation, leading to improved speech intelligibility and understanding for listeners. The hypothesis is that more blurry sources should have a lower clarity value. Indeed, loudspeaker D had a lower clarity value throughout the path compared to the rest of the loudspeakers. However, looking at the difference between playback methods for loudspeaker D, they do not reveal the expected results. Both auralization methods obtained lower blur ratings but the C50 values are higher for auralizations than for measurements. This means that clarity cannot explain the difference in blur evaluation. A similar situation occurs for C80 - both auralization methods have higher C80 values than measurements for loudspeaker D.

4.2.2 Apparent Source Width (ASW)

ASW is a measure of the spatial extent of the auditory event (Berg & Rumsey, 2003). An increase in the perceived width of the source could be associated with an increase in blur. In previous studies, ASW has been shown to be directly correlated with $[1 - IACC_{E3}]$ (Beranek, 1995; Okano et al., 1998). The results of $[1 - IACC_{E3}]$ calculation for different methods of playback and speaker D are shown in Figure 52. The results indicate that ASW cannot explain the blur ratings. GA method has a lower value of $[1 -$

$IACC_{E3}$] than the reference which means that the sound image for the GA method was narrower than for the reference. However, [$1 - IACC_{E_3}$] significantly fluctuates based on the head's orientation relative to the sound source and may also change with distance (Neidhardt et al., 2022). Therefore, matching IACCearly might only be relevant for individuals directly facing the sound source. Furthermore, ASW varies depending on the angle at which reflections occur (Johnson & Lee, 2019). Using the IACC metric may also result in significantly inaccurate predictions when single reflections overwhelmingly dominate the sound field (Blau, 2004).

4.2.3 Localizability

Localizability is a spatial attribute that refers to the ease of localizing the sound (Berg & Rumsey, 2003; Nicol et al., 2014). When the source is difficult to localize, it could be perceived as more blurry. The scatter plot depicting responses to the localization question (Figure 33) reveals differences in the ellipses calculated for each rendering method during the standing phase. The majority of the responses for loudspeaker D were biased towards the listening point, thus underestimating the distance. The difference between the GA and R methods is particularly evident where the ellipse extends more significantly and is directed toward the listening point. This proves that the virtual source for the GA method was perceived as closer to the listener than the real loudspeaker and virtual source rendered with method SRIR. The hypothesis is that underestimation of source distance was the primary cause of lower ratings for plausibility, blur, and localization error for the GA method. Looking at the results of externalization evaluation (Figure 36), it is clear that the GA method for loudspeaker D was also statistically significantly less externalized than the other two methods. This observation supports the stated hypothesis as the externalization perception is closely linked with the perception of distance (Best et al., 2020). Besides that, the small angle between the loudspeakers (refer to Figure 46) made it almost impossible to differentiate the loudspeakers based on the angular position. Listeners had to rely almost solely on the distance perception. Previous literature recognizes distance localization ”blur” as an attribute of auditory distance judgment which has high variability (Zahorik et al., 2005). This variability can be partly reduced when visual anchor is available (Anderson & Zahorik, 2014).

4.3 Distance Perception

The primary cues for distance perception are intensity and direct-to-reverberant ratio (DRR) (Zahorik et al., 2005). The difference in total energy of the signal, indicative of sound intensity, is illustrated in Plot 48. Importantly, at the standing point, the energy of the SRIR method and R method is identical, while the energy of the GA method surpasses that of the real loudspeaker by approximately 2 dB. These values correlate with the results of plausibility and blur ratings in the standing phase which are the same for real playback and SRIR method and lower for the GA method (see Figure 56). The correlation between intensity and perceptual ratings is still working also in the walking phase. Importantly, the energy between method SRIR and KU 100 measurements is very similar only at the point M07. After the standing point, it diverges and obtains values around 2-3 dB above the measurements. This observation correlates with the ratings of SRIR which for the walking phase are rated lower than real reference.

The second distance cue that could explain the underestimation of distance is the direct-to-reverberant ratio (Figure 49). However, the DRR does not explain the difference in ratings as the values for GA and SRIR method for loudspeaker D in standing point are very similar. These observations align with previous research indicating that for speech signals, intensity often carries more significance than DRR in judgments of distance (Zahorik, 2002).

More importantly, other studies showed that DRR provides an absolute cue about distance (Mershon & Bowers, 1979) while intensity needs to be compared relative to other presentations at different distances to be useful (Mershon & King, 1975). These findings provide more insight into the cause of loudspeaker recognition errors during the listening test. As the stimuli were presented in pairs, the intensity values of the stimuli were compared. When the intensity of one of the stimuli was different than expected for a given position it led to the localization error. The intensity between auralizations and real loudspeakers at the standing point was different for all the loudspeakers. This means that for all of the pairs when real sound was presented with virtual it possibly led to the errors of localization. The plot indicating errors of loudspeaker recognition in the standing phase shows that the biggest amount of errors was obtained by pairs SRIR-R, GA-R for loudspeakers B, C, and D confirming this hypothesis (refer to Figure 57). Moreover, the ratings of plausibility were highest for the GA-GA method which also supports this observation (refer to Figure 31).

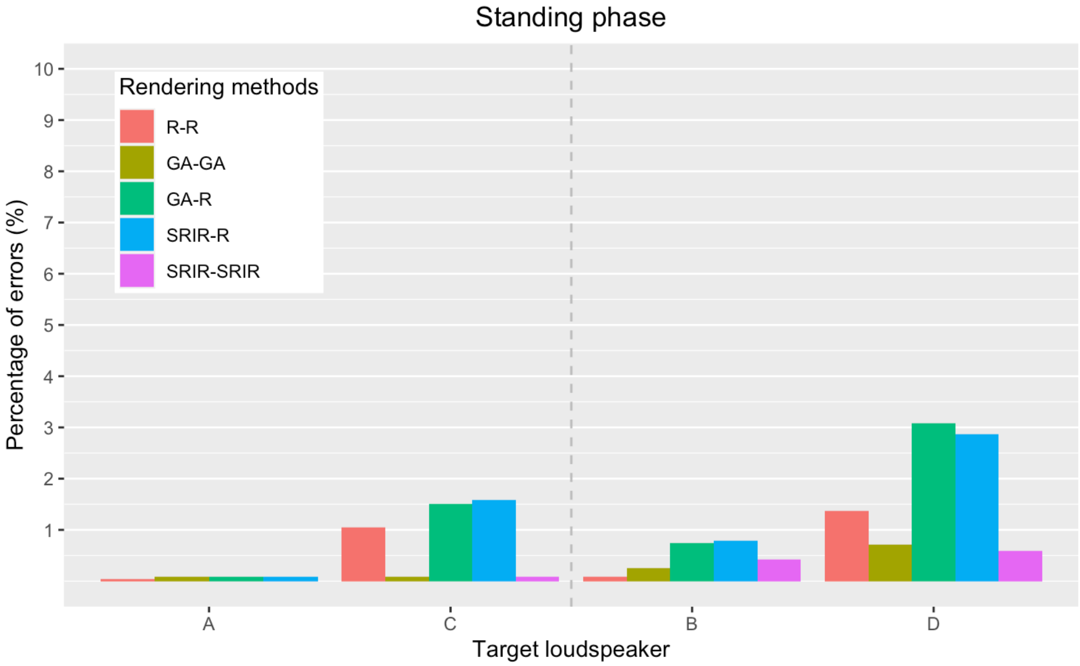


Figure 57: Percentage of loudspeaker recognition errors for different rendering method pairs in standing phase

4.4 Interaction of Self-Motion and Auditory Cues

The experimental design of the study allowed us to compare conditions when the listener was moving or remained stationary. When the listener is moving, more cues are available which give more details about the acoustic characteristics of the room and the sources. However, the analysis of the data revealed that for loudspeakers A, and C there was no statistically significant difference between plausibility evaluation for different phases (see Figure 29 and 30). As discussed before, lower plausibility judgment was closely related to the ambiguity of perceived distance. Contrary to results for loudspeakers A-C - previous studies showed that the movement of the listener improves the accuracy of distance perception in comparison to standing condition (Speigle & Loomis, 1993; Ashmead et al., 1995). However, the condition of the standing phase of the current study presents a specific case where all of the loudspeakers stood at the same line as the listener at different distances. As a result, this design allowed participants to listen to the same source from three different distances. Consequently, these settings allowed to reliably compare the different methods of playback and provide enough cues to be comparable with walking conditions. Contrary to this, loudspeaker D was standing in a different location and had a different orientation than other loudspeakers, thus the number of cues available during the standing phase was more limited than for other loudspeakers which resulted in much larger evaluation differences between walking and standing phases. Another difference between the phases was that in the standing phase, real loudspeaker B was rated lower than loudspeakers A and C. Besides that loudspeaker C rendered with method GA was rated lower than loudspeakers A and B. The cause of these lower ratings will be explained in subsequent sections.

4.5 Source Orientation

The findings underscore the significance of considering the orientation of sound sources, particularly evident with loudspeaker D. Unlike its counterparts positioned closer to the walking path and oriented perpendicular to it, loudspeaker D was situated farther away and aligned parallel to the path (see Figure 46). In the standing phase, the real loudspeaker D was perceived as less plausible compared to real loudspeaker B. It was caused by the fact that the reference provided by loudspeaker B was insufficient for accurately judging loudspeaker D. Listeners became accustomed to the variations in DRR and intensity between loudspeakers A, B, and C, which were positioned similarly in a line, differing mainly in distance. Therefore, the unfamiliar directivity of loudspeaker

D during the standing phase might have created wrong expectations about its sound characteristics, contributing to its lower plausibility ratings. This aligns with previous research demonstrating the impact of source directivity on distance perception (Wendt et al., 2017; Laitinen et al., 2015).

During the walking phase, judgments became more reliable as participants had access to more cues of loudspeaker D. The lower ratings for loudspeaker D could be attributed to incorrect intensity values, possibly caused by the influence of reverberated sound. Analysis of DRR values indicated that during the walking phase, loudspeaker D was heard from a distance where the energy of direct and reverberated sound was similar and close to critical distance, unlike loudspeakers A, B, and C, where direct sound predominated for most of the path (see Figure 49). For loudspeakers A, B, and C the change in acoustic cues was much larger when walking along the path because they were heard from close distances and in frontal direction. This orientation was associated with a fuller spectrum of sound as indicated by the directivity pattern of the loudspeaker and auralizations. When crossing the loudspeaker in the walking phase, the change in spectral content was significant. In contrast, loudspeaker D was never heard from the front. Consequently, its sound spectrum had a lower magnitude of high frequencies. When approaching the loudspeaker along the path, the change in the acoustic cues was smaller as the intensity increase associated with the approaching movement was contradicted by the decrease of the high frequencies influenced by the loudspeaker directivity pattern. In consequence, even the real loudspeaker D was rated as less plausible, as there were fewer acoustic cues leading to a weaker parallax effect.

Additionally, previous research showed that the lower the energy of the direct sound, the smaller differences of early reflections or reverberation are perceived (Buchholz et al., 2001). Consequently, the inaccuracy of early reflections pattern and late reverberation of auralizations likely played a more significant role for loudspeaker D than other loudspeakers leading to lower plausibility scores.

4.6 Visual Cues

Interestingly, during the standing phase, real loudspeaker B received lower ratings for blur and localization error compared to the auralizations. Analysis of the responses to the localization question (Figure 33), along with bar plots of loudspeaker recognition errors (Figure 35), revealed that this discrepancy was often due to participants perceiving real loudspeaker B in the position of loudspeaker C. The ambiguity surrounding the position of loudspeaker B might have stemmed from intensity differences. Real loudspeaker B was consistently presented in pair with loudspeaker D, rendered either virtually or physically. As the total signal energy emitted by real loudspeaker B, positioned closer, was lower than the energy of the auralizations for loudspeaker D, situated farther away, this could have led to an overestimation of distance. Additionally, this overestimation might have been influenced by the visual cue of the standing loudspeaker, contributing to what is known as the ”ventriloquism effect.” Previous studies, such as the one by Zahorik (2003), have demonstrated the significant impact of visual capture on auditory perception over considerable distances. Likewise, research by Mershon and King (1975) observed a similar effect, wherein an occluded sound source positioned closer to listeners than a visible dummy loudspeaker led to an overestimation of the distance to the sound source, perceiving it as being at the farther dummy loudspeaker. However, other studies have shown that sources positioned farther away than the visual target were more likely to be perceived as coincident than those positioned closer, suggesting that the intensity difference effect was strong enough to perceptually relocate the real sound source to the more distant visual target.

Surprisingly, this effect did not impact plausibility ratings. One possible explanation is that the implausibility of loudspeaker D consistently caused loudspeaker B to be perceived as real. Another hypothesis is that changing the perceived position did not affect the ”realism” of the sound.

4.7 Early Reflections

During the walking phase, loudspeaker C, rendered using the GA method, received lower scores for plausibility, blur, and localization error compared to the GA and R methods. Analysis of responses to the localization question (Figure 32) and localization error evaluation (Figure 30) revealed issues with the accuracy of loudspeaker C localization. Since the direct sound for both auralization methods was rendered similarly, the discrepancy must have originated from differences in early reflections or late reverberation, influencing sound localization perception. This effect aligns with the phenomenon of summing localization, suggesting that early reflections arriving within 1–7ms after the direct sound can shift the apparent source position (Neidhardt et al., 2022). Further examination of early reflection patterns between rendering methods is necessary to explore this hypothesis.

In the implementation of the SRIR method, the early reflection pattern remained constant, with only the relative intensity of reverberation modulated based on the listener-source distance (refer to Chapter V). Results indicate that maintaining the spatio-temporal pattern of reflections did not significantly impact plausibility, particularly for loudspeakers A, B, and C (see Figure 30). Furthermore, keeping the same pattern of early reflections resulted in better localizability of virtual sources than simulating the reflections with the GA method which seemed to create more localization ambiguities and higher blur. This finding supports previous research by Neidhardt et al. (2018) and Neidhardt and Kamandi (2022), which found that preserving the spatio-temporal pattern of early reflections did not significantly affect the plausibility of virtual sound sources. However, as observed in the current experiment, in cases where all loudspeakers were directed towards the listener (loudspeakers A, B, C), the direct sound played a more pivotal role in virtual sound source perception, shadowing the inaccuracies of early reflections pattern. Previous studies have shown that for loudspeakers facing away from the listener, the pattern of early reflections becomes more critical, and keeping it constant can lead to lower plausibility ratings (Neidhardt & Zerlik, 2021). Further investigation of the influence of early reflections rendering accuracy is required to understand its significance in sound perception.

4.8 Non-Individualized HRTFs

Blur and localization error were not always correlated with plausibility ratings. For loudspeaker A in the standing phase, the difference between rendering methods for blur and localization error was very small while the difference in plausibility is more significant. As the blur ratings were associated mostly with distance blur, other factors led participants to perceive the implausibility of auralizations. One of the important factors might have been the implementation of non-individualized HRTFs. As stated before (see Section 2.2.1), height disparity between participants' ear level and the center of the loudspeaker significantly influenced subjective ratings. A greater absolute height difference correlated with lower plausibility, blur, externalization scores, higher timbre difference ratings, and increased localization error values. This effect was attributed to the use of generalized HRTFs, which compensated for height differences by placing virtual sources below or above the horizontal plane. Non-individualized HRTFs have been linked to degraded vertical localization accuracy and increased ambiguity along the cone of confusion, impacting the perception of blur, externalization, and localization error, consequently affecting plausibility. However, as the current study revealed the tuning of internal reference affects the perception of sound inaccuracies indicating the importance of context in the assessment of virtual sound. Further research is needed to explore the importance of non-individualized HRTFs in different AAR scenarios. However, it is important to acknowledge that the scope of the objective analysis did not fully cover the potential effects of binaural rendering, such as the conversion between HOA and binaural for simulating room effects which could also impact blur and plausibility perception. Further research is needed to explore additional objective measures of acoustic simulation quality.

5 Limitations and Future Work

The study's primary limitation lies in its focus on evaluating sound perception within a single medium-sized space, which leaves uncertain how space size influences the importance of various acoustic characteristics on sound perception. By confining the acoustic space to one specific room, the study's results were constrained, as room size impacts the relative significance of acoustic simulation accuracy. Specifically, room dimensions greatly affect decay time (Rungta et al., 2016) and the magnitude of differences across room sections. Moreover, the room's small size in the study limited the range of source-listener distances, consistently placing the listener within the critical distance where reverberation equals direct sound level. Consequently, emphasis was placed on the radiation properties of the sound source. Future research should investigate how room size affects the impact of various acoustic attributes on plausibility judgments.

Additionally, the results were restricted to specific loudspeaker positions and directions within the room, with all loudspeakers facing the listener. This setup influenced the importance of direct sound and loudspeaker radiation patterns. The results revealed the ability to shape the inner reference in plausibility judgments depending on the type of sounds presented to the listener. It seems that the less similarity is between the two sounds within one room, the more difficult is the accurate plausibility judgment. This effect needs further exploration with regard to different source positions within the room.

The study focused on two different acoustic rendering methods which differ simultaneously by several features. Thus it prevented detailed conclusions regarding which rendering aspect contributed most to ratings. Further investigation is necessary to determine the relative importance of early reflections versus reverberation simulation. Questions remain about the extent to which simplifying the simulation while maintaining plausibility is feasible. It appears that accurately replicating early reflection patterns may be less crucial than assumed, and the key attributes of reverberation need further studies. Importantly, the study emphasized the critical role of accurately simulating distance to achieve plausible virtual sound sources in augmented acoustic reality (AAR) contexts, considering that distance perception relies on both direct sound and reverberation.

The study employed only male voice stimuli due to the listening test's length. However, the choice of stimuli type may affect the perceptual importance of distance cues (Zahorik, 2002) and in consequence plausibility. Previous research indicated that using noise stimuli led to significantly lower plausibility ratings (Neidhardt et al., 2018), likely causing higher sensitivity to rendering inconsistencies. Thus, further studies should explore alternative stimulus types in AAR scenarios.

CHAPTER VIII

CONCLUSIONS AND FUTURE WORK

1 Summary of Contributions

In Chapter I we formulated research questions that will be addressed by this dissertation. This section will outline answers to each of these questions.

(1) How Does a Subject's Freedom of Movement Affect the Perceptual Evaluation of an AR Sound Scene? While walking, participants could effectively utilize dynamic acoustic cues such as changes in intensity, and spectrum by changing azimuth to the sound source, and DRR. Additionally, the parallax effect was pronounced for loudspeakers A, B, and C along the walking path, enhancing the presence of sound sources. The primary factors influencing plausibility judgments during motion appeared to be the coherence between self-motion and auditory cues, as well as the alignment between visual and auditory cues.

In contrast, during the standing phase, participants were confined to a single static perspective of the sources. Consequently, participants judged mostly the perceptual aspects of the sound focusing on the intensity, spectrum, and DRR, and relied more heavily on comparing the two sources presented together in each trial. This was proved by the results of analysis showing that the lower plausibility judgments for loudspeakers

B and D, even in real playback scenarios were caused by the inaccuracy of intensity cues in the auralizations.

However, the ratings of loudspeakers A and C were very similar between the two phases. This was due to the unique experimental setup, where loudspeakers A, B, and C were aligned in a single line at varying distances. This setup facilitated reliable evaluation of different rendering methods for these loudspeakers as it mimicked a situation where one source could be heard from different distances. Consequently, the positioning of loudspeaker D, which diverged from the others, led to more pronounced differences in evaluation due to the limited cues available during the standing phase.

(2) What Is the Correlation Between Plausibility and Other Perceptual Attributes of Sound? The analysis explored the correlation between plausibility difference and other attribute differences, revealing that each attribute independently contributes to explaining variance in plausibility. Blur ratings accounted for the largest portion of variance, followed by localization error, which was partially correlated with blur, and then externalization, timbre, and reverberation. Blur and localization error ratings were primarily influenced by the ambiguity between perceived sound source location and visual anchors, a crucial factor in plausibility evaluation. Externalization emerged as the next significant factor.

The most important conclusion is that plausibility is a complex subjective percept. As it relates to the comparison of sound to the inner reference, it encompasses different aspects of sound. All evaluated factors played a role in plausibility perception. The highest correlation with plausibility was observed for blur and localization error, highlighting the importance of congruency between visual anchors and sound. Externalization followed as a crucial aspect of binaural sound, prone to disruption and sensitive to room divergence. Surprisingly, reverberation emerged as the least significant attribute, consistent with previous findings indicating listeners' limited sensitivity to reverberation changes.

(3) Do the Properties of Real Reference Affect Plausibility Judgment? Yes, the properties of the real reference do affect plausibility judgment. Plausibility evaluation is based on the agreement between perceived sound event and internal reference shaped by previous sound experiences. The experiment proved that in the context of AAR, the internal reference can be tuned by the real sound environment around the listener. The results of the experiment showed that manipulation of the real reference properties affected plausibility perception. By positioning two pairs of loudspeakers differently, the study enabled a comparison of plausibility perception when stimuli were presented with similar or different real counterparts. The results revealed that when real loudspeakers were positioned similarly, participants found it easier to distinguish between auralizations compared to when the real counterparts were positioned differently. This suggests that the properties of the real reference, such as their positioning, can impact the perceived plausibility of virtual sounds. Therefore, adjusting the properties of the real reference, such as altering the loudspeaker's position within the pair, can be valuable in adapting the test to different real-life scenarios and objectives, where rendering accuracy may vary. The results indicated also that the inclusion of real references significantly affected the perception of virtual sources. From all of the pairs of rendering methods presented in one trial, pair of two auralizations GA received the highest scores of plausibility. This indicates that without comparison to the real source, virtual source can be perceived more plausible.

(4) How Do Objective Measures of Acoustical Parameters Correspond to Subjective Evaluation of Acoustic Processing? The analysis showed that ratings of blur were mainly correlated with localizability. The imperfect localizability was caused by the errors of distance estimation for virtual and sometimes real sources. The main acoustic cue for distance estimation was sound intensity. As the intensity values were compared between loudspeakers, the discrepancy between real and virtual loudspeakers compared in pairs led to localization errors and sometimes problems with loudspeaker recognition.

However, it has to be noted that the scope of the objective analysis was not able to encompass possible effects of the binaural rendering (in particular conversion between HOA and binaural for rendering room effects) as well as non-individual HRTFs which may also contribute to the blur and plausibility perception. There is a need for further research on other objective measures of acoustic simulation quality.

(5) How Does the Position of the Source in the Room and Orientation Influence the Assessment of the Auralizations? The positioning and orientation of the sound sources had varying impacts on the assessment. The position primarily influenced the distance between the loudspeaker and the listener, consequently affecting the DRR and sound intensity. Meanwhile, the orientation altered the angle between the loudspeaker and the listener, thus modifying the sound spectrum due to the directivity pattern.

Loudspeakers A, B, and C were placed close to the walking path and were positioned perpendicular to it. This setup provided participants with a strong parallax effect, enhancing the plausibility of virtual loudspeakers. Additionally, it minimized the significance of early reflections and reverberations, as they were overshadowed by direct sound.

In contrast, loudspeaker D was situated farther from the path and oriented parallel to it. Here, early reflections and reverberations played a more substantial role in sound perception, influenced not only by distance but also by orientation, resulting in the decrease of high frequencies. Consequently, this led to a diminished parallax effect.

(6) Is the Proposed Methodology an Effective Method for Evaluating Plausibility in 6DOF AR Environments? The proposed methodology appears to be an effective method for evaluating plausibility in 6DoF AR environments. The experimental design focused on replicating real-life scenarios of auditory augmented reality (AAR), where real sounds coexist with virtual ones, making direct comparisons challenging. By utilizing pairs of real and virtual loudspeakers with identical acoustical characteristics but varying positions, the study successfully imitated real-world AAR conditions.

The results highlight the subjectivity of sound plausibility, emphasizing its dependence on the specific context in which the sound is presented. The ability to tune the internal reference by including real counterparts in the evaluation task proved crucial in understanding plausibility perception. Moreover, the methodology enabled subtle differences in plausibility to be detected through the use of categorical scales, revealing a nuanced ”grey area” in plausibility perception that may be overlooked by binary evaluation paradigms.

Careful consideration of participant movement within the 6DoF space is crucial, and the methodology addressed this by implementing a predetermined walking path. This controlled movement ensured consistent evaluation conditions, allowing participants to experience similar hearing conditions. Furthermore, the study effectively evaluated the chosen auralization methods. Both methods achieved the intended goals, eliciting subtle but significant differences in perception between real and virtual sounds, despite simplifications of the implementation.

Overall, the proposed methodology offered a comprehensive and adaptable approach to evaluating plausibility in 6DoF AR environments, providing valuable insights into the plausibility perception and its implications for AAR systems.

(7) How Do the Participants’ Speed of Walking and Amplitude of Yaw Movement Affect the Evaluation? The analysis of tracking data across both phases uncovered a consistent pattern in participants’ behavior. During the walking phase, participants exhibited faster movement when exposed to real playback, correlating with higher scores for sound attributes such as plausibility, blur, localization error, and loudspeaker recognition rate. Similarly, in the standing phase, reduced yaw movement amplitude during real playback corresponded to elevated ratings for the same attributes. Participants walked faster and minimized head movements when encountering easier tasks with real playback, while they adopted slower speeds and increased head movements during auralization playback, indicative of a more challenging task.

These findings suggest a meaningful correlation between stimulus presentation and participant behavior, implying that analyzing listener behavior could provide insights into the perceived nature of the stimulus. However, further research in this area is needed to broaden the understanding of these relationships.

2 Implications for Sound Design in AR

This section will briefly discuss the implications of the conclusions drawn from the dissertation study on sound design practices in AR.

The findings of the study underscored the significance of aligning visual and acoustic cues for an immersive AR experience. Notably, the accurate portrayal of distance emerged as a critical factor in determining plausibility judgments. While distance perception primarily relies on intensity cues, the role of parameters such as DRR may vary depending on the stimulus type. Consequently, detailed attention to both direct sound and reverberant energy is essential for convincingly rendering virtual sound sources in AR environments.

Furthermore, the position and orientation of sound sources can influence the relative importance of direct and reverberated sound. Close-range sources need prioritizing the rendering of direct sound, whereas distant sources require careful consideration of early reflections and late reverberation. Additionally, when directional sound sources face away, the pattern of early reflections becomes particularly crucial, posing a challenge for sound design.

The context of the AR scene also needs consideration. If the scenario anticipates significant movement of the user, accurately simulating the parallax effect becomes pivotal for plausibility. This means prioritizing the ”near field” effects, including accurate reproduction of directivity patterns and faithful rendering of near field acoustic effects (e.g. near field binaural cues). Conversely, in more static scenarios, the sound perception is based more on a comparison of audio cues between real and virtual sources within a room. Consequently, the consistency between them gains importance.

Given the unpredictable nature of the user’s environment in AR experiences, the sound design must accommodate scenarios with varying numbers of real sound sources. The likelihood of users hearing real sounds similar to virtual ones affects the detectability of discrepancies in the reproduced sounds. The closer the similarity between virtual and real sounds, the easier it becomes for users to perceive inconsistencies.

Lastly, the perception of plausibility depends on multiple attributes. Besides localization errors, blur, and reverberance which were already discussed, aspects like externalization and accuracy of the directivity pattern also contribute significantly to sound scene plausibility. Ensuring fidelity across these attributes should enhance the overall immersive quality of the AR environment.

3 Future Directions

There are several interesting new research directions inspired by this study.

3.1 Varying the Signal

The experimental design of the study primarily focused on comparing two sound sources. The objective was to introduce differences in loudspeaker positioning within the room to simulate real-world AR scenarios where the virtual sound is always presented with real sounds from the environment. We found a significant influence of the position difference within the pair of sources on perceptual ratings. However, there are alternative approaches to replicate AR environments. Rather than altering position, exploring how adjustments to the signal itself influence evaluations could be valuable. In such cases, the source position would remain constant, but alterations could be made to the type of stimuli (such as different voice timbres or variations in samples). This approach could provide insights into how different types of variations in real counterparts impact plausibility judgments.

3.2 Exploring Simplifications of Auralization Methods

There is a limited number of research studies that validate the significance of various simulation parameters in 6DoF environments. Currently, it remains unclear how factors such as the order of reflections (particularly for mirrored sources), the number of FDN channels, and the order of Ambisonics of SRIR influence plausibility judgments. Additionally, it is important to investigate how much the auralizations can be simplified without compromising sound plausibility. This includes exploring alternatives such as substituting late reverberation from SRIR with omnidirectional IR, employing lower Ambisonics order for early reflections, reducing the order of reflections and FDN channels in GA methods, and decreasing HOA order for beamforming utilized in directivity modeling.

3.3 Stimuli Types

The experiment solely concentrated on speech stimuli. The choice of the speech stimuli type was motivated by its familiarity, also in spectral and spatial behavior. Yet it is crucial to extend validation to other stimulus types using the same methodology. It is expected that noise stimuli might diminish the plausibility of auralizations. It may stimulate larger spectral bandwidth and in consequence reveal limitations of the auralization methods, especially for radiation pattern, room acoustic modeling, etc.) Conversely, the impact of music stimuli on the plausibility of sound sources using identical rendering methods could vary, potentially either decreasing or increasing plausibility depending on the specific characteristics of the music.

3.4 Participants Selection

This study was conducted with audio experts and involved a task centered on a direct focus on sound qualities. Nevertheless, it might be beneficial to validate how the plausibility of virtual sources would be assessed with a group of naive subjects who do not have expertise in audio, and without specifically focusing on listening which closer resembles the real applications of audio AR.

3.5 Behavioral Measures in Plausibility Evaluation

The test results revealed that the choice of rendering method could be inferred from analyzing subjects’ walking speed and head movements. This indicates a significant potential for measuring plausibility through implicit means. Consequently, the duration of listening tests could be substantially reduced, enabling the inclusion of more conditions.

3.6 Sound Source Orientation

Research on the evaluation of directional sound sources facing away from the listener in 6DoF is limited. However, existing knowledge suggests that in such scenarios, the pattern of early reflections significantly influences the perception of plausibility. This study demonstrated that even a slight orientation change, where the sound source is not directly in front of the listener, can reduce plausibility. Further studies should explore various sound source orientations and positions within the room to deepen our understanding of this effect.

3.7 Non-Individualized HRTFs in 6DoF

The majority of research on the perceptual influence of non-individualized HRTFs has been conducted in static or 3DoF conditions. This study suggested that differences in height compensated by non-individualized HRTFs influenced plausibility judgments. Further investigation into this effect is necessary, including varying the elevation of the sound sources.

4 Conclusions

In this dissertation, we presented a study focused on plausibility evaluation in the AR context. We proposed a novel methodology for the study focused on resembling real-life AR environment scenarios. The study's comprehensive analysis provides a nuanced understanding of the interactions between participant movement and sound perception in 6DoF AR environments. We analyzed the influence of loudspeaker position and rendering method on plausibility judgment. The study outcomes revealed the correlation of plausibility with other sound attributes and indicated the core acoustic parameters that were associated with subjective assessment. We found and discussed the influence of the real counterpart on the plausibility judgment. The insights gained contribute to the refinement of experimental methodologies and deepen our understanding of the plausibility perception in AR context.

REFERENCES

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, *65*(4), 943–950. 19.

Andéol, G., Savel, S., & Guillaume, A. (2014). Perceptual factors contribute more than acoustic factors to sound localization abilities with virtual sources. *Frontiers in Neuroscience*, *8*(DEC), 1–17. 25.

Anderson, P. W., & Zahorik, P. (2014). Auditory/visual distance estimation: Accuracy and variability. *Frontiers in Psychology*, *5*(SEP), 1–11. 224.

Ashmead, D. H., Davis, D. F. L., & Northington, A. (1995). Contribution of Listeners' Approaching Motion to Auditory Distance Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 239–256. 226.

Astolfi, A., Corrado, V., & Griginis, A. (2008). Comparison between measured and calculated parameters for the acoustical characterization of small classrooms. *Applied Acoustics*, *69*(11). 92.

Avid. (2019). *Pro Tools - Music Software*. `https://www.avid.com/pro-tools`. (Accessed on 08/01/2019) 70.

Bailey, W., & Fazenda, B. M. (2018). The effect of visual cues and binaural rendering method on plausibility in virtual environments. *Proceedings of the 144th AES Convention*. 28 and 33.

Barron, M., & Lee, L. J. (1988). Energy relations in concert auditoriums. I. *Journal of the Acoustical Society of America*, *84*(2), 618–628. 18 and 95.

Barron, M., & Marshall, A. H. (1981). Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, *77*(2), 211–232. 23.

Bech, S. (1992). Selection and training of subjects for listening tests on sound-reproducing equipment. *J. Audio Eng. Soc*, *40*(7/8), 590–610. 39.

Bech, S. (1995). Timbral aspects of reproduced sound in small rooms. II. *Journal of the Acoustical Society of America*, *97*(3), 1717–1726. 22.

Begault, D. R. (1999). Auditory and Non-Auditory Factors that Potentially Influence Virtual Acoustic Imagery. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction* (pp. 13–26). 28.

Begault, D. R., & Trejo, L. J. (2000). 3-d sound for virtual reality and multimedia. 63.

Begault, D. R., Wenzel, E. M., & Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society. Audio Engineering Society*, *49*(10), 904–916. 23, 25, and 136.

Beranek, L. L. (1995). Comparison between subjective judgments of concert halls' quality and objective measurements of acoustical attributes. *Acoustical Physics*, *41*(5), 620–629. 214 and 222.

Berg, J. (2002). Systematic evaluation of perceived spatial quality in surround sound systems (Doctoral dissertation, Luleå tekniska universitet). 41.

Berg, J., & Rumsey, F. (2000, Feb). In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors. In *Audio Engineering Society Convention 108*. 40.

Berg, J., & Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. In *AES International Conference: Multichannel Audio, The New Reality* (pp. 184–198). Banff, Canada. 221, 222, and 223.

Berg, J., & Rumsey, F. (2006). Identification of quality attributes of spatial audio by repertory grid technique. *J. Audio Eng. Soc*, *54*(5), 365–379. 44.

Berger, C. C., Gonzalez-Franco, M., Tajadura-Jiménez, A., Florencio, D., & Zhang, Z. (2018). Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Frontiers in Neuroscience*, *12*(FEB). 26.

Bergstrom, I., Azevedo, S., Papiotis, P., Saldanha, N., & Slater, M. (2017). The Plausibility of a String Quartet Performance in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, *23*(4), 1332–1339. 33.

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., & Kopčo, N. (2020). Sound externalization: A review of recent research. *Trends in Hearing*, *24*, 2331216520948390. 137, 189, 204, and 224.

Blau, M. (2004). Correlation of apparent source width with objective measures in synthetic sound fields. *Acta Acustica united with Acustica*, *90*(4), 720–730. 223.

Blauert, J., & Butler, R. A. (1985). Spatial Hearing: The Psychophysics of Human Sound Localization by Jens Blauert. *The Journal of the Acoustical Society of America*. 25 and 201.

Boone, M. M., Verheijen, E. N. G., & van Tol, P. F. (1995). Spatial sound-field reproduction by wave-field synthesis. *J. Audio Eng. Soc*, *43*(12), 1003–1012. 65.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. 109.

Brinkman, W.-P., Hoekstra, A. R. D., & van Egmond, R. (2015). The effect of 3D audio and other audio techniques on virtual reality experience. *Annual Review of Cybertherapy and Telemedicine*, 44–48. 58 and 62.

Brinkmann, F., Lindau, A., & Weinzierl, S. (2017). On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, *142*(4), 1784–1795. 102.

Brungart, D. (1998). Control of perceived distance in virtual audio displays. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, *3*(3), 1101–1104. 29.

Brungart, D. S., Simpson, B. D., & Kordik, A. J. (2005). The detectability of headtracker latency in virtual audio displays. *Proceedings of the 11th International Conference on Auditory Display (ICAD2005)*, 37–42. 27.

Brungart, D. S., Simpson, B. D., Mckinley, R. L., Kordik, A. J., Dallman, R. C., & Ovenshire, D. a. (2004). The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. In *Proceedings of ICAD 04- Tenth Meeting of the International Conference on Auditory Display*. Sydney, Australia. 26.

Buchholz, J. M., Mourjopoulos, J., & Blauert, J. (2001). Room masking: understanding and modelling the masking of room reflections. In *Audio Engineering Society Convention 110*. Amsterdam, The Netherlands. 229.

Carpentier, T. (2021). Spat : a comprehensive toolbox for sound spatialization in Max. *Ideas Sonicas*, *13*(24). 92.

Carpentier, T., Bahu, H., Noisternig, M., & Warusfel, O. (2014, Sept.). Measurement of a head-related transfer function database with high spatial resolution. In *7th Forum Acusticum (EAA)*. Poland. 96.

Carpentier, T., & Einbond, A. (2022, Apr). Spherical correlation as a similarity measure for 3D radiation patterns of musical instruments. In *16ème Congrès Français d'Acoustique*. Marseille (FR). 89.

Choisel, S., & Wickelmaier, F. (2006). Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *J. Audio Eng. Soc*, *54*(9), 815–826. 44.

Churchill, E. F., & Snowdon, D. (1998). Collaborative virtual environments: an introductory review of issues and systems. *Virtual Reality*, *3*(1), 3–15. 59.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409–412. 28.

Colomes, C., Le Bagousse, S., & Paquier, M. (2010, Nov). Families of sound attributes for assessment of spatial audio. In *Audio Engineering Society Convention 129*. 39.

De Sena, E., Haciihabibo⊠lu, H., Cvetković, Z., & Smith, J. O. (2015). Efficient Synthesis of Room Acoustics via Scattering Delay Networks. *IEEE Transactions on Audio, Speech and Language Processing*, *23*(9), 1478–1492. 22.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology*, *6*, 26. 59.

Draper, J. V., Kaber, D. B., & Usher, J. M. (1998). Telepresence. *Human Factors*, *40*(3), 354. 31.

Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., & Wenzel, E. M. (1992a). On the Externalization of Auditory Images. *Presence: Teleoperators and Virtual Environments*, *1*(2), 251–257. 189.

Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., & Wenzel, E. M. (1992b, May). On the externalization of auditory images. *Presence: Teleoper. Virtual Environ.*, *1*(2), 251–257. 51.

EBU 3286–E. (1997). *Assessment methods for the subjective evaluation of the quality of sound programme material – Music*. EBU Tech. 3286–E. 51.

EBU 562-3. (1990). *Subjective Assessment of Sound Quality*. 42.

Elorza, D. O. (2005). Room acoustics modeling using the ray- tracing method : implementation and evaluation (Licentiate thesis, University of Turku). , 110. 19.

Enge, K., Frank, M., & Höldrich, R. (2020). Listening experiment on the plausibility of acoustic modeling in virtual reality. In *Fortschritte der Akustik - DAGA* (pp. 13–16). Hannover, Germany. 200 and 201.

Engel, I., Henry, C., Garí, S. V. A., Robinson, P. W., Poirier-quinot, D., & Picinali, L. (2019). Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening. *EAA Spatial Audio Signal Processing Symposium*, 121–126. 20, 24, and 33.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108* (pp. 199–202). 85 and 86.

Fediuk, R., Amran, M., Vatin, N., Vasilev, Y., Lesovik, V., & Ozbakkaloglu, T. (2021). Acoustic properties of innovative concretes: A review. *Materials*, *14*(2). 92.

Felnhofer, A., Kothgassner, O. D., Schmidt, M., Heinzle, A. K., Beutl, L., Hlavacs, H., & Kryspin-Exner, I. (2015). Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human Computer Studies*, *82*, 48–56. 56.

Fenton, S., & Wakefield, J. (2012, Apr). Objective profiling of perceived punch and clarity in produced music. In *Audio Engineering Society Convention 132*. 51.

Furness, R. K. (1990, May). Ambisonics-an overview. In *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*. 43.

Geluso, P. (2012, Apr). Capturing height: The addition of z microphones to stereo and surround microphone arrays. In *Audio Engineering Society Convention 132*. 46.

Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, *21*(1), 2–10. 67.

Gibbs, T. (2007). *The Fundamentals of Sonic Art and Sound Design*. Ava Publishing. 2.

Gochfeld, D., Brenner, C., Layng, K., Herscher, S., DeFanti, C., Olko, M., ... Perlin, K. (2018). Holojam in wonderland: Immersive mixed reality theater. *Leonardo*, *51*(4), 362-367. 57 and 60.

Google. (2018). *Resonance Audio Unity SDK API Reference*. `https://resonance-audio.github.io/resonance-audio/`. (Accessed on 03/24/2024) 16.

Greenblatt, A., Abel, J., & Berners, D. (2010). A Hybrid Reverberation Crossfading Technique. In *ICASSP*. 93.

Guastavino, C., & Katz, B. F. (2004). Perceptual evaluation of multi-dimensional spatial audio reproduction. *The Journal of the Acoustical Society of America*, *116*(2), 1105–1115. 40 and 51.

Guezenoc, C., & Séguier, R. (2018). HRTF Individualization: A Survey. *AES 145th Convention*, 1–10. 25 and 63.

Gupta, R., Ranjan, R., He, J., & Gan, W.-S. (2018). On the use of closed-back headphones for active hear-through equalization in augmented reality applications. In *International Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA. 61 and 64.

Gutierrez-Parera, P., Lopez, J. J., & Aguilera, E. (2015). On the influence of headphone quality in the spatial immersion produced by binaural recordings. In *Audio Engineering Society Convention 138*. 64.

Hendrickx, E., Stitt, P., Messonnier, J. C., Lyzwa, J. M., Katz, B. F., & de Boishéraud, C. (2017). Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *The Journal of the Acoustical Society of America*, *141*(3), 2011. 201.

Herscher, S., DeFanti, C., Vitovitch, N. G., Brenner, C., Xia, H., Layng, K., & Perlin, K. (2019, October). Cavrn: An exploration and evaluation of a collective audience virtual reality nexus. *UIST 2019: 32nd ACM User Interface Software and Technology Symposium*. 59, 68, and 69.

Honda, A., Shibata, H., Hidaka, S., Gyoba, J., Iwaya, Y., & Suzuki, Y. (2013). Effects of head movement and proprioceptive feedback in training of sound localization. *i-Perception*, *4*(4), 253–264. 197.

Horowitz, S., & Looney, S. R. (2014). *The essential guide to game audio: the theory and practice of sound for games*. Routledge. 61 and 62.

Horsburgh, A. J., McAlpine, K. B., & Clark, D. F. (2011, Feb). A perspective on the adoption of ambisonics for games. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. 43.

IEC 60268. (1997). *Sound System Equipment–– Part 13: Listening Tests on Loudspeakers*. 42.

ITU-R BS.1116-1. (1997). *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. 39.

ITU-R BS.1534-1. (2003). *Method for the subjective assessment of intermediate quality level of coding systems*. 39.

Johnson, D., & Lee, H. (2019). Perceptual threshold of apparent source width in relation to the azimuth of a single reflection. *The Journal of the Acoustical Society of America*, *145*(4), EL272–EL276. 223.

Jot, J.-M., Audfray, R., Hertensteiner, M., & Schmidt, B. (2021). Rendering spatial sound for interoperable experiences in the audio metaverse. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (p. 1-15). 95.

Jot, J.-M., & Chaigne, A. (1991). Digital delay networks for designing artificial reverberators. In *Audio Engineering Society Convention 90*. 22.

Kamekawa, T., & Marui, A. (2010, Oct). Developing common attributes to evaluate spatial impression of surround sound recording. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. 41 and 51.

Kaplanis, N., Bech, S., Jensen, S. H., & Van Waterschoot, T. (2014). Perception of reverberation in small rooms: A literature study. In *Proceedings of the AES International Conference*. Helsinki, Finland. 23.

Karjalainen, M., Huopaniemi, J., & Välimäki, V. (1995). Direction-Dependent Physical Modeling of Musical Instruments. In *Int. Congr. on Acoustics (ICA\'95)* (Vol. 3, pp. 561–563). 15.

Kato, M., Uematsu, H., Kashino, M., & Hirahara, T. (2003). The effect of head motion on the accuracy of sound localization. *Acoustical Science and Technology*, *24*(5), 315–317. 197.

Kelly, G. (1991). *The Psychology of Personal Constructs*. Routledge. 39.

Klein, F., Gari, S. V., Arend, J. M., & Robinson, P. W. (2021). Towards determining thresholds for room divergence: A pilot study on detection thresholds. *2021 Immersive and 3D Audio: From Architecture to Automotive, I3DA 2021*, 1–7. 204.

Kobayashi, M., & Ueno, K. (2015). The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study Abstract. *Presence, 24*(2), 163–174. 58.

Kronlachner, M., & Zotter, F. (2014). Spatial transformations for the enhancement of Ambisonic recordings. *2nd International Conference on Spatial Audio*(2), 1–5. 90.

Laitinen, M.-V., Politis, A., Huhtakallio, I., & Pulkki, V. (2015). Controlling the perceived distance of an auditory object by manipulation of loudspeaker directivity. *The Journal of the Acoustical Society of America, 137*(6), EL462–EL468. 228.

Larsson, P., Västfjäll, D., Kleiner, M., Vastfjall, D., & Kleiner, M. (2001). Ecological acoustics and the multi-modal perception of rooms: Real and unreal experiences of auditory-visual virtual environments. *International Conference on Auditory Display*(May 2014), 245–249. 28.

Layng, K., Perlin, K., Herscher, S., Brenner, C., & Meduri, T. (2019). Cave: Making collective virtual narrative. *Leonardo, 52*(4), 349-356. 57 and 68.

Leclère, T., Lavandier, M., & Perrin, F. (2019). On the externalization of sound sources with headphones without reference to a real source. *The Journal of the Acoustical Society of America, 146*(4), 2309–2320. 137.

Lee, K. S., Abel, J. S., Välimäki, V., Stilson, T., & Berners, D. P. (2012). The switched convolution reverberator. *AES: Journal of the Audio Engineering Society, 60*(4), 227–236. 22.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkmann, F., & Weinzierl, S. (2014). A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica, 100*(5), 984–994. 82, 102, and 103.

Lindau, A., Hohn, T., & Weinzierl, S. (2007). Binaural resynthesis for comparative studies of acoustical environments. *Audio Engineering Society - 122nd Audio Engineering Society Convention 2007, 3*, 1394–1403. 200.

Lindau, A., & Weinzierl, S. (2012). Assesing the Plausibility of Virtual Acoustic Environments. *Acta Acustica united with Acustica, 98*(5), 804–810. 3, 31, 32, 199, and 200.

Lorho, G. (2005, Oct). Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction. In *Audio Engineering Society Convention 119*. 41.

Martin, A., Jin, C., & Member, A. E. S. (2009). Psychoacoustic Evaluation of Systems for Delivering Spatialized Augmented-Reality Audio *. , *57*(12). 64.

Martin, V., Viaud-Delmon, I., & Warusfel, O. (2022). Effect of environment-related cues on auditory distance perception in the context of audio-only augmented reality. *Applied*

Sciences (Switzerland), *12*(1). 28.

Mason, R., & Rumsey, F. (2000, Feb). An assessment of the spatial performance of virtual home theatre algorithms by subjective and objective methods. In *Audio Engineering Society Convention 108*. 41.

Massé, P., Carpentier, T., Warusfel, O., & Noisternig, M. (2020). Denoising directional room impulse responses with spatially anisotropic late reverberation tails. *Applied Sciences (Switzerland)*, *10*(3). 86 and 93.

Mendonça, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P., & Santos, J. A. (2012). On the improvement of localization accuracy with non-individualized HRTF-based sounds. *AES: Journal of the Audio Engineering Society*, *60*(10), 821–830. 25 and 194.

Merimaa, J. (2010). Modification of HRTF filters to reduce timbral effects in binaural synthesis, part 2: Individual HRTFs. *129th Audio Engineering Society Convention 2010*, *2*, 1330–1342. 194.

Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, *8*(3), 311–322. 225.

Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, *18*(6), 409–415. 225 and 229.

Miles, R. N. (1984). Sound field in a rectangular enclosure with diffusely reflecting boundaries. *Journal of Sound and Vibration*, *92*(2), 203–226. 20.

Monson, B. B., Hunter, E. J., & Story, B. H. (2012). Horizontal directivity of low- and high-frequency energy in speech and singing. *The Journal of the Acoustical Society of America*, *132*(1), 433–441. 15.

Mueller-Tomfelde, C. (2002). Hybrid Sound Reproduction in Audio-Augmented Reality. *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 1–6. 65.

Multiplayer Game Development Made Easy | Photon Engine. (2019). `https://www.photonengine.com/`. (Accessed on 08/01/2019) 69.

Neidhardt, A., & Kamandi, S. (2022). Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room. In *Proc. 152nd Conf. Audio Eng. Soc.* 200, 201, and 231.

Neidhardt, A., & Knoop, N. (2017). Binaural walk-through scenarios with actual self-walking using an HTC Vive. In *Proceedings of the DAGA 2017* (pp. 283–286). 20, 34, 102, 200, and 201.

Neidhardt, A., Schneiderwind, C., & Klein, F. (2022). Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical

Framework. *Trends in Hearing*, *26*. 223 and 230.

Neidhardt, A., Tommy, A. I., & Pereppadan, A. D. (2018). Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. In *Proceedings of the 144th AES Convention*. Milan, Italy. 20, 23, 28, 35, 102, 200, 231, and 234.

Neidhardt, A., & Zerlik, A. M. (2021). The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR. *Frontiers in Virtual Reality*, *2*(September), 1–17. 132, 196, 199, 200, and 231.

Neumeyer, D. (2009). Diegetic/nondiegetic: A theoretical model. *Music and the Moving Image*, *2*(1), 26–39. 68.

Nicol, R., Gros, L., Colomes, C., Warusfel, O., Noisternig, M., Bahu, H., ... Simon, L. S. R. (2014). A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering. *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, 100–106. 223.

Nosal, E.-M., Hodgson, M., & Ashdown, I. (2004). Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms. *The Journal of the Acoustical Society of America*, *116*(2), 970–980. 20.

Nowak, J., & Klockgether, S. (2017). Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations. *The Journal of the Acoustical Society of America*, *142*(3), 1634–1645. 20 and 94.

Okano, T., Beranek, L. L., & Hidaka, T. (1998). Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *The Journal of the Acoustical Society of America*, *104*(1), 255–265. 214 and 222.

Olive, S. E., & Toole, F. E. (1989). Detection of reflections in typical rooms. *AES: Journal of the Audio Engineering Society*, *37*(7-8), 539–553. 22.

Otondo, F., & Rindel, J. H. (2005). A new method for the radiation representation of musical instruments in auralizations. *Acta Acustica united with Acustica*, *91*(5), 902–906. 15.

Paterson, J., & Kadel, O. (2019). Immersive Audio Post-production for 360° Video: Workflow Case Studies. In *2019 AES International Conference on Immersive and Interactive Audio*. York, UK. 61.

Perrett, S., & Noble, W. (1997). The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics*, *59*(7), 1018–1026. 197.

Picinali, L., Wallin, A., Levtov, Y., & Poirier-Quinot, D. (2017). Comparative perceptual evaluation between different context. In *Audio Engineering Society Convention 142*. Berlin, Germany. 23.

Pike, C., & Melchior, F. (2013). An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio. In *Audio Engineering Society Convention 134*. 67.

Pike, C., Melchior, F., & Tew, T. (2014). Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room. *AES 55th International Conference: Spatial Audio*, 1–8. 32 and 199.

Poirier-Quinot, D., Katz, B., & Noisternig, M. (2017). EVERTims: Open source framework for real-time auralization in VR. *ACM International Conference Proceeding Series, Part F1319*, 323–328. 92.

Pörschmann, C., Arend, J. M., & Neidhardt, A. (2017). A Spherical Near-Field HRTF Set for Auralization and Psychoacoustic Research. *Aes142*, eBrief 322. 20.

Pörschmann, C., & Wiefling, S. (2015). Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses. *International Conference on Spatial Audio*(December 2016). 20 and 21.

Postma, B. N. J., Demontis, H., & Katz, B. F. G. (2017). Subjective Evaluation of Dynamic Voice Directivity for Auralizations. *Acta Acustica united with Acustica*, *103*, 181–184. 16 and 28.

Postma, B. N. J., & Katz, B. F. G. (2016). Perceptive and objective evaluation of calibrated room acoustic simulation auralizations. *The Journal of the Acoustical Society of America*, *140*(6), 4326–4337. 16.

Postma, B. N. J., & Katz, B. F. G. (2017). The influence of visual distance on the room-acoustic experience of auralizations. *The Journal of the Acoustical Society of America*, *142*(5). 29.

Rämö, J., & Välimäki, V. (2014). An allpass hear-through headset. In *2014 22nd European Signal Processing Conference (EUSIPCO)* (pp. 1123–1127). 60 and 61.

Rindel, J. H., Otondo, F., & Christensen, C. L. (2004). Sound Source Representation for Auralization. In *International Symposium on Room Acoustics*. Kyoto, Japan. 15.

Robotham, T., Rummukainen, O. S., & Habets, E. A. P. (2019). Towards the Perception of Sound Source Directivity Inside Six-Degrees-of-Freedom Virtual Reality. In *5th International Conference on Spatial Audio ICSA* (pp. 1–8). Ilmenau, Germany. 17.

Roginska, A., & Geluso, P. (2017). *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*. Taylor & Francis. 1.

Rumsey, F. (1998, May). Controlled subjective assessments of 2-to-5-channel surround sound processing algorithms. In *Audio Engineering Society Convention 104*. 39.

Rumsey, F. (2007). Basic Psychoacoustics for Surround Recording. In *AES 22nd UK Conference* (pp. 1–9). University of Surrey, Guildford, UK. 65.

Rungta, A., Rust, S., Morales, N., Klatzky, R., Lin, M., & Manocha, D. (2016). Psychoacoustic

Characterization of Propagation Effects in Virtual Environments. *ACM Transactions on Applied Perception*, *13*(4), 1–18. 233.

Samoylenko, E. (1996). Systematic analysis of verbalizations produced in comparing musical timbres. *International Journal of Psychology*, *31*(6), 255-278. 49.

Sandvad, J. (1996). Dynamic aspects of auditory virtual environments. In *100th Audio Engineering Society Convention.* Copenhagen, Denmark. 26.

Savioja, L., & Svensson, U. P. (2015). Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, *138*(2), 708–730. 19.

Scharine, A. A., Cave, K. D., & Letowski, T. R. (1999). Auditory Perception and Cognitive Performance. In C. E. Rash, M. B. Russo, T. R. Letowski, & E. T. Schmeisser (Eds.), *Helmet Mounted Displays - Sensation, Perception and Cognitive Issues* (pp. 391–490). Fort Rucker, Alabama: U.S. Army Aeromedical Research Laboratory. 65.

Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N. C., & Nordahl, R. (2018). Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE Computer Graphics and Applications*, *38*(2), 31–43. 56.

Shabtai, N. R., Behler, G., Vorländer, M., & Weinzierl, S. (2017). Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments. *The Journal of the Acoustical Society of America*, *141*(2), 1246–1256. 16.

Shaw, M. L., & Gaines, B. R. (1989). Comparing conceptual structures: consensus, conflict, correspondence and contrast. *Knowledge Acquisition*, *1*(4), 341 - 363. 39.

Shinn-Cunningham, B. (2000). Distance cues for virtual auditory space. *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, 227–230. 66 and 68.

Shinn-Cunningham, B., & Ram, S. (2003). Identifying where you are in a room: Sensitivity to room acoustics. *Proceedings of the 9th International Conference on Auditory Display (ICAD2003)*, 21–24. 132 and 204.

Shivappa, S., Morrell, M., Sen, D., Peters, N., & Salehin, S. M. A. (2016, Sep). Efficient, compelling, and immersive vr audio experience using scene based audio/higher order ambisonics. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. 43.

Silzle, A. (2007, May). Quality taxonomies for auditory virtual environments. In *Audio Engineering Society Convention 122*. 42.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549–3557. 31.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*, 603-616. 31.

Southern, A., & Murphy, D. (2009). Low complexity directional sound sources for finite Difference Time Domain room acoustic models. *126th Audio Engineering Society Convention 2009*, *3*, 1126–1135. 16.

Speigle, J. M., & Loomis, J. M. (1993). Auditory distance perception by translating observers. *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium, VRAIS 1993*, 92–99. 226.

Susal, J., Krauss, K., Tsingos, N., & Altman, M. (2016, Sep). Immersive audio for vr. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. 43.

Takanen, M., Hiipakka, M., & Pulkki, V. (2012a). Audibility of coloration artifacts in hrtf filter designs. In *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. 66.

Takanen, M., Hiipakka, M., & Pulkki, V. (2012b). Audibility of coloration artifacts in HRTF filter designs. *Proceedings of the AES International Conference*, 8–16. 194.

Tatlow, S. (2024). Authenticity in sound design for virtual reality. In J. Cook, A. Kolassa, A. Robinson, & A. Whittaker (Eds.), *History as Fantasy in Music, Sound, Image, and Media* (1st ed., pp. 161–183). New York, USA: Routledge. 55.

Thery, D., Poirier-Quinot, D., Postma, B. N., & Katz, B. F. (2017). Impact of the visual rendering system on subjective auralization assessment in VR. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10700 LNCS*, 105–118. 28.

Toole, F. E. (1985). Subjective measurements of loudspeaker sound quality and listener performance. *J. Audio Eng. Soc*, *33*(1/2), 2–32. 39.

Unity Real-Time Development Platform | 3D, 2D VR & AR Visualizations. (2019). `https://unity.com/`. (Accessed on 08/01/2019) 69.

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2004). Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments. In *Proc. of Seventh Annual Workshop Presence 2004* (pp. 141–147). 25.

Valve. (2019). *Steam Audio.* `https://valvesoftware.github.io/steam-audio/`. (Accessed on 08/01/2019) 16 and 70.

Vorländer, M. (2008). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer. 93.

Wagner, I., Mccall, R., Morrison, A., & Valle, M. (2009). On the Role of Presence in Mixed Reality. *Presence, 18*(4), 249–276. 30.

Wallach, H. (1940). The Role of Head Movements and Vestibular and Visual Cues in Sound Localization. *Journal of Experimental Psychology, 27*(4), 339–368. 25 and 197.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1973). The Precedence Effect in Sound Localization (Tutorial Reprint). *Jaes, 21*(10), 817–826. 22.

Wendt, F., Zotter, F., Frank, M., & Höldrich, R. (2017). Auditory distance control using a variable-directivity loudspeaker. *Applied Sciences (Switzerland), 7*(7). 228.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America, 94*(1), 111–123. 194.

Werner, S., Klein, F., Mayenfels, T., & Brandenburg, K. (2016). A summary on acoustic room divergence and its effect on externalization of auditory events. *8th International Conference on Quality of Multimedia Experience, QoMEX*. 23, 136, and 204.

Werner, S., Klein, F., Neidhardt, A., Sloma, U., Schneiderwind, C., & Brandenburg, K. (2021). Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation. *Applied Sciences (Switzerland), 11*(3), 1–20. 20, 26, and 36.

Wersényi, G. (2009). Effect of emulated head-tracking for reducing localization errors in virtual audio simulation. *IEEE Transactions on Audio, Speech and Language Processing, 17*(2), 247–252. 25.

Whittington, W. (2007). *Sound design and science fiction*. University of Texas Press. 64.

Wightman, F., Kistler, D., & Andersen, K. (1994). Reassessment of the role of head movements in human sound localization. *The Journal of the Acoustical Society of America, 95*(5 Supplement), 3003–3004. 197 and 201.

Wightman, F. L., & Kistler, D. J. (2018). Factors Affecting the Relative Salience of Sound Localization Cues. In *Binaural and Spatial Hearing in Real and Virtual Environments* (pp. 1–24). 25.

Wirler, S. A., Meyer-Kahlen, N., & Schlecht, S. J. (2020). Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes. In *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*. Virtual. 35, 102, and 200.

Woszczyk, W., Bech, S., & Hansen, V. (1995, Oct). Interaction between audio-visual factors in a home theater system: Definition of subjective attributes. In *Audio Engineering Society Convention 99*. 39.

Yadav, M., Cabrera, D., Miranda, L., Martens, W. L., Lee, D., & Collins, R. (2013). Investigating auditory room size perception with autophonic stimuli. In *Audio Engineering Society Convention 135*. New York, USA. 23.

Yairi, S., Iwaya, Y., & Suzuki, Y. (2007). Estimation of detection threshold of system latency of virtual auditory display. *Applied Acoustics*, *68*(8), 851–863. 27, 197, and 201.

Zacharov, N., & Koivuniemi, K. (2001, Nov). Unravelling the perception of spatial sound reproduction: Analysis; external preference mapping. In *Audio Engineering Society Convention 111*. 41.

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, *111*(4), 1832–1846. 225 and 234.

Zahorik, P. (2003, apr). Auditory and visual distance perception: The proximity-image effect revisited. *The Journal of the Acoustical Society of America*, *113*(4 Supplement), 2270. 229.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, *91*(3), 409–420. 224.

Zhao, J., Zhang, B., Yan, Z., Wang, J., & Fei, Z. (2017). A study on the factors affecting audio-video subjective experience in virtual reality environments. In *International Conference on Virtual Reality and Visualization, ICVRV 2017* (pp. 303–306). IEEE. 58.

APPENDIX A

INFORMATION NOTICES

Information notice (Phase I)

The main goal of augmented reality audio is to superimpose virtual sound sources onto the real environment. In this type of application it is important that the sources are blending seamlessly with the environment so that it is difficult to distinguish the virtual sound from the real sources.

During this study we aim to evaluate the performance of different audio rendering systems in augmented reality context. The data gathered from you in the experiment will be anonymized and will not be used in future studies.

The session begins with a quick questionnaire on your demographic characteristics and background (age, sex, audio proficiency). You will then participate in a training sequence, the objective of which is to familiarize yourself with the experiment, its protocol and the response interface. You will be asked to stand on the point marked on the floor. You will wear headphones with a small tracking device attached which will send your position and head movements data in real time. During the trial you will be asked to walk following the path drawn on the floor back and forth while listening to speech stimuli, i.e. a phrase spoken by a male voice in english. The sentence will be repeated twice from two loudspeakers and for a total duration of 24s.

You will start walking when you see the green light in front of you and hear the sound stimulus. When you reach the end of the path and see an arrow sign on the iPad, you will rotate back and wait for the green light to start walking back. When walking, try to look ahead and adjust the speed so that you finish the path when the stimuli ends.

After finishing the path, you will be asked to respond to a questionnaire on the laptop. The first question will ask you to rate the plausibility that the sound played was coming from the real loudspeaker. You will be able to rate the plausibility on the scale using a slider. The next question will require identifying the localization of both audio clips. You will have to drag the circles labeled 1 or 2 which refer to the first and second stimulus respectively, and drop them on the picture presenting the room and the loudspeakers. The question about externalization will require choosing one of the 3 circle areas which indicate how externalized the sound seemed to be. The four next questions will display a slider on a scale which will allow you to rate the given perceptual attribute.

After each question you need to click the “Next” button to move to the next question. You can also walk the path again during the questionnaire but only one time during each trial. In order to do that, click on the “Play Stimuli Again” button which will be visible through the whole questionnaire. At any point during the test, you can leave your comment, just click on the button ”Leave a Comment”.

The training sequence includes four trials, after which the experience will begin. The experience consists of 40 trials with the same protocol as described above. There is no right or wrong answer, we are interested in your perception of sound depending on the audio rendering system. It is important to give identical concentration for each stimulus. You are free to take a break, whenever you need it.

The data collected will be for the purpose of evaluating the performance of sound systems tested. You have been assigned a code number, the stored data will thus be anonymous. The publication of the results will not include any individual results.

Thank you for your participation.

Total duration of the experience: about 60 min

Information notice (Phase II)

The main goal of augmented reality audio is to superimpose virtual sound sources onto the real environment. In this type of application it is important that the sources are blending seamlessly with the environment so that it is difficult to distinguish the virtual sound from the real sources.

During this study we aim to evaluate the performance of different audio rendering systems in augmented reality context. The data gathered from you in the experiment will be anonymized and will not be used in future studies.

The session begins with a quick questionnaire on your demographic characteristics and background (age, sex, audio proficiency). You will then participate in a training sequence, the objective of which is to familiarize yourself with the experiment, its protocol and the response interface. You will be asked to stand on the point marked on the floor. You will wear headphones with a small tracking device attached which will send your position and head movements data in real time. During the trial you will be asked to listen to speech stimuli, i.e. a phrase spoken by a male voice in English. The sentence will be repeated twice from two loudspeakers and for a total duration of 24s.

When sound playback stops, you will be asked to respond to a questionnaire on the laptop. The first question will ask you to rate the plausibility that the sound played was coming from the real loudspeaker. You will be able to rate the plausibility on the scale using a slider. The next question will require identifying the localization of both audio clips. You will have to drag the circles labeled 1 or 2 which refer to the first and second stimulus respectively, and drop them on the picture presenting room and the loudspeakers. The question about externalization will require choosing one of the 3 circle areas which indicate how externalized the sound seemed to be. The four next questions will display a slider on a scale which will allow you to rate the given perceptual attribute. After each question you need to click the "Next" button to move to the next question. You can also walk the path again during the questionnaire but only for one time. In order to do that, click on the "Play Stimuli Again" button which will be visible through the whole questionnaire. At any point during the test, you can leave your comment, just click on the button "Leave a Comment".

The training sequence includes four trials, after which the experience will begin. The experience consists of 40 trials with the same protocol as described above. There is no right or wrong answer, we are interested in your perception of sound depending on the audio rendering system. It is important to give identical concentration for each stimulus. You are free to take a break, whenever you need it.

The data collected will be for the purpose of evaluating the performance of sound systems tested. You have been assigned a code number, the stored data will thus be anonymous. The publication of the results will not include any individual results.

Thank you for your participation.

Total duration of the experience: about 60 min

Demographic questionnaire

1. What is your age?
2. What is your gender?
3. Do you have any hearing impairment (Yes or no)?
4. Please indicate the number of years of formal musical training that you received
5. Have you ever previously participated in audio listening tests?
6. Have you ever previously participated in spatial audio listening tests?
7. What is your occupation?
8. What is your height in meters? (e.g. 1.71)

APPENDIX B

SIMULATIONS DIAGRAMS

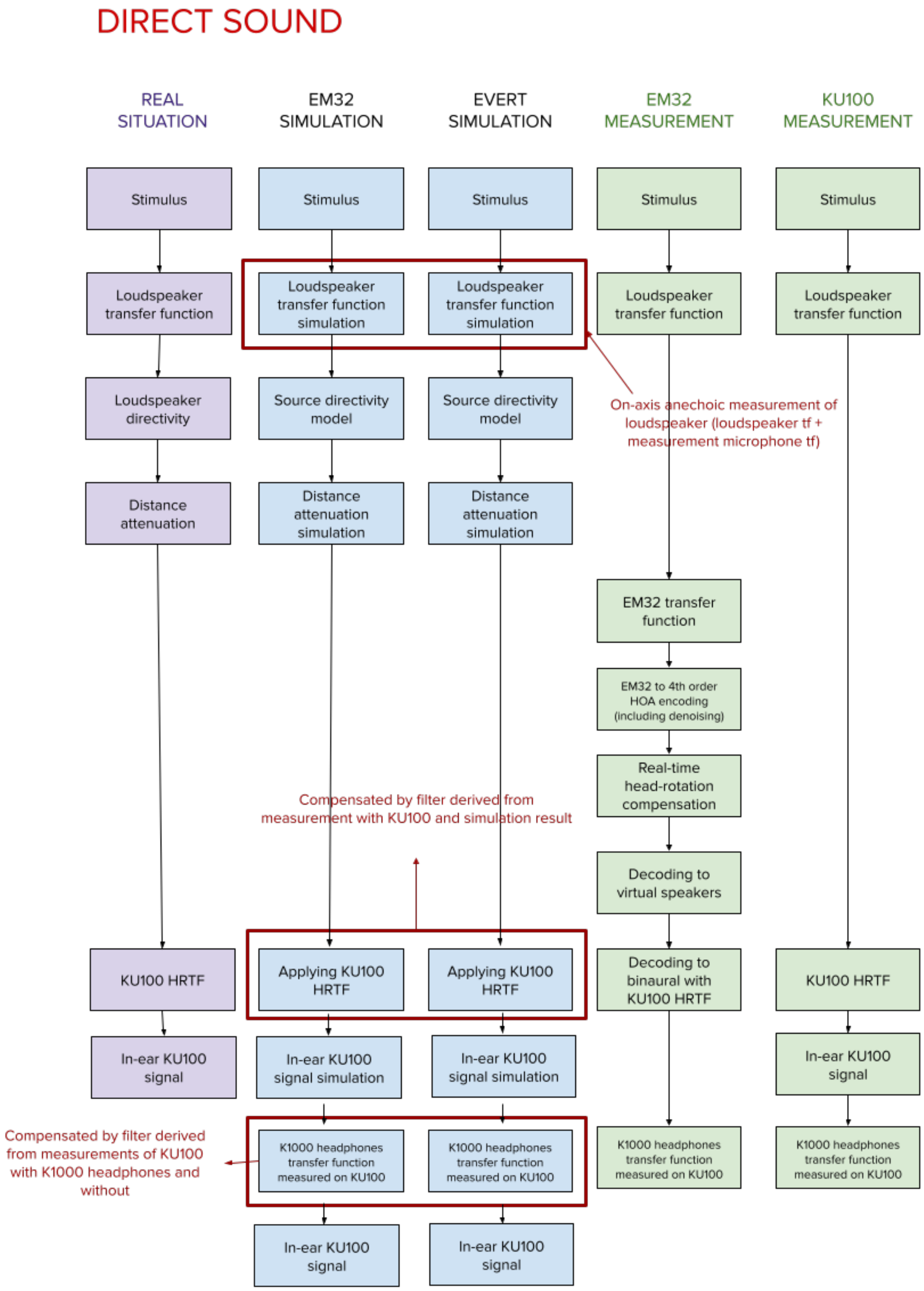
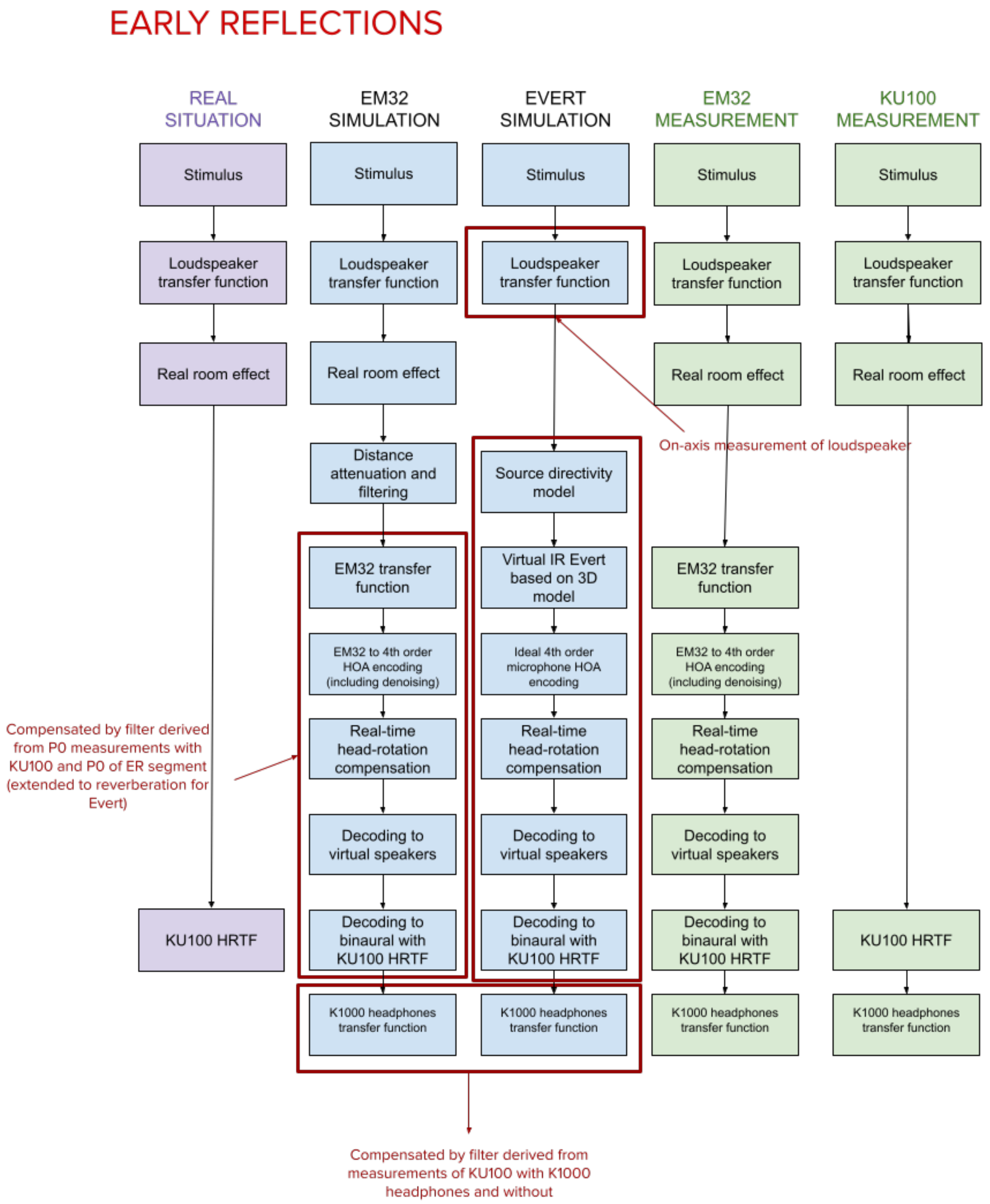

DIRECT SOUND
REAL SITUATION
EM32 SIMULATION
EVERT SIMULATION
EM32 MEASUREMENT
KU100 MEASUREMENT
Stimulus
Stimulus
Stimulus
Stimulus
Stimulus
Loudspeaker transfer function
Loudspeaker transfer function simulation
Loudspeaker transfer function simulation
Loudspeaker transfer function
Loudspeaker transfer function
Loudspeaker directivity
Source directivity model
Source directivity model
On-axis anechoic measurement of loudspeaker (loudspeaker tf + measurement microphone tf)
Distance attenuation
Distance attenuation simulation
Distance attenuation simulation
EM32 transfer function
EM32 to 4th order HOA encoding (including denoising)
Real-time head-rotation compensation
Compensated by filter derived from measurement with KU100 and simulation result
Decoding to virtual speakers
KU100 HRTF
Applying KU100 HRTF
Applying KU100 HRTF
Decoding to binaural with KU100 HRTF
KU100 HRTF
In-ear KU100 signal
In-ear KU100 signal simulation
In-ear KU100 signal simulation
In-ear KU100 signal
Compensated by filter derived from measurements of KU100 with K1000 headphones and without
K1000 headphones transfer function measured on KU100
K1000 headphones transfer function measured on KU100
K1000 headphones transfer function measured on KU100
K1000 headphones transfer function measured on KU100
In-ear KU100 signal
In-ear KU100 signal

Figure 58: Elements of the direct sound segment sound path for real situation, simulations and measurements


EARLY REFLECTIONS
REAL SITUATION
EM32 SIMULATION
EVERT SIMULATION
EM32 MEASUREMENT
KU100 MEASUREMENT
Stimulus
Stimulus
Stimulus
Stimulus
Stimulus
Loudspeaker transfer function
Loudspeaker transfer function
Loudspeaker transfer function
Loudspeaker transfer function
Loudspeaker transfer function
Real room effect
Real room effect
Real room effect
Real room effect
Distance attenuation and filtering
Source directivity model
EM32 transfer function
Virtual IR Evert based on 3D model
EM32 transfer function
EM32 to 4th order HOA encoding (including denoising)
Ideal 4th order microphone HOA encoding
EM32 to 4th order HOA encoding (including denoising)
Real-time head-rotation compensation
Real-time head-rotation compensation
Real-time head-rotation compensation
Decoding to virtual speakers
Decoding to virtual speakers
Decoding to virtual speakers
KU100 HRTF
Decoding to binaural with KU100 HRTF
Decoding to binaural with KU100 HRTF
Decoding to binaural with KU100 HRTF
KU100 HRTF
K1000 headphones transfer function
K1000 headphones transfer function
K1000 headphones transfer function
K1000 headphones transfer function

On-axis measurement of loudspeaker

Compensated by filter derived from P0 measurements with KU100 and P0 of ER segment (extended to reverberation for Evert)

Compensated by filter derived from measurements of KU100 with K1000 headphones and without

Figure 59: Elements of the early reflections segment sound path for real situation, simulations and measurements

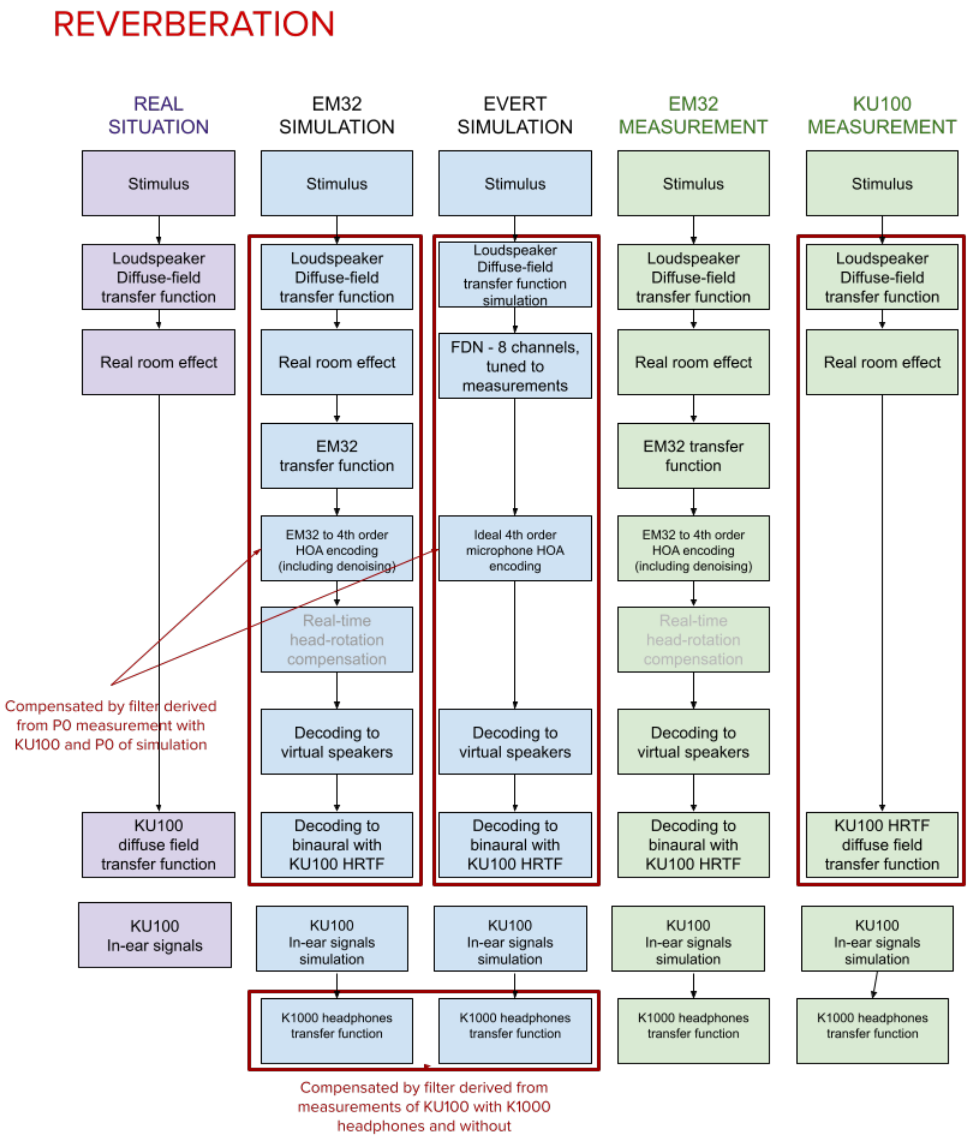

REVERBERATION
REAL SITUATION
EM32 SIMULATION
EVERT SIMULATION
EM32 MEASUREMENT
KU100 MEASUREMENT
Stimulus
Loudspeaker Diffuse-field transfer function
Real room effect
KU100 diffuse field transfer function
KU100 In-ear signals
Stimulus
Loudspeaker Diffuse-field transfer function
Real room effect
EM32 transfer function
EM32 to 4th order HOA encoding (including denoising)
Real-time head-rotation compensation
Decoding to virtual speakers
Decoding to binaural with KU100 HRTF
KU100 In-ear signals simulation
K1000 headphones transfer function
Stimulus
Loudspeaker Diffuse-field transfer function simulation
FDN - 8 channels, tuned to measurements
Ideal 4th order microphone HOA encoding
Decoding to virtual speakers
Decoding to binaural with KU100 HRTF
KU100 In-ear signals simulation
K1000 headphones transfer function
Stimulus
Loudspeaker Diffuse-field transfer function
Real room effect
EM32 transfer function
EM32 to 4th order HOA encoding (including denoising)
Real-time head-rotation compensation
Decoding to virtual speakers
Decoding to binaural with KU100 HRTF
KU100 In-ear signals simulation
K1000 headphones transfer function
Stimulus
Loudspeaker Diffuse-field transfer function
Real room effect
KU100 HRTF diffuse field transfer function
KU100 In-ear signals simulation
K1000 headphones transfer function
Compensated by filter derived from P0 measurement with KU100 and P0 of simulation
Compensated by filter derived from measurements of KU100 with K1000 headphones and without

Figure 60: Elements of the reverberation segment sound path for real situation, simulations and measurements

APPENDIX C

STATISTICAL ANALYSIS - WALKING PHASE

1 Plausibility

Table 29

Model selection for plausibility ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method*speaker+order+speed	16	10716.76	0.00	0.68	-5342.30
method*speaker+order+speed+height	17	10718.50	1.74	0.28	-5342.15
method*speaker+speed	15	10722.54	5.77	0.04	-5346.19
method*speaker+order	15	10730.36	13.59	0.00	-5350.10
method*speaker+yearsofmusic	31	10737.33	20.56	0.00	-5337.35
method*speaker	14	10737.58	20.82	0.00	-5354.72
method*speaker+method_pair	18	10738.04	21.28	0.00	-5350.91
method*speaker+audio_test	15	10738.64	21.88	0.00	-5354.24
method*speaker+spatial_audio_test	15	10739.31	22.54	0.00	-5354.58
method*speaker+height	15	10739.34	22.57	0.00	-5354.59
method*speaker+index	17	10741.92	25.16	0.00	-5353.86
method+speaker	8	10811.80	95.03	0.00	-5397.87
speaker	6	10985.84	269.08	0.00	-5486.91
method	5	11077.09	360.32	0.00	-5533.53
null	3	11235.98	519.21	0.00	-5614.99

Table 30

Pairwise comparison for plausibility ratings

contrast	estimate	SE	df	t.ratio	p.value
Method = GA					
A - B	0.08	0.13	3126.01	0.602	0.9315
A - C	0.31	0.13	3126.03	2.339	0.0896
A - D	1.69	0.13	3126.02	12.797	<.0001
B - C	0.23	0.13	3126.03	1.741	0.3024
B - D	1.61	0.13	3126.01	12.227	<.0001
C - D	1.38	0.13	3126.06	10.487	<.0001

Method = R					
A - B	0.20	0.09	3126.02	2.175	0.1304
A - C	-0.09	0.09	3126.10	-0.975	0.7636
A - D	0.53	0.09	3126.74	5.744	<.0001
B - C	-0.29	0.09	3126.18	-3.152	0.0089
B - D	0.33	0.09	3126.93	3.565	0.0021
C - D	0.62	0.09	3126.30	6.733	<.0001
Method = SRIR					
A - B	0.07	0.13	3126.11	0.539	0.9495
A - C	0.15	0.13	3126.36	1.118	0.6782
A - D	1.21	0.13	3126.12	9.216	<.0001
B - C	0.08	0.13	3126.10	0.580	0.9380
B - D	1.14	0.13	3126.01	8.660	<.0001
C - D	1.06	0.13	3126.10	8.057	<.0001

Results are averaged over the levels of: Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 31

Pairwise comparison for plausibility ratings

contrast	estimate	SE	df	t.ratio	p.value
Ldspkr = A					
GA - R	-0.35	0.11	3126.15	-3.057	0.0064
GA - SRIR	-0.03	0.13	3126.06	-0.266	0.9617
R - SRIR	0.31	0.11	3126.25	2.769	0.0156
Ldspkr = B					
GA - R	-0.23	0.11	3126.08	-1.993	0.1141
GA - SRIR	-0.04	0.13	3126.03	-0.332	0.9411
R - SRIR	0.18	0.11	3126.03	1.609	0.2419
Ldspkr = C					
GA - R	-0.75	0.11	3126.17	-6.576	<.0001
GA - SRIR	-0.20	0.13	3126.07	-1.487	0.2972
R - SRIR	0.55	0.11	3126.08	4.835	<.0001
Ldspkr = D					
GA - R	-1.50	0.11	3127.22	-13.191	<.0001
GA - SRIR	-0.51	0.13	3126.04	-3.877	0.0003
R - SRIR	0.99	0.11	3126.83	8.700	<.0001

Results are averaged over the levels of: Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

2 Blur

Table 32

Model selection for blur ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method*speaker+index+speed+order	19	10518.40	0.00	0.99	-5240.08
method*speaker+index+speed	18	10527.42	9.02	0.01	-5245.60
method*speaker+index+order	18	10532.64	14.24	0.00	-5248.21
method*speaker+index	17	10540.00	21.60	0.00	-5252.90
method*speaker+speed	15	10545.71	27.30	0.00	-5257.78
method*speaker+order	15	10547.00	28.60	0.00	-5258.42
method*speaker	14	10554.31	35.90	0.00	-5263.09
method*speaker+audio_test	15	10556.14	37.74	0.00	-5262.99
method*speaker+height	15	10556.20	37.80	0.00	-5263.03
method*speaker+spatial_audio_test	15	10556.26	37.85	0.00	-5263.05
method*speaker+method_pair	18	10562.23	43.83	0.00	-5263.01
method*speaker+years_of_music	31	10564.64	46.24	0.00	-5251.00
method+speaker	8	10620.42	102.02	0.00	-5302.19
speaker	6	10747.49	229.09	0.00	-5367.73
method	5	11587.94	1069.54	0.00	-5788.96
null	3	11679.30	1160.90	0.00	-5836.65

Table 33

Pairwise comparison for blur ratings

contrast	estimate	SE	df	t.ratio	p.value
Method = GA					
A - B	-0.06	0.13	3129.05	-0.490	0.9613
A - C	-0.43	0.13	3129.06	-3.367	0.0043
A - D	-2.41	0.13	3129.06	-18.911	<.0001
B - C	-0.37	0.13	3129.07	-2.884	0.0206
B - D	-2.35	0.13	3129.04	-18.466	<.0001
C - D	-1.98	0.13	3129.10	-15.589	<.0001
Method = R					
A - B	-0.26	0.09	3129.06	-2.910	0.0190
A - C	-0.10	0.09	3129.12	-1.121	0.6763

A - D	-1.49	0.09	3129.61	-16.609	<.0001
B - C	0.16	0.09	3129.21	1.793	0.2769
B - D	-1.23	0.09	3129.79	-13.672	<.0001
C - D	-1.39	0.09	3129.26	-15.524	<.0001
Method = SRIR					
A - B	-0.13	0.13	3129.12	-1.064	0.7115
A - C	-0.12	0.13	3129.37	-0.946	0.7801
A - D	-2.23	0.13	3129.14	-17.559	<.0001
B - C	0.01	0.13	3129.14	0.114	0.9995
B - D	-2.10	0.13	3129.05	-16.461	<.0001
C - D	-2.11	0.13	3129.13	-16.533	<.0001

Results are averaged over the levels of: Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 34

Pairwise comparison for blur ratings

contrast	estimate	SE	df	t.ratio	p.value
Ldspkr = A					
GA - R	0.29	0.11	3129.17	2.631	0.0232
GA - SRIR	0.05	0.13	3129.10	0.429	0.9035
R - SRIR	-0.24	0.11	3129.32	-2.149	0.0803
Ldspkr = B					
GA - R	0.09	0.11	3129.14	0.839	0.6785
GA - SRIR	-0.02	0.13	3129.07	-0.141	0.9892
R - SRIR	-0.11	0.11	3129.07	-1.002	0.5755
Ldspkr = C					
GA - R	0.62	0.11	3129.22	5.629	<.0001
GA - SRIR	0.36	0.13	3129.10	2.849	0.0123
R - SRIR	-0.26	0.11	3129.11	-2.322	0.0529
Ldspkr = D					
GA - R	1.22	0.11	3130.15	10.998	<.0001
GA - SRIR	0.23	0.13	3129.07	1.839	0.1571
R - SRIR	-0.98	0.11	3129.79	-8.867	<.0001

Results are averaged over the levels of: Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 35

Pairwise comparison for blur ratings

contrast	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	0.19	0.06	3131.40	2.936	0.0176
(1-12) - (25-36)	0.26	0.06	3130.48	4.033	0.0003
(1-12) - (37-48)	0.29	0.06	3130.95	4.540	<.0001
(13-24) - (25-36)	0.07	0.06	3129.18	1.087	0.6975
(13-24) - (37-48)	0.10	0.06	3129.08	1.618	0.3687
(25-36) - (37-48)	0.03	0.06	3129.08	0.540	0.9493

Results are averaged over the levels of: Method, Ldspkr, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

3 Localization Error

Table 36

Model selection for localization error ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method*speaker+index+order	18	778.49	0.00	0.99	-371.13
method*speaker+order	15	787.59	9.10	0.01	-378.72
method*speaker+index	17	799.91	21.42	0.00	-382.85
method*speaker	14	808.92	30.43	0.00	-390.39
method*speaker+spatial_audio_test	15	809.90	31.41	0.00	-389.87
method*speaker+speed	15	810.26	31.77	0.00	-390.05
method*speaker+audio_test	15	810.81	32.32	0.00	-390.33
method*speaker+height	15	810.83	32.34	0.00	-390.34
method*speaker+method_pair	18	813.00	34.51	0.00	-388.38
method+speaker	8	833.35	54.86	0.00	-408.65
method	5	837.06	58.57	0.00	-413.52
speaker	6	894.97	116.48	0.00	-441.47
null	3	898.19	119.70	0.00	-446.09

Table 37

Pairwise comparison for localization error ratings

contrast	estimate	SE	df	t.ratio	p.value
Method = GA					
A - B	-0.03	0.03	2996.13	-1.022	0.7365
A - C	-0.02	0.03	2996.12	-0.807	0.8510
A - D	-0.11	0.03	2996.23	-4.069	0.0003
B - C	0.01	0.03	2996.18	0.211	0.9967
B - D	-0.09	0.03	2996.31	-3.057	0.0121
C - D	-0.09	0.03	2996.18	-3.256	0.0063
Method = R					
A - B	-0.03	0.02	2996.19	-1.353	0.5292
A - C	0.02	0.02	2996.10	1.065	0.7110
A - D	0.04	0.02	2996.15	1.970	0.1996
B - C	0.05	0.02	2996.21	2.417	0.0742
B - D	0.06	0.02	2996.31	3.336	0.0048
C - D	0.02	0.02	2996.17	0.892	0.8090
Method = SRIR					
A - B	0.01	0.03	2996.24	0.467	0.9663
A - C	0.04	0.03	2996.16	1.311	0.5561
A - D	-0.07	0.03	2996.17	-2.424	0.0728
B - C	0.02	0.03	2996.14	0.860	0.8253

B - D	-0.08	0.03	2996.14	-2.931	0.0179
C - D	-0.11	0.03	2996.17	-3.763	0.0010

Results are averaged over the levels of: Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 38

Pairwise comparison for localization error ratings

contrast	estimate	SE	df	t.ratio	p.value
Ldspkr = A					
GA - R	0.03	0.02	2996.16	1.425	0.3282
GA - SRIR	-0.02	0.03	2996.24	-0.878	0.6544
R - SRIR	-0.06	0.02	2996.26	-2.406	0.0427
Ldspkr = B					
GA - R	0.04	0.02	2996.37	1.503	0.2896
GA - SRIR	0.02	0.03	2996.12	0.606	0.8168
R - SRIR	-0.02	0.02	2996.27	-0.808	0.6983
Ldspkr = C					
GA - R	0.08	0.02	2996.14	3.212	0.0038
GA - SRIR	0.03	0.03	2996.15	1.245	0.4270
R - SRIR	-0.04	0.02	2996.19	-1.767	0.1808
Ldspkr = D					
GA - R	0.19	0.02	2996.34	7.709	<.0001
GA - SRIR	0.02	0.03	2996.14	0.735	0.7427
R - SRIR	-0.17	0.02	2996.24	-6.880	<.0001

Results are averaged over the levels of: Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 39

Pairwise comparison of trial index for localization error ratings

contrast	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	0.05	0.01	2996.26	3.496	0.0027
(1-12) - (25-36)	0.04	0.01	2996.37	2.677	0.0375
(1-12) - (37-48)	0.04	0.01	2996.28	3.150	0.0090
(13-24) - (25-36)	-0.01	0.01	2996.28	-0.825	0.8429
(13-24) - (37-48)	-0.00	0.01	2996.27	-0.337	0.9869
(25-36) - (37-48)	0.01	0.01	2996.37	0.489	0.9617

Results are averaged over the levels of: Method, Ldspkr, Order

Degrees-of-freedom method: kenward-roger

```
P value adjustment: tukey method for comparing a family of 4 estimates
```

4 Externalization

Table 40

Model selection for externalization ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method*speaker+order+method_pair	18	2574.90	0.00	0.50	-1269.34
method*speaker+order	14	2575.34	0.44	0.40	-1273.60
method+speaker+order+method_pair	12	2579.43	4.53	0.05	-1277.66
method+speaker+order	8	2579.84	4.94	0.04	-1281.90
method*speaker	13	2637.51	62.61	0.00	-1305.69
method*speaker+speed	14	2638.14	63.25	0.00	-1305.01
method*speaker+height	14	2639.15	64.25	0.00	-1305.51
method+speaker+method_pair	11	2641.44	66.54	0.00	-1309.68
method+speaker	7	2641.64	66.74	0.00	-1313.80
method+speaker+spatial_audio_test	8	2641.88	66.98	0.00	-1312.92
method+speaker+audio_test	8	2642.93	68.03	0.00	-1313.44
method*speaker+index	16	2643.20	68.30	0.00	-1305.51
method+speaker+years_music	24	2651.01	76.11	0.00	-1301.31
speaker	5	2724.29	149.39	0.00	-1357.14
method	4	2738.43	163.53	0.00	-1365.21
null	2	2817.14	242.24	0.00	-1406.57

Table 41

Pairwise comparison of rendering method for externalization ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
Ldspkr = A					
GA - R	-0.50	0.24	Inf	-2.096	0.0907
GA - SRIR	0.37	0.27	Inf	1.403	0.3392
R - SRIR	0.87	0.23	Inf	3.732	0.0006
Ldspkr = B					
GA - R	-0.80	0.23	Inf	-3.444	0.0017
GA - SRIR	0.02	0.26	Inf	0.089	0.9956
R - SRIR	0.82	0.23	Inf	3.548	0.0011
Ldspkr = C					
GA - R	-0.89	0.24	Inf	-3.757	0.0005
GA - SRIR	0.12	0.27	Inf	0.454	0.8924
R - SRIR	1.01	0.24	Inf	4.259	0.0001
Ldspkr = D					
GA - R	-1.88	0.28	Inf	-6.768	<.0001
GA - SRIR	-0.68	0.29	Inf	-2.372	0.0466
R - SRIR	1.20	0.29	Inf	4.135	0.0001

Results are averaged over the levels of: Order
Results are given on the log odds ratio (not the response) scale.
P value adjustment: tukey method for comparing a family of 3 estimates

Table 42

Pairwise comparison of loudspeaker position for externalization ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
Method = GA					
A - B	0.49	0.27	Inf	1.837	0.2558
A - C	0.16	0.27	Inf	0.616	0.9271
A - D	-0.29	0.27	Inf	-1.064	0.7117
B - C	-0.32	0.26	Inf	-1.227	0.6095
B - D	-0.78	0.27	Inf	-2.888	0.0202
C - D	-0.45	0.27	Inf	-1.678	0.3351
Method = R					
A - B	0.19	0.20	Inf	0.958	0.7732
A - C	-0.23	0.20	Inf	-1.154	0.6560
A - D	-1.68	0.24	Inf	-6.916	<.0001
B - C	-0.42	0.20	Inf	-2.110	0.1499
B - D	-1.86	0.24	Inf	-7.725	<.0001
C - D	-1.44	0.24	Inf	-5.925	<.0001
Method = SRIR					
A - B	0.14	0.26	Inf	0.534	0.9507
A - C	-0.09	0.26	Inf	-0.330	0.9876
A - D	-1.35	0.28	Inf	-4.751	<.0001
B - C	-0.23	0.26	Inf	-0.859	0.8260
B - D	-1.49	0.28	Inf	-5.230	<.0001
C - D	-1.26	0.29	Inf	-4.414	0.0001

Results are averaged over the levels of: Order
Results are given on the log odds ratio (not the response) scale.
P value adjustment: tukey method for comparing a family of 4 estimates

5 Loudspeaker Recognition Rate

Table 43

Model selection for speaker recognition error

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method*speaker*order+index+speed	30	1147.04	0.00	0.94	-543.22
method*speaker*order+index	29	1152.51	5.47	0.06	-546.98
method*speaker*order	26	1168.43	21.39	0.00	-557.99
method*speaker*order+height	27	1170.36	23.32	0.00	-557.94
method*speaker+order+index	18	1202.31	55.27	0.00	-583.05
method*speaker+order	15	1218.47	71.43	0.00	-594.16
method*speaker+order+height+	16	1220.39	73.35	0.00	-594.11
method*speaker	14	1234.15	87.10	0.00	-603.01
method*speaker+spatial_audio_test	15	1234.89	87.84	0.00	-602.37
method*speaker+audio_test	15	1235.35	88.31	0.00	-602.60
method*speaker+height	15	1236.06	89.02	0.00	-602.96
method*speaker+method_pair	18	1238.34	91.30	0.00	-601.06
method+speaker	8	1264.77	117.73	0.00	-624.36
method	5	1274.96	127.92	0.00	-632.47
speaker	6	1286.50	139.46	0.00	-637.24
null	3	1296.54	149.50	0.00	-645.27

Table 44

Pairwise comparison for loudspeaker recognition rate

contrast	estimate	SE	df	t.ratio	p.value
Method = GA, Order = 1					
A - B	0.04	0.04	3140.10	1.065	0.7109
A - C	-0.04	0.04	3140.08	-1.101	0.6891
A - D	0.20	0.04	3140.11	4.835	<.0001
B - C	-0.09	0.04	3140.13	-2.165	0.1334
B - D	0.15	0.04	3140.12	3.771	0.0010
C - D	0.24	0.04	3140.19	5.927	<.0001
Method = R, Order = 1					
A - B	0.01	0.03	3140.80	0.294	0.9912
A - C	-0.09	0.03	3143.74	-3.047	0.0125
A - D	-0.08	0.03	3142.90	-2.814	0.0253
B - C	-0.10	0.03	3141.35	-3.343	0.0047
B - D	-0.09	0.03	3140.78	-3.110	0.0102
C - D	0.01	0.03	3140.23	0.231	0.9957
Method = SRIR, Order = 1					
A - B	-0.07	0.04	3141.20	-1.751	0.2976

A - C	-0.16	0.04	3141.25	-4.031	0.0003
A - D	0.11	0.04	3140.75	2.738	0.0316
B - C	-0.09	0.04	3140.20	-2.298	0.0987
B - D	0.18	0.04	3140.21	4.487	<.0001
C - D	0.28	0.04	3140.21	6.741	<.0001
Method = GA, Order = 2					
A - B	0.04	0.04	3140.12	0.976	0.7633
A - C	0.05	0.04	3140.17	1.246	0.5974
A - D	0.09	0.04	3140.15	2.248	0.1109
B - C	0.01	0.04	3140.22	0.270	0.9931
B - D	0.05	0.04	3140.07	1.281	0.5751
C - D	0.04	0.04	3140.23	1.015	0.7405
Method = R, Order = 2					
A - B	0.02	0.03	3141.41	0.676	0.9061
A - C	0.08	0.03	3141.38	2.853	0.0226
A - D	0.04	0.03	3140.31	1.229	0.6085
B - C	0.06	0.03	3140.12	2.174	0.1307
B - D	0.02	0.03	3142.29	0.548	0.9470
C - D	-0.05	0.03	3142.36	-1.628	0.3629
Method = SRIR, Order = 2					
A - B	-0.01	0.04	3140.43	-0.362	0.9837
A - C	0.01	0.04	3140.48	0.282	0.9922
A - D	-0.00	0.04	3140.30	-0.040	1.0000
B - C	0.03	0.04	3140.59	0.639	0.9193
B - D	0.01	0.04	3140.10	0.321	0.9885
C - D	-0.01	0.04	3140.34	-0.320	0.9887

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 45

Pairwise comparison for loudspeaker recognition rate

contrast	estimate	SE	df	t.ratio	p.value
Method = GA, Ldspkr = A					
1 - 2	-0.08	0.04	3140.10	-1.850	0.0644
Method = R, Ldspkr = A					
1 - 2	-0.09	0.03	3140.41	-3.238	0.0012
Method = SRIR, Ldspkr = A					
1 - 2	-0.09	0.04	3140.76	-2.332	0.0197
Method = GA, Ldspkr = B					
1 - 2	-0.08	0.04	3140.13	-1.947	0.0517

Method = R, Ldspkr = B					
1 - 2	-0.08	0.03	3142.09	-2.845	0.0045
Method = SRIR, Ldspkr = B					
1 - 2	-0.04	0.04	3140.56	-0.936	0.3496
Method = GA, Ldspkr = C					
1 - 2	0.02	0.04	3140.10	0.492	0.6227
Method = R, Ldspkr = C					
1 - 2	0.08	0.03	3146.16	2.656	0.0079
Method = SRIR, Ldspkr = C					
1 - 2	0.08	0.04	3140.10	1.995	0.0462
Method = GA, Ldspkr = D					
1 - 2	-0.18	0.04	3140.13	-4.425	<.0001
Method = R, Ldspkr = D					
1 - 2	0.02	0.03	3140.82	0.808	0.4193
Method = SRIR, Ldspkr = D					
1 - 2	-0.21	0.04	3140.15	-5.088	<.0001

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

Table 46

Pairwise comparison for loudspeaker recognition rate

contrast	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	-0.05	0.01	3148.12	-3.209	0.0074
(1-12) - (25-36)	-0.06	0.01	3145.35	-4.319	0.0001
(1-12) - (37-48)	-0.06	0.01	3146.63	-4.283	0.0001
(13-24) - (25-36)	-0.02	0.01	3140.49	-1.097	0.6915
(13-24) - (37-48)	-0.02	0.01	3140.25	-1.086	0.6983
(25-36) - (37-48)	0.00	0.01	3140.18	0.007	1.0000

Results are averaged over the levels of: Order, Method, Ldspkr

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

6 Timbre Difference

Table 47

Model selection for timbre difference ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method+speaker+index	11	5367.42	0.00	0.63	-2672.63
method+speaker	8	5370.58	3.16	0.13	-2677.24
method*speaker	12	5371.49	4.07	0.08	-2673.65
method+speaker+spatial_audio_test	9	5371.88	4.46	0.07	-2676.88
method+speaker+audio_test	9	5372.60	5.18	0.05	-2677.24
method*speaker+height	13	5372.89	5.47	0.04	-2673.33
speaker	4	5436.42	69.00	0.00	-2714.20
method	7	5619.22	251.80	0.00	-2802.57
null	3	5673.62	306.20	0.00	-2833.80
height	4	5675.00	307.58	0.00	-2833.49

Table 48

Pairwise comparison loudspeaker position pair for timbre difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(A-C) - (B-D)	-1.05	0.06	1556.04	-16.538	<.0001

Results are averaged over the levels of: method_pair, Index_id

Degrees-of-freedom method: kenward-roger

Table 49

Pairwise comparison of rendering method pair for timbre difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(GA-GA) - (GA-R)	-0.39	0.13	1556.04	-3.106	0.0165
(GA-GA) - (R-R)	0.35	0.14	1556.04	2.534	0.0837
(GA-GA) - (SRIR-R)	-0.15	0.13	1556.04	-1.196	0.7537
(GA-GA) - (SRIR-SRIR)	0.37	0.16	1556.05	2.392	0.1180
(GA-R) - (R-R)	0.74	0.10	1556.05	7.780	<.0001
(GA-R) - (SRIR-R)	0.24	0.08	1556.04	3.109	0.0164
(GA-R) - (SRIR-SRIR)	0.77	0.13	1556.05	6.044	<.0001
(R-R) - (SRIR-R)	-0.50	0.10	1556.05	-5.243	<.0001
(R-R) - (SRIR-SRIR)	0.02	0.14	1556.05	0.170	0.9998
(SRIR-R) - (SRIR-SRIR)	0.52	0.13	1556.05	4.145	0.0003

Results are averaged over the levels of: Ldspkr_pair, Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

Table 50

Pairwise comparison of trial index for timbre difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	-0.13	0.09	1556.05	-1.400	0.4992
1 - 3	0.12	0.09	1556.04	1.343	0.5355
1 - 4	0.11	0.09	1556.05	1.163	0.6506
2 - 3	0.25	0.09	1556.04	2.726	0.0328
2 - 4	0.24	0.09	1556.04	2.517	0.0577
3 - 4	-0.01	0.09	1556.04	-0.152	0.9987

Results are averaged over the levels of: method_pair, Ldspkr_pair

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

7 Reverberation Difference

Table 51

Model selection for reverberation difference ratings in walking phase

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method+speaker+audio_test	9	5185.59	0.00	0.33	-2583.74
method+speaker+spatial_audio_test	9	5185.91	0.32	0.28	-2583.90
method+speaker	8	5186.54	0.95	0.20	-2585.22
method+speaker+height	9	5188.38	2.79	0.08	-2585.13
method*speaker	12	5189.20	3.61	0.05	-2582.50
method+speaker+index	11	5189.20	3.61	0.05	-2583.52
speaker	4	5227.20	41.61	0.00	-2609.59
method	7	5821.91	636.33	0.00	-2903.92
null	3	5845.67	660.08	0.00	-2919.83
height	4	5847.51	661.92	0.00	-2919.74

Table 52

Pairwise comparison of loudspeaker position pair for reverberation difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(A-C) - (B-D)	-1.69	0.06	1553.02	-28.038	<.0001

Results are averaged over the levels of: method_pair

Degrees-of-freedom method: kenward-roger

Table 53

Pairwise comparison of rendering method pair for reverberation difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(GA-GA) - (GA-R)	-0.03	0.12	1553.02	-0.300	0.9982
(GA-GA) - (R-R)	0.54	0.13	1553.02	4.219	0.0003
(GA-GA) - (SRIR-R)	0.19	0.12	1553.02	1.658	0.4606
(GA-GA) - (SRIR-SRIR)	0.43	0.15	1553.02	2.908	0.0303
(GA-R) - (R-R)	0.57	0.09	1553.02	6.348	<.0001
(GA-R) - (SRIR-R)	0.23	0.07	1553.02	3.095	0.0171
(GA-R) - (SRIR-SRIR)	0.46	0.12	1553.02	3.972	0.0007
(R-R) - (SRIR-R)	-0.35	0.09	1553.02	-3.825	0.0013
(R-R) - (SRIR-SRIR)	-0.11	0.13	1553.03	-0.855	0.9130
(SRIR-R) - (SRIR-SRIR)	0.24	0.12	1553.03	2.021	0.2564

Results are averaged over the levels of: Ldspkr_pair

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

8 Plausibility Difference

Table 55

Pairwise comparison of loudspeaker position pair for plausibility difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(A-C) - (B-D)	-0.77	0.06	1553.02	-12.362	<.0001

Results are averaged over the levels of: method_pair

Degrees-of-freedom method: kenward-roger

Table 56

Pairwise comparison of rendering method pair for plausibility difference ratings in walking phase

contrast	estimate	SE	df	t.ratio	p.value
(GA-GA) - (GA-R)	-0.16	0.12	1553.02	-1.312	0.6838
(GA-GA) - (R-R)	0.42	0.13	1553.02	3.142	0.0147
(GA-GA) - (SRIR-R)	0.04	0.12	1553.02	0.297	0.9983
(GA-GA) - (SRIR-SRIR)	0.05	0.15	1553.04	0.298	0.9983
(GA-R) - (R-R)	0.58	0.09	1553.03	6.131	<.0001
(GA-R) - (SRIR-R)	0.19	0.08	1553.02	2.543	0.0819
(GA-R) - (SRIR-SRIR)	0.20	0.12	1553.03	1.684	0.4439
(R-R) - (SRIR-R)	-0.38	0.09	1553.03	-4.058	0.0005
(R-R) - (SRIR-SRIR)	-0.37	0.13	1553.05	-2.790	0.0424
(SRIR-R) - (SRIR-SRIR)	0.01	0.12	1553.05	0.081	1.0000

Results are averaged over the levels of: Ldspkr_pair

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

Table 54

Model selection for plausibility difference ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
method+speaker+index	11	5273.23	0.00	0.37	-2625.53
method+speaker+height	9	5273.88	0.65	0.27	-2627.88
method+speaker	8	5273.98	0.75	0.26	-2628.94
index+method*speaker	15	5277.47	4.24	0.04	-2623.58
height+method*speaker	13	5278.17	4.94	0.03	-2625.97
method*speaker	12	5278.26	5.03	0.03	-2627.03
speaker	4	5303.27	30.04	0.00	-2647.62
method	7	5418.15	144.92	0.00	-2702.04
isdiff	4	5429.98	156.75	0.00	-2710.98
height	4	5443.76	170.53	0.00	-2717.87
null	3	5443.86	170.63	0.00	-2718.92

9 Speed

Table 57

Model selection for speed of walking

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
speaker+method+index+order	12	-8909.11	0.00	0.56	4466.61
speaker*method+index+order	18	-8908.61	0.50	0.44	4472.42
speaker+method+index	11	-8897.98	11.13	0.00	4460.03
speaker*method*order	26	-8870.09	39.02	0.00	4461.27
speaker+method+order	9	-8852.85	56.26	0.00	4435.45
speaker*method+order+height	16	-8850.26	58.85	0.00	4441.22
speaker+method+method_pair	15	-8843.46	65.65	0.00	4436.80
speaker+method	8	-8841.97	67.14	0.00	4429.01
speaker*method	14	-8841.35	67.76	0.00	4434.74
speaker*method+height	15	-8839.37	69.74	0.00	4434.76
method	5	-8835.37	73.74	0.00	4422.70
order	4	-8831.70	77.41	0.00	4419.86
speaker	6	-8827.60	81.51	0.00	4419.81
null	3	-8820.98	88.13	0.00	4413.50

Table 58

Pairwise comparison of loudspeaker position for speed of walking in walking phase

contrast	estimate	SE	df	t.ratio	p.value
A - B	-0.00	0.00	3122.03	-0.108	0.9995
A - C	-0.01	0.00	3122.03	-2.423	0.0730

A - D	-0.01	0.00	3122.03	-2.710	0.0342
B - C	-0.01	0.00	3122.03	-2.314	0.0951
B - D	-0.01	0.00	3122.03	-2.602	0.0460
C - D	-0.00	0.00	3122.03	-0.288	0.9917

Results are averaged over the levels of: Method, Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 59

Pairwise comparison of rendering methods for speed of walking in walking phase

contrast	estimate	SE	df	t.ratio	p.value
GA - R	-0.01	0.00	3122.06	-4.262	0.0001
GA - SRIR	-0.00	0.00	3122.03	-0.648	0.7933
R - SRIR	0.01	0.00	3122.05	3.517	0.0013

Results are averaged over the levels of: Ldspkr, Index_id, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 60

Pairwise comparison of trial index for speed of walking in walking phase

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	0.02	0.00	3122.04	7.326	<.0001
1 - 3	0.02	0.00	3122.03	5.428	<.0001
1 - 4	0.02	0.00	3122.03	6.041	<.0001
2 - 3	-0.01	0.00	3122.04	-1.900	0.2283
2 - 4	-0.00	0.00	3122.04	-1.243	0.5996
3 - 4	0.00	0.00	3122.03	0.652	0.9147

Results are averaged over the levels of: Method, Ldspkr, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 61

Pairwise comparison of playback order for speed of walking in walking phase

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	0.01	0.00	3122.03	3.624	0.0003

Results are averaged over the levels of: Method, Ldspkr, Index_id

Degrees-of-freedom method: kenward-roger

APPENDIX D

STATISTICAL ANALYSIS - WALKING AND STANDING PHASES

1 Plausibility

Table 62

Model selection for plausibility ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*method*speaker+order+height+method_pair	32	16463.66	0.00	0.71	-8199.61
phase*method*speaker+order+method_pair	31	16465.97	2.32	0.22	-8201.78
phase*method*speaker+order+height	28	16469.59	5.93	0.04	-8206.62
phase*method*speaker+order+index	30	16471.82	8.17	0.01	-8205.72
phase*method*speaker+order	27	16471.90	8.24	0.01	-8208.79
phase*method*speaker	26	16479.26	15.61	0.00	-8213.48
height+order+method*speaker	16	16483.08	19.43	0.00	-8225.48
index+order+method*speaker	18	16485.21	21.56	0.00	-8224.54
order+method*speaker	15	16485.22	21.57	0.00	-8227.56
phase*method*speaker+years_music	41	16487.82	24.17	0.00	-8202.55
phase*method*speaker+spatial_test	42	16489.80	26.14	0.00	-8202.52
phase*method*speaker+audio_test	42	16489.83	26.17	0.00	-8202.53
index+method*speaker	17	16492.44	28.78	0.00	-8229.16
method*speaker	14	16492.44	28.79	0.00	-8232.18
phase+method*speaker	15	16494.37	30.72	0.00	-8232.14
order*method*speaker	26	16498.82	35.16	0.00	-8223.26
method+speaker	8	16593.81	130.15	0.00	-8288.89
speaker	6	16807.73	344.07	0.00	-8397.86
method	5	16898.28	434.62	0.00	-8444.13
height	4	17096.86	633.21	0.00	-8544.43
null	3	17098.33	634.67	0.00	-8546.16

Table 63

Pairwise comparison results for plausibility ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value
Ldspkr = A, Method = GA					

1 - 2	-0.09	0.15	4802.55	-0.609	0.5425
Ldspkr = B, Method = GA					
1 - 2	-0.25	0.15	4802.55	-1.599	0.1099
Ldspkr = C, Method = GA					
1 - 2	0.03	0.15	4802.55	0.217	0.8286
Ldspkr = D, Method = GA					
1 - 2	-0.20	0.15	4802.55	-1.321	0.1867
Ldspkr = A, Method = R					
1 - 2	-0.01	0.11	4803.68	-0.111	0.9113
Ldspkr = B, Method = R					
1 - 2	0.21	0.11	4803.32	1.936	0.0529
Ldspkr = C, Method = R					
1 - 2	0.07	0.11	4803.70	0.645	0.5191
Ldspkr = D, Method = R					
1 - 2	0.37	0.11	4803.70	3.442	0.0006
Ldspkr = A, Method = SRIR					
1 - 2	-0.03	0.15	4802.56	-0.200	0.8413
Ldspkr = B, Method = SRIR					
1 - 2	-0.18	0.15	4802.55	-1.175	0.2401
Ldspkr = C, Method = SRIR					
1 - 2	0.24	0.15	4802.55	1.582	0.1137
Ldspkr = D, Method = SRIR					
1 - 2	-0.55	0.15	4802.20	-3.591	0.0003

Results are averaged over the levels of: Order

Degrees-of-freedom method: kenward-roger

Table 64

Pairwise comparison results for plausibility ratings

Ldspkr_pairwise	estimate	SE	df	t.ratio	p.value
Method = R, Phase = 1					
A - B	0.23	0.11	4801.40	2.116	0.1480
A - C	-0.03	0.11	4801.41	-0.263	0.9936
A - D	0.43	0.11	4801.41	3.944	0.0005
B - C	-0.26	0.11	4801.41	-2.383	0.0805
B - D	0.20	0.11	4801.41	1.825	0.2617
C - D	0.45	0.11	4801.40	4.215	0.0001
Method = GA, Phase = 1					
A - B	0.03	0.15	4801.40	0.163	0.9985
A - C	0.35	0.15	4801.40	2.250	0.1101
A - D	1.62	0.15	4801.40	10.494	<.0001
B - C	0.32	0.15	4801.40	2.091	0.1562
B - D	1.59	0.15	4801.40	10.349	<.0001

C - D	1.27	0.15	4801.40	8.258	<.0001
Method = SRIR, Phase = 1					
A - B	0.11	0.15	4801.40	0.715	0.8911
A - C	0.07	0.15	4801.43	0.474	0.9647
A - D	1.07	0.15	4801.43	6.946	<.0001
B - C	-0.04	0.15	4801.41	-0.238	0.9953
B - D	0.96	0.15	4801.41	6.224	<.0001
C - D	0.99	0.15	4801.40	6.440	<.0001
Method = R, Phase = 2					
A - B	0.45	0.11	4801.42	4.157	0.0002
A - C	0.05	0.11	4801.40	0.492	0.9608
A - D	0.81	0.11	4801.40	7.478	<.0001
B - C	-0.40	0.11	4801.42	-3.665	0.0014
B - D	0.36	0.11	4801.42	3.314	0.0051
C - D	0.76	0.11	4801.40	6.985	<.0001
Method = GA, Phase = 2					
A - B	-0.13	0.15	4801.40	-0.827	0.8418
A - C	0.47	0.15	4801.40	3.090	0.0108
A - D	1.51	0.15	4801.40	9.835	<.0001
B - C	0.60	0.15	4801.40	3.917	0.0005
B - D	1.63	0.15	4801.40	10.662	<.0001
C - D	1.03	0.15	4801.40	6.745	<.0001
Method = SRIR, Phase = 2					
A - B	-0.04	0.15	4801.40	-0.261	0.9938
A - C	0.35	0.15	4801.40	2.263	0.1069
A - D	0.54	0.15	4801.44	3.543	0.0023
B - C	0.39	0.15	4801.40	2.524	0.0564
B - D	0.58	0.15	4801.44	3.804	0.0008
C - D	0.20	0.15	4801.44	1.284	0.5734

Results are averaged over the levels of: Order, method_pair

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 65

Pairwise comparison results for plausibility ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1, Ldspkr = A					
R - GA	0.59	0.14	4801.44	4.160	0.0001
R - SRIR	0.38	0.14	4801.40	2.715	0.0183
GA - SRIR	-0.21	0.17	4801.43	-1.215	0.4445
Phase = 2, Ldspkr = A					
R - GA	0.50	0.14	4801.40	3.591	0.0010

R - SRIR	0.36	0.14	4801.40	2.568	0.0277
GA - SRIR	-0.14	0.17	4801.40	-0.842	0.6768
Phase = 1, Ldspkr = B					
R - GA	0.38	0.14	4801.42	2.722	0.0179
R - SRIR	0.26	0.14	4801.40	1.857	0.1515
GA - SRIR	-0.12	0.17	4801.41	-0.721	0.7513
Phase = 2, Ldspkr = B					
R - GA	-0.07	0.14	4801.41	-0.520	0.8616
R - SRIR	-0.13	0.14	4801.41	-0.925	0.6248
GA - SRIR	-0.06	0.17	4801.40	-0.334	0.9404
Phase = 1, Ldspkr = C					
R - GA	0.96	0.14	4801.44	6.857	<.0001
R - SRIR	0.48	0.14	4801.45	3.416	0.0019
GA - SRIR	-0.48	0.17	4801.40	-2.813	0.0137
Phase = 2, Ldspkr = C					
R - GA	0.92	0.14	4801.40	6.581	<.0001
R - SRIR	0.65	0.14	4801.40	4.656	<.0001
GA - SRIR	-0.27	0.17	4801.40	-1.585	0.2522
Phase = 1, Ldspkr = D					
R - GA	1.78	0.14	4801.44	12.673	<.0001
R - SRIR	1.02	0.14	4801.45	7.248	<.0001
GA - SRIR	-0.76	0.17	4801.40	-4.430	<.0001
Phase = 2, Ldspkr = D					
R - GA	1.20	0.14	4801.40	8.551	<.0001
R - SRIR	0.09	0.14	4801.44	0.670	0.7807
GA - SRIR	-1.11	0.17	4801.42	-6.480	<.0001

Results are averaged over the levels of: Order, Method_type

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 66

Pairwise comparison results of rendering method pair for plausibility ratings

contrast	estimate	SE	df	t.ratio	p.value
(R-R) - (GA-GA)	-0.28	0.10	4801.42	-2.726	0.0503
(R-R) - (GA-R)	0.01	0.07	4801.41	0.122	1.0000
(R-R) - (SRIR-R)	0.05	0.07	4801.41	0.736	0.9480
(R-R) - (SRIR-SRIR)	0.09	0.10	4801.41	0.841	0.9179
(GA-GA) - (GA-R)	0.29	0.08	4801.41	3.604	0.0029
(GA-GA) - (SRIR-R)	0.33	0.10	4801.41	3.218	0.0114
(GA-GA) - (SRIR-SRIR)	0.37	0.13	4801.40	2.852	0.0354
(GA-R) - (SRIR-R)	0.04	0.07	4801.41	0.619	0.9721
(GA-R) - (SRIR-SRIR)	0.08	0.10	4801.40	0.768	0.9399

(SRIR-R) - (SRIR-SRIR)	0.04	0.08	4801.41	0.477	0.9895

Results are averaged over the levels of: Order, Phase, Method, Ldspkr

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

2 Blur

Table 67

Model selection for blur ratings in walking phase

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*method*speaker+index+height+order	31	16284.89	0.00	0.49	-8111.24
phase*method*speaker+index+height	30	16285.09	0.20	0.45	-8112.35
phase*method*speaker+index+height+method_pair	34	16289.38	4.49	0.05	-8110.44
phase*method*speaker+index+order	30	16294.26	9.37	0.00	-8116.94
phase*method*speaker+index	29	16294.45	9.57	0.00	-8118.04
phase*method*speaker+height	27	16302.56	17.67	0.00	-8124.12
phase*method*speaker+order	27	16311.68	26.79	0.00	-8128.68
phase*method*speaker	26	16311.86	26.97	0.00	-8129.78
phase*method*speaker+audio_test	27	16313.16	28.27	0.00	-8129.42
phase*method*speaker+spatial_test	27	16313.83	28.94	0.00	-8129.76
index+method*speaker	17	16580.68	295.79	0.00	-8273.27
phase+method*speaker	15	16584.21	299.32	0.00	-8277.05
height+method*speaker	15	16588.99	304.10	0.00	-8279.44
order+method*speaker	15	16595.65	310.76	0.00	-8282.77
method*speaker	14	16595.68	310.79	0.00	-8283.79
order*method*speaker	26	16601.06	316.18	0.00	-8274.39
phase*method*speaker+years_music	41	16618.07	333.18	0.00	-8267.67
method+speaker	8	16706.59	421.71	0.00	-8345.28
speaker	6	16805.90	521.01	0.00	-8396.94
method	5	18002.61	1717.72	0.00	-8996.30
index	6	18054.40	1769.51	0.00	-9021.19
height	4	18072.99	1788.10	0.00	-9032.49
null	3	18076.88	1791.99	0.00	-9035.44

Table 68

Pairwise comparison results for blur ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1, Ldspkr = A					
R - GA	-0.27	0.13	4801.55	-2.075	0.0951
R - SRIR	-0.27	0.13	4801.53	-2.079	0.0943
GA - SRIR	0.00	0.15	4801.55	0.008	1.0000
Phase = 2, Ldspkr = A					
R - GA	0.08	0.13	4801.53	0.601	0.8196
R - SRIR	0.09	0.13	4801.53	0.665	0.7840
GA - SRIR	0.01	0.15	4801.53	0.055	0.9983
Phase = 1, Ldspkr = B					

R - GA	-0.10	0.13	4801.54	-0.800	0.7033
R - SRIR	-0.26	0.13	4801.53	-2.037	0.1035
GA - SRIR	-0.16	0.15	4801.54	-1.068	0.5339
Phase = 2, Ldspkr = B					
R - GA	0.60	0.13	4801.53	4.602	<.0001
R - SRIR	0.30	0.13	4801.53	2.313	0.0541
GA - SRIR	-0.30	0.15	4801.53	-1.984	0.1162
Phase = 1, Ldspkr = C					
R - GA	-0.82	0.13	4801.55	-6.333	<.0001
R - SRIR	-0.37	0.13	4801.57	-2.833	0.0128
GA - SRIR	0.46	0.15	4801.54	3.015	0.0073
Phase = 2, Ldspkr = C					
R - GA	-0.86	0.13	4801.53	-6.589	<.0001
R - SRIR	-0.96	0.13	4801.53	-7.382	<.0001
GA - SRIR	-0.10	0.15	4801.53	-0.687	0.7710
Phase = 1, Ldspkr = D					
R - GA	-1.37	0.13	4801.55	-10.487	<.0001
R - SRIR	-1.01	0.13	4801.57	-7.768	<.0001
GA - SRIR	0.35	0.15	4801.54	2.328	0.0521
Phase = 2, Ldspkr = D					
R - GA	-0.59	0.13	4801.53	-4.564	<.0001
R - SRIR	-0.08	0.13	4801.53	-0.612	0.8136
GA - SRIR	0.51	0.15	4801.53	3.425	0.0018

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 69

Pairwise comparison results for blur ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
Ldspkr = A, Method = R					
1 - 2	0.42	0.11	4803.49	3.936	0.0001
Ldspkr = B, Method = R					
1 - 2	-0.85	0.11	4803.49	-7.978	<.0001
Ldspkr = C, Method = R					
1 - 2	-0.72	0.11	4803.51	-6.760	<.0001
Ldspkr = D, Method = R					
1 - 2	-0.23	0.11	4803.51	-2.128	0.0334
Ldspkr = A, Method = GA					
1 - 2	0.77	0.15	4802.52	5.090	<.0001
Ldspkr = B, Method = GA					

1 - 2	-0.14	0.15	4802.52	-0.954	0.3403
Ldspkr = C, Method = GA					
1 - 2	-0.75	0.15	4802.52	-4.979	<.0001
Ldspkr = D, Method = GA					
1 - 2	0.55	0.15	4802.52	3.630	0.0003
Ldspkr = A, Method = SRIR					
1 - 2	0.77	0.15	4802.53	5.163	<.0001
Ldspkr = B, Method = SRIR					
1 - 2	-0.28	0.15	4802.52	-1.869	0.0616
Ldspkr = C, Method = SRIR					
1 - 2	-1.31	0.15	4802.52	-8.674	<.0001
Ldspkr = D, Method = SRIR					
1 - 2	0.71	0.15	4802.52	4.705	<.0001

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

Table 70

Pairwise comparison results for blur ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
Method = R, Phase = 1					
A - B	-0.25	0.11	4801.53	-2.386	0.0799
A - C	-0.07	0.11	4801.54	-0.634	0.9212
A - D	-1.47	0.11	4801.54	-13.883	<.0001
B - C	0.19	0.11	4801.54	1.756	0.2949
B - D	-1.22	0.11	4801.54	-11.492	<.0001
C - D	-1.40	0.11	4801.53	-13.269	<.0001
Method = GA, Phase = 1					
A - B	-0.09	0.15	4801.54	-0.570	0.9410
A - C	-0.62	0.15	4801.54	-4.114	0.0002
A - D	-2.56	0.15	4801.54	-16.987	<.0001
B - C	-0.53	0.15	4801.53	-3.550	0.0022
B - D	-2.48	0.15	4801.53	-16.445	<.0001
C - D	-1.94	0.15	4801.53	-12.896	<.0001
Method = SRIR, Phase = 1					
A - B	-0.25	0.15	4801.54	-1.654	0.3484
A - C	-0.17	0.15	4801.55	-1.110	0.6831
A - D	-2.21	0.15	4801.55	-14.713	<.0001
B - C	0.08	0.15	4801.54	0.537	0.9499
B - D	-1.97	0.15	4801.54	-13.046	<.0001
C - D	-2.05	0.15	4801.53	-13.539	<.0001
Method = R, Phase = 2					
A - B	-1.52	0.11	4801.53	-14.292	<.0001

A - C	-1.20	0.11	4801.53	-11.307	<.0001
A - D	-2.11	0.11	4801.53	-19.896	<.0001
B - C	0.32	0.11	4801.53	2.979	0.0154
B - D	-0.59	0.11	4801.53	-5.604	<.0001
C - D	-0.91	0.11	4801.53	-8.582	<.0001
Method = GA, Phase = 2					
A - B	-1.00	0.15	4801.53	-6.638	<.0001
A - C	-2.14	0.15	4801.53	-14.231	<.0001
A - D	-2.78	0.15	4801.53	-18.546	<.0001
B - C	-1.14	0.15	4801.53	-7.594	<.0001
B - D	-1.79	0.15	4801.53	-11.909	<.0001
C - D	-0.65	0.15	4801.53	-4.313	0.0001
Method = SRIR, Phase = 2					
A - B	-1.30	0.15	4801.53	-8.674	<.0001
A - C	-2.25	0.15	4801.53	-14.969	<.0001
A - D	-2.28	0.15	4801.53	-15.169	<.0001
B - C	-0.95	0.15	4801.53	-6.298	<.0001
B - D	-0.98	0.15	4801.53	-6.498	<.0001
C - D	-0.03	0.15	4801.53	-0.200	0.9971

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 71

Pairwise comparison for blur ratings

contrast	estimate	SE	df	t.ratio	p.value
Phase = 1, Ldspkr = A					
R - GA	-0.27	0.13	4801.55	-2.075	0.0951
R - SRIR	-0.27	0.13	4801.53	-2.079	0.0943
GA - SRIR	0.00	0.15	4801.55	0.008	1.0000
Phase = 2, Ldspkr = A					
R - GA	0.08	0.13	4801.53	0.601	0.8196
R - SRIR	0.09	0.13	4801.53	0.665	0.7840
GA - SRIR	0.01	0.15	4801.53	0.055	0.9983
Phase = 1, Ldspkr = B					
R - GA	-0.10	0.13	4801.54	-0.800	0.7033
R - SRIR	-0.26	0.13	4801.53	-2.037	0.1035
GA - SRIR	-0.16	0.15	4801.54	-1.068	0.5339
Phase = 2, Ldspkr = B					
R - GA	0.60	0.13	4801.53	4.602	<.0001
R - SRIR	0.30	0.13	4801.53	2.313	0.0541

GA - SRIR	-0.30	0.15	4801.53	-1.984	0.1162
Phase = 1, Ldspkr = C					
R - GA	-0.82	0.13	4801.55	-6.333	<.0001
R - SRIR	-0.37	0.13	4801.57	-2.833	0.0128
GA - SRIR	0.46	0.15	4801.54	3.015	0.0073
Phase = 2, Ldspkr = C					
R - GA	-0.86	0.13	4801.53	-6.589	<.0001
R - SRIR	-0.96	0.13	4801.53	-7.382	<.0001
GA - SRIR	-0.10	0.15	4801.53	-0.687	0.7710
Phase = 1, Ldspkr = D					
R - GA	-1.37	0.13	4801.55	-10.487	<.0001
R - SRIR	-1.01	0.13	4801.57	-7.768	<.0001
GA - SRIR	0.35	0.15	4801.54	2.328	0.0521
Phase = 2, Ldspkr = D					
R - GA	-0.59	0.13	4801.53	-4.564	<.0001
R - SRIR	-0.08	0.13	4801.53	-0.612	0.8136
GA - SRIR	0.51	0.15	4801.53	3.425	0.0018

Results are averaged over the levels of: Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 72

Pairwise comparison results of trial index for blur ratings

Index_id_pairwise	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	0.06	0.05	4801.53	1.044	0.7235
(1-12) - (25-36)	0.14	0.05	4801.53	2.683	0.0368
(1-12) - (37-48)	0.24	0.05	4801.53	4.531	<.0001
(13-24) - (25-36)	0.09	0.05	4801.53	1.645	0.3535
(13-24) - (37-48)	0.19	0.05	4801.53	3.500	0.0026
(25-36) - (37-48)	0.10	0.05	4801.53	1.871	0.2408

Results are averaged over the levels of: Phase, Method, Ldspkr

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

3 Externalization

Table 73

Model selection for externalization ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*speaker+method+order+height+method_pair	17	3580.40	0.00	0.59	-1773.14
phase*speaker+method+order+height	13	3582.19	1.79	0.24	-1778.06
phase*speaker+method*order+height	15	3583.09	2.69	0.15	-1776.49
phase*speaker+method+order	12	3587.68	7.28	0.02	-1781.81
phase*speaker+method+height	12	3634.80	54.40	0.00	-1805.37
phase*speaker+method+method_type	15	3638.48	58.08	0.00	-1804.19
phase*speaker+method	11	3640.12	59.72	0.00	-1809.03
phase*speaker+method+audio_test	12	3641.58	61.18	0.00	-1808.76
phase*speaker+method+spatial_audio_test	12	3642.10	61.70	0.00	-1809.02
phase*speaker+method+years_music	26	3647.93	67.53	0.00	-1797.82
phase*speaker+method+index	58	3690.97	110.57	0.00	-1786.76
method*speaker+phase	14	3724.74	144.34	0.00	-1848.32
method*speaker	13	3808.55	228.15	0.00	-1891.24
method+speaker	7	3833.72	253.32	0.00	-1909.85
speaker	5	3934.48	354.08	0.00	-1962.23
method	4	4012.08	431.68	0.00	-2002.03
phase	3	4032.14	451.74	0.00	-2013.07
null	2	4107.43	527.03	0.00	-2051.72

Table 74

Pairwise comparison for externalization ratings

Ldspkr_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1					
A - B	0.20	0.15	Inf	1.317	0.5519
A - C	-0.22	0.15	Inf	-1.437	0.4761
A - D	-1.04	0.17	Inf	-6.272	<.0001
B - C	-0.42	0.15	Inf	-2.746	0.0307
B - D	-1.24	0.17	Inf	-7.494	<.0001
C - D	-0.82	0.17	Inf	-4.901	<.0001
Phase = 2					
A - B	-1.91	0.17	Inf	-10.946	<.0001
A - C	-2.29	0.19	Inf	-12.174	<.0001
A - D	-2.03	0.18	Inf	-11.371	<.0001
B - C	-0.38	0.21	Inf	-1.837	0.2559
B - D	-0.12	0.20	Inf	-0.594	0.9341
C - D	0.26	0.21	Inf	1.249	0.5953

Results are averaged over the levels of: Method, Order

Results are given on the log odds ratio (not the response) scale.

P value adjustment: tukey method for comparing a family of 4 estimates

Table 75

Pairwise comparison for externalization ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value
Ldspkr = A					
1 - 2	0.45	0.15	Inf	3.005	0.0027
Ldspkr = B					
1 - 2	-1.66	0.18	Inf	-9.440	<.0001
Ldspkr = C					
1 - 2	-1.62	0.19	Inf	-8.479	<.0001
Ldspkr = D					
1 - 2	-0.54	0.19	Inf	-2.813	0.0049

Results are averaged over the levels of: Method, Order

Results are given on the log odds ratio (not the response) scale.

Table 76

Pairwise comparison for externalization ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
R - GA	0.97	0.10	Inf	9.359	<.0001
R - SRIR	0.84	0.10	Inf	8.158	<.0001
GA - SRIR	-0.12	0.11	Inf	-1.079	0.5274

Results are averaged over the levels of: Phase, Ldspkr, Order

Results are given on the log odds ratio (not the response) scale.

P value adjustment: tukey method for comparing a family of 3 estimates

4 Localization Error

Table 77

Model selection for localization error ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*method*speaker+order+index+height	31	-470.79	0.00	0.85	266.61
phase*method*speaker+order+index+height+method_pair	35	-467.24	3.55	0.14	268.90
phase*method*speaker+order+height	28	-461.32	9.48	0.01	258.83
phase*method*speaker+order+index	30	-451.87	18.92	0.00	256.14
phase*method*speaker+order	27	-442.45	28.34	0.00	248.39
phase*method*speaker+height	27	-435.63	35.17	0.00	244.98
phase*method*speaker+index	29	-426.28	44.51	0.00	242.33
phase*method*speaker	26	-416.95	53.84	0.00	234.63
phase*method*speaker+spatial_test	27	-415.44	55.35	0.00	234.89
phase*method*speaker+audio_test	27	-415.02	55.78	0.00	234.67
phase*speaker+method*speaker	18	-366.96	103.84	0.00	201.55
phase+method*speaker	15	-354.31	116.48	0.00	192.21
phase*method*speaker+years_music	41	-302.71	168.08	0.00	192.73
order*method*speaker	26	-268.74	202.06	0.00	160.52
order+method*speaker	15	-230.99	239.81	0.00	130.54
method*speaker	14	-206.81	263.98	0.00	117.45
method+speaker	8	-156.44	314.35	0.00	86.24
method	5	-113.61	357.19	0.00	61.81
speaker	6	-95.98	374.82	0.00	54.00
null	3	-54.54	416.25	0.00	30.27

Table 78

Pairwise comparison for localization error ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value
Ldspkr = A, Method = R					
1 - 2	0.10	0.02	4632.06	5.400	<.0001
Ldspkr = B, Method = R					
1 - 2	0.06	0.02	4631.84	3.078	0.0021
Ldspkr = C, Method = R					
1 - 2	0.08	0.02	4631.40	4.496	<.0001
Ldspkr = D, Method = R					
1 - 2	-0.06	0.02	4631.85	-2.971	0.0030
Ldspkr = A, Method = GA					
1 - 2	0.08	0.03	4630.57	3.191	0.0014
Ldspkr = B, Method = GA					
1 - 2	0.17	0.03	4630.68	6.314	<.0001
Ldspkr = C, Method = GA					

1 - 2	0.14	0.03	4630.63	5.158	<.0001
Ldspkr = D, Method = GA					
1 - 2	0.09	0.03	4630.46	3.167	0.0016
Ldspkr = A, Method = SRIR					
1 - 2	0.10	0.03	4630.30	3.882	0.0001
Ldspkr = B, Method = SRIR					
1 - 2	0.14	0.03	4630.61	5.217	<.0001
Ldspkr = C, Method = SRIR					
1 - 2	0.05	0.03	4630.87	1.835	0.0666
Ldspkr = D, Method = SRIR					
1 - 2	0.16	0.03	4630.87	5.921	<.0001

Results are averaged over the levels of: Order, Index_id
Degrees-of-freedom method: kenward-roger

Table 79

Pairwise comparison for localization error ratings

Ldspkr_pairwise	estimate	SE	df	t.ratio	p.value
Method = R, Phase = 1					
A - B	-0.01	0.02	4629.48	-0.714	0.8917
A - C	0.04	0.02	4629.49	2.212	0.1202
A - D	0.07	0.02	4629.49	3.947	0.0005
B - C	0.06	0.02	4629.49	2.929	0.0180
B - D	0.09	0.02	4629.54	4.674	<.0001
C - D	0.03	0.02	4629.50	1.715	0.3159
Method = GA, Phase = 1					
A - B	-0.04	0.03	4629.46	-1.546	0.4102
A - C	-0.04	0.03	4629.45	-1.595	0.3818
A - D	-0.13	0.03	4629.50	-5.021	<.0001
B - C	-0.00	0.03	4629.49	-0.063	0.9999
B - D	-0.09	0.03	4629.54	-3.486	0.0028
C - D	-0.09	0.03	4629.49	-3.394	0.0039
Method = SRIR, Phase = 1					
A - B	0.02	0.03	4629.52	0.772	0.8671
A - C	0.06	0.03	4629.56	2.062	0.1658
A - D	-0.07	0.03	4629.50	-2.603	0.0457
B - C	0.04	0.03	4629.47	1.320	0.5502
B - D	-0.09	0.03	4629.47	-3.430	0.0034
C - D	-0.13	0.03	4629.48	-4.709	<.0001
Method = R, Phase = 2					
A - B	-0.06	0.02	4629.50	-2.980	0.0153
A - C	0.02	0.02	4629.44	1.331	0.5428
A - D	-0.08	0.02	4629.46	-4.436	0.0001

B - C	0.08	0.02	4629.48	4.297	0.0001
B - D	-0.03	0.02	4629.48	-1.396	0.5021
C - D	-0.11	0.02	4629.45	-5.773	<.0001
Method = GA, Phase = 2					
A - B	0.04	0.03	4629.46	1.596	0.3807
A - C	0.01	0.03	4629.44	0.416	0.9758
A - D	-0.13	0.03	4629.50	-4.965	<.0001
B - C	-0.03	0.03	4629.47	-1.177	0.6417
B - D	-0.18	0.03	4629.49	-6.517	<.0001
C - D	-0.14	0.03	4629.52	-5.356	<.0001
Method = SRIR, Phase = 2					
A - B	0.05	0.03	4629.44	2.078	0.1604
A - C	0.00	0.03	4629.44	0.032	1.0000
A - D	-0.02	0.03	4629.47	-0.584	0.9369
B - C	-0.05	0.03	4629.44	-2.043	0.1724
B - D	-0.07	0.03	4629.47	-2.632	0.0423
C - D	-0.02	0.03	4629.48	-0.615	0.9274

Results are averaged over the levels of: Order, Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

Table 80

Pairwise comparison for localization error ratings

method_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1, Ldspkr = A					
R - GA	-0.01	0.02	4629.48	-0.575	0.8337
R - SRIR	-0.06	0.02	4629.60	-2.497	0.0336
GA - SRIR	-0.05	0.03	4629.53	-1.691	0.2088
Phase = 2, Ldspkr = A					
R - GA	-0.03	0.02	4629.44	-1.323	0.3825
R - SRIR	-0.06	0.02	4629.44	-2.475	0.0357
GA - SRIR	-0.03	0.03	4629.44	-0.992	0.5821
Phase = 1, Ldspkr = B					
R - GA	-0.04	0.02	4629.58	-1.772	0.1792
R - SRIR	-0.02	0.02	4629.51	-1.071	0.5324
GA - SRIR	0.02	0.03	4629.46	0.612	0.8137
Phase = 2, Ldspkr = B					
R - GA	0.07	0.02	4629.47	2.948	0.0090
R - SRIR	0.05	0.02	4629.48	2.369	0.0470
GA - SRIR	-0.01	0.03	4629.46	-0.517	0.8630
Phase = 1, Ldspkr = C					
R - GA	-0.10	0.02	4629.48	-4.188	0.0001

R - SRIR	-0.04	0.02	4629.53	-1.918	0.1337
GA - SRIR	0.05	0.03	4629.48	1.969	0.1200
Phase = 2, Ldspkr = C					
R - GA	-0.04	0.02	4629.45	-1.919	0.1334
R - SRIR	-0.08	0.02	4629.44	-3.524	0.0012
GA - SRIR	-0.04	0.03	4629.44	-1.371	0.3560
Phase = 1, Ldspkr = D					
R - GA	-0.22	0.02	4629.55	-9.569	<.0001
R - SRIR	-0.20	0.02	4629.52	-8.794	<.0001
GA - SRIR	0.02	0.03	4629.46	0.664	0.7844
Phase = 2, Ldspkr = D					
R - GA	-0.08	0.02	4629.52	-3.440	0.0017
R - SRIR	0.01	0.02	4629.49	0.475	0.8831
GA - SRIR	0.09	0.03	4629.52	3.380	0.0021

Results are averaged over the levels of: Order, Index_id

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Table 81

Pairwise comparison for localization error ratings

Index_id_pairwise	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	0.03	0.01	4629.56	3.490	0.0027
(1-12) - (25-36)	0.03	0.01	4629.49	3.256	0.0062
(1-12) - (37-48)	0.03	0.01	4629.58	2.704	0.0347
(13-24) - (25-36)	-0.00	0.01	4629.55	-0.232	0.9956
(13-24) - (37-48)	-0.01	0.01	4629.49	-0.770	0.8682
(25-36) - (37-48)	-0.01	0.01	4629.59	-0.542	0.9488

Results are averaged over the levels of: Method, Ldspkr, Phase, Order

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 4 estimates

5 Loudspeaker Recognition Rate

Table 82

Model selection for loudspeaker recognition rate

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*speaker+method+order	12	3206.27	0.00	0.44	-1591.10
phase*speaker+order	10	3207.09	0.81	0.29	-1593.52
phase*speaker+method+order+method_type	17	3208.81	2.54	0.12	-1587.34
phase*speaker+method*order	14	3209.33	3.06	0.10	-1590.62
phase*speaker+order+method_type	14	3210.56	4.29	0.05	-1591.24
phase*speaker+method	11	3221.51	15.24	0.00	-1599.73
phase*speaker+method+spatial_test	12	3221.75	15.48	0.00	-1598.84
phase*speaker+method+height	12	3222.00	15.73	0.00	-1598.97
phase*speaker	9	3222.31	16.04	0.00	-1602.14
phase*speaker+method+years_music	26	3231.87	25.60	0.00	-1589.79
phase*speaker+method+index	58	3244.63	38.36	0.00	-1563.60
phase+method*speaker	14	3266.27	60.00	0.00	-1619.09
order+method*speaker	14	3273.27	67.00	0.00	-1622.59
method*speaker	13	3288.38	82.11	0.00	-1631.15
phase*method+speaker	10	3321.65	115.38	0.00	-1650.80
method+speaker	7	3362.17	155.90	0.00	-1674.08
speaker	5	3362.94	156.67	0.00	-1676.46
method	4	3563.09	356.82	0.00	-1777.54
null	2	3563.55	357.28	0.00	-1779.78

Table 83

Pairwise comparison for loudspeaker recognition rate

Phase_pairwise	estimate	SE	df	t.ratio	p.value
S = A					
1 - 2	-2.39	0.40	Inf	-5.920	<.0001
S = B					
1 - 2	0.47	0.17	Inf	2.735	0.0062
S = C					
1 - 2	0.13	0.21	Inf	0.641	0.5215
S = D					
1 - 2	1.36	0.15	Inf	8.801	<.0001

Results are averaged over the levels of: Order

Results are given on the log odds ratio (not the response) scale.

Table 84

Pairwise comparison for loudspeaker recognition rate

Ldspkr_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1					
A - B	0.12	0.19	Inf	0.653	0.9144
A - C	-0.31	0.20	Inf	-1.541	0.4127
A - D	0.27	0.18	Inf	1.477	0.4516
B - C	-0.44	0.20	Inf	-2.183	0.1278
B - D	0.15	0.18	Inf	0.827	0.8416
C - D	0.59	0.20	Inf	2.984	0.0151
Phase = 2					
A - B	2.98	0.40	Inf	7.523	<.0001
A - C	2.21	0.41	Inf	5.432	<.0001
A - D	4.02	0.39	Inf	10.272	<.0001
B - C	-0.77	0.18	Inf	-4.213	0.0001
B - D	1.04	0.15	Inf	7.149	<.0001
C - D	1.81	0.17	Inf	10.527	<.0001

Results are averaged over the levels of: Order

Results are given on the log odds ratio (not the response) scale.

P value adjustment: tukey method for comparing a family of 4 estimates

6 Timbre Difference

Table 85

Model selection for timbre difference ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*method*speaker+height+index	26	8148.13	0.00	0.41	-4047.77
phase*method*speaker+height	23	8148.57	0.45	0.33	-4051.05
phase*speaker+height+method+index	14	8150.14	2.01	0.15	-4060.98
phase*speaker+height+method	11	8150.95	2.82	0.10	-4064.42
phase*method*speaker+index	25	8157.52	9.39	0.00	-4053.49
phase*method*speaker	22	8157.93	9.81	0.00	-4056.75
phase*speaker+method	10	8160.17	12.04	0.00	-4070.04
phase*method*speaker+audio_test	38	8170.68	22.56	0.00	-4046.71
phase*method*speaker+spatial_test	38	8172.99	24.86	0.00	-4047.87
phase*method*speaker+years_music	37	8173.45	25.33	0.00	-4049.13
phase+method*speaker	13	8205.97	57.84	0.00	-4089.91
height+method*speaker	13	8448.04	299.91	0.00	-4210.94
method*speaker	12	8450.12	302.00	0.00	-4213.00
index+method*speaker	15	8450.25	302.13	0.00	-4210.03
method+speaker	8	8456.62	308.49	0.00	-4220.28
method	7	8544.66	396.53	0.00	-4265.30
speaker	4	8558.95	410.83	0.00	-4275.47
height	4	8641.29	493.16	0.00	-4316.64
null	3	8642.98	494.86	0.00	-4318.49

Table 86

Pairwise comparison for timbre difference ratings

Ldspkr_pair_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1					
(A-C) - (B-D)	-0.94	0.07	2382.24	-12.551	<.0001
Phase = 2					
(A-C) - (B-D)	-0.14	0.07	2382.24	-1.902	0.0573

Results are averaged over the levels of: Method_pair

Degrees-of-freedom method: kenward-roger

Table 87

Pairwise comparison for timbre difference ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value
Ldspkr_pair = A-C					
1 - 2	-1.27	0.07	2386.48	-16.973	<.0001

S_type = B-D					
1 - 2	-0.47	0.07	2386.48	-6.338	<.0001

Results are averaged over the levels of: Method_pair

Degrees-of-freedom method: kenward-roger

Table 88

Pairwise comparison for timbre difference ratings

contrast	estimate	SE	df	t.ratio	p.value
(R-R) - (GA-GA)	-0.55	0.11	2382.24	-4.922	<.0001
(R-R) - (GA-R)	-0.81	0.08	2382.24	-10.315	<.0001
(R-R) - (SRIR-R)	-0.48	0.08	2382.24	-6.122	<.0001
(R-R) - (SRIR-SRIR)	-0.08	0.11	2382.24	-0.761	0.9417
(GA-GA) - (GA-R)	-0.26	0.10	2382.24	-2.598	0.0710
(GA-GA) - (SRIR-R)	0.07	0.10	2382.24	0.649	0.9668
(GA-GA) - (SRIR-SRIR)	0.46	0.13	2382.24	3.604	0.0030
(GA-R) - (SRIR-R)	0.33	0.06	2382.24	5.135	<.0001
(GA-R) - (SRIR-SRIR)	0.73	0.10	2382.24	7.156	<.0001
(SRIR-R) - (SRIR-SRIR)	0.40	0.10	2382.24	3.909	0.0009

Results are averaged over the levels of: Ldspkr_pair, Phase

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

7 Reverberation Difference

Table 89

Model selection for reverberation difference ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*method*speaker	22	8299.64	0.00	0.47	-4127.61
phase*speaker+method	10	8301.42	1.78	0.19	-4140.67
phase*method*speaker+height	23	8301.62	1.97	0.18	-4127.58
phase*speaker+height+method	11	8303.38	3.73	0.07	-4140.63
phase*method*speaker+index	25	8304.33	4.68	0.05	-4126.89
phase*speaker+method+index	13	8306.17	6.53	0.02	-4140.01
phase*method*speaker+height+index	26	8306.30	6.66	0.02	-4126.86
phase*speaker+height+method+index	14	8308.13	8.49	0.01	-4139.98
phase*method*speaker+audio_test	38	8321.31	21.66	0.00	-4122.02
phase*method*speaker+years_music	37	8322.09	22.45	0.00	-4123.45
phase*method*speaker+spatial_test	38	8324.03	24.39	0.00	-4123.39
phase+method*speaker	13	8431.39	131.74	0.00	-4202.62
method*speaker	12	8520.83	221.18	0.00	-4248.35
height+method*speaker	13	8522.07	222.43	0.00	-4247.96
index+method*speaker	15	8525.77	226.12	0.00	-4247.78
method+speaker	8	8529.36	229.72	0.00	-4256.65
speaker	4	8577.13	277.48	0.00	-4284.55
method	7	8809.25	509.60	0.00	-4397.60
null	3	8850.83	551.19	0.00	-4422.41
height	4	8852.27	552.62	0.00	-4422.12

Table 90

Pairwise comparison for reverberation difference ratings

Ldspkr_pair_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1					
(A-C) - (B-D)	-1.57	0.09	2394.55	-17.101	<.0001
Phase = 2					
(A-C) - (B-D)	-0.27	0.09	2394.55	-2.977	0.0029

Results are averaged over the levels of: method_pair

Degrees-of-freedom method: kenward-roger

Table 91

Pairwise comparison for reverberation difference ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value

Ldspkr_pair = A-C					
1 - 2	-1.16	0.09	2397.61	-12.634	<.0001
Ldspkr_pair = B-D					
1 - 2	0.14	0.09	2397.61	1.476	0.1401

Results are averaged over the levels of: method_pair

Degrees-of-freedom method: kenward-roger

Table 92

Pairwise comparison for reverberation difference ratings

method_pair_pairwise	estimate	SE	df	t.ratio	p.value
(R-R) - (GA-GA)	-0.51	0.12	2394.55	-4.433	0.0001
(R-R) - (GA-R)	-0.60	0.08	2394.55	-7.360	<.0001
(R-R) - (SRIR-R)	-0.43	0.08	2394.55	-5.270	<.0001
(R-R) - (SRIR-SRIR)	-0.16	0.12	2394.55	-1.391	0.6338
(GA-GA) - (GA-R)	-0.09	0.11	2394.55	-0.845	0.9165
(GA-GA) - (SRIR-R)	0.08	0.11	2394.55	0.774	0.9382
(GA-GA) - (SRIR-SRIR)	0.35	0.13	2394.55	2.634	0.0646
(GA-R) - (SRIR-R)	0.17	0.07	2394.55	2.559	0.0784
(GA-R) - (SRIR-SRIR)	0.44	0.11	2394.55	4.177	0.0003
(SRIR-R) - (SRIR-SRIR)	0.27	0.11	2394.55	2.559	0.0785

Results are averaged over the levels of: Ldspkr_pair, Phase

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 5 estimates

8 Plausibility Difference

Table 93

Model selection for plausibility difference ratings

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
phase*speaker+height+method+index	14	7985.97	0.00	0.42	-3978.90
phase*speaker+method+index	13	7986.36	0.39	0.34	-3980.10
phase*speaker+height+method	11	7988.39	2.42	0.13	-3983.14
phase*speaker+method	10	7988.79	2.82	0.10	-3984.35
phase*method*speaker+height+index	26	7995.47	9.50	0.00	-3971.44
phase*method*speaker+index	25	7995.83	9.87	0.00	-3972.64
phase*method*speaker+height	23	7997.80	11.83	0.00	-3975.67
phase*method*speaker	22	7998.18	12.21	0.00	-3976.88
phase*method*speaker+years_music	37	8003.70	17.74	0.00	-3964.26
phase*method*speaker+audio_test	38	8004.82	18.85	0.00	-3963.78
phase*method*speaker+spatial_test	38	8005.53	19.56	0.00	-3964.13
index+method*speaker	15	8024.87	38.90	0.00	-3997.33
method+speaker	8	8026.57	40.60	0.00	-4005.25
height+method*speaker	13	8027.41	41.44	0.00	-4000.63
method*speaker	12	8027.72	41.75	0.00	-4001.79
phase+method*speaker	13	8029.61	43.65	0.00	-4001.73
method	7	8072.61	86.65	0.00	-4029.28
speaker	4	8086.63	100.66	0.00	-4039.31
height	4	8131.13	145.17	0.00	-4061.56
null	3	8131.34	145.37	0.00	-4062.67

Table 94

Pairwise comparison for plausibility difference ratings

Ldspkr_pair_pairwise	estimate	SE	df	t.ratio	p.value
Phase = 1					
(A-C) - (B-D)	-0.69	0.07	2384.04	-9.476	<.0001
Phase = 2					
(A-C) - (B-D)	-0.03	0.07	2384.04	-0.364	0.7158

Results are averaged over the levels of: method_pair, Index_id

Degrees-of-freedom method: kenward-roger

Table 95

Pairwise comparison for plausibility difference ratings

Phase_pairwise	estimate	SE	df	t.ratio	p.value

Ldspkr_pair = A-C					
1 - 2	-0.31	0.07	2384.04	-4.299	<.0001
Ldspkr_pair = B-D					
1 - 2	0.35	0.07	2384.04	4.814	<.0001

Results are averaged over the levels of: method_pair, Index_id
Degrees-of-freedom method: kenward-roger

Table 96

Pairwise comparison for plausibility difference ratings

method_pair_pairwise	estimate	SE	df	t.ratio	p.value
(R-R) - (GA-GA)	-0.55	0.11	2384.04	-4.981	<.0001
(R-R) - (GA-R)	-0.62	0.08	2384.04	-8.077	<.0001
(R-R) - (SRIR-R)	-0.41	0.08	2384.04	-5.386	<.0001
(R-R) - (SRIR-SRIR)	-0.32	0.11	2384.04	-2.931	0.0034
(GA-GA) - (GA-R)	-0.07	0.10	2384.04	-0.654	0.5130
(GA-GA) - (SRIR-R)	0.14	0.10	2384.04	1.368	0.1714
(GA-GA) - (SRIR-SRIR)	0.23	0.13	2384.04	1.825	0.0681
(GA-R) - (SRIR-R)	0.21	0.06	2384.04	3.288	0.0010
(GA-R) - (SRIR-SRIR)	0.30	0.10	2384.04	2.906	0.0037
(SRIR-R) - (SRIR-SRIR)	0.09	0.10	2384.04	0.880	0.3790

Results are averaged over the levels of: Ldspkr_pair, Phase, Index_id
Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 5 estimates

9 Yaw Movement - Standing Phase

Table 97

Model selection for amplitude of yaw movement

Model	K	AICc	Delta AICc	AICc weight	log-Likelihood
speaker+method+index+order	12	6363.40	0.00	0.66	-3169.63
speaker+method+index	11	6364.84	1.45	0.32	-3171.37
method+index+order	9	6372.31	8.92	0.01	-3177.12
speaker*method+index	17	6373.24	9.85	0.00	-3169.49
method+index	8	6373.73	10.33	0.00	-3178.83
index	6	6373.74	10.35	0.00	-3180.85
speaker+index+order	7	6392.34	28.94	0.00	-3189.15
speaker	6	6393.73	30.33	0.00	-3190.85
speaker+method	8	6395.63	32.23	0.00	-3189.78
speaker+method+method_type	12	6398.70	35.30	0.00	-3187.28
order	4	6401.10	37.70	0.00	-3196.54
speaker*method+order	15	6402.36	38.96	0.00	-3186.08
null	3	6402.45	39.06	0.00	-3198.22
speaker*method	14	6403.75	40.35	0.00	-3187.79
method	5	6404.35	40.96	0.00	-3197.16
speaker*method+order+height	16	6404.38	40.98	0.00	-3186.07
speaker*method+height	15	6405.77	42.37	0.00	-3187.78

Table 98

Pairwise comparison of trial index for amplitude of yaw movement

Index_id_pairwise	estimate	SE	df	t.ratio	p.value
(1-12) - (13-24)	-0.09	0.05	2379.03	-1.729	0.3088
(1-12) - (25-36)	-0.09	0.05	2379.03	-1.743	0.3016
(1-12) - (37-48)	-0.30	0.05	2379.03	-5.841	<.0001
(13-24) - (25-36)	-0.00	0.05	2379.03	-0.019	1.0000
(13-24) - (37-48)	-0.21	0.05	2379.03	-4.143	0.0002
(25-36) - (37-48)	-0.21	0.05	2379.03	-4.146	0.0002

Results are averaged over the levels of: Method, Ldspkr

Degrees-of-freedom method: kenward-roger

Results are given on the log (not the response) scale.

P value adjustment: tukey method for comparing a family of 4 estimates

Table 99

Pairwise comparison of rendering method for amplitude of yaw movement

method_pairwise	estimate	SE	df	t.ratio	p.value

R - GA	-0.09	0.04	2379.03	-2.008	0.0447
R - SRIR	-0.02	0.04	2379.03	-0.552	0.5813
GA - SRIR	0.07	0.05	2379.03	1.273	0.2031

Results are averaged over the levels of: Ldspkr, Index_id

Degrees-of-freedom method: kenward-roger

Results are given on the log (not the response) scale.

Table 100

Pairwise comparison of loudspeaker position for amplitude of yaw movement

Ldspkr_pairwise	estimate	SE	df	t.ratio	p.value
A - B	0.15	0.05	2379.03	2.907	0.0193
A - C	0.18	0.05	2379.03	3.595	0.0019
A - D	0.14	0.05	2379.03	2.662	0.0391
B - C	0.04	0.05	2379.03	0.688	0.9018
B - D	-0.01	0.05	2379.03	-0.245	0.9948
C - D	-0.05	0.05	2379.03	-0.933	0.7871

Results are averaged over the levels of: Method, Index_id

Degrees-of-freedom method: kenward-roger

Results are given on the log (not the response) scale.

P value adjustment: tukey method for comparing a family of 4 estimates