Auditory distance perception of static sources in the context of Audio-only Augmented Reality : an investigation of acoustic and non-acoustic cues

Vincent Martin

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité Sciences Cognitives
ED158 - Ecole doctorale Cerveau Cognition Comportement
Sciences et Technologie de la Musique et du Son (UMR 9912)
Institut de Recherche et de Coordination Acoustique Musique
Equipe Espace Acoustiques et Cognitifs.

AUDITORY DISTANCE PERCEPTION OF STATIC SOURCES IN AUDIO-ONLY AUGMENTED REALITY

AN INVESTIGATION OF ACOUSTIC AND NON-ACOUSTIC CUES

VINCENT MARTIN
DIRIGÉ PAR : ISABELLE VIAUD-DELMON & OLIVIER WARUSFEL

FÉVRIER 2022

ABSTRACT

This thesis aims to investigate a variety of effects linking the auditory distance perception of virtual sound sources to the context of audio-only augmented reality (AAR) applications. It focuses on the ways in which its specific perceptual context and primary objectives impose constraints on the design of the distance rendering approach used to generate virtual sound sources for AAR applications.

AAR is a term that refers to a set of technologies that aim to merge computer-generated auditory content into a user's actual acoustic environment. AAR systems have fundamental requirements that distinguish them from conventional human-computer interfaces: the user must have perception of his environment through all sensory modalities, and an audio playback system must enable a seamless integration of virtual sound events within the user's environment. Different challenges arise as a result of these critical requirements.

Among the different technical challenges, one is to process sound using an artificial spatialization technique in order to monitor the apparent position of virtual sound sources and reproduce the acoustic properties of the room. It is well established that the procedure chosen has a direct effect on the reproduction of the room effect's properties.

The precision required for acoustic cue reproduction is closely related to the human auditory system's ability to infer a distance percept from sound signals.

The first part of the thesis' motivations concerns the critical role of acoustic cue reproduction in the auditory distance perception of virtual sound sources in the context of audio-only augmented reality. In comparison to other dimensions, auditory distance perception is understudied. It is based on a range of cues categorised as objective/acoustic, and cognitive/subjective. We examined which distinct strategies for weighting auditory cues are used by the auditory system to create the perception of sound distance. By considering different spatial and temporal segmentations, we attempted to characterize how early energy is perceived in relation to reverberation.

The second part of the thesis's motivations focuses on how, in AAR applications, environmental-related cues could impact the perception of virtual sound sources. In AAR applications, the geometry of the environment is not always completely considered. In particular, the calibration effect induced by the perception of the visual environment on the auditory perception is generally overlooked. We also became interested in the particular instance in which co-occurring real sound sources whose placements are unknown to the user could affect the auditory distance perception of virtual sound sources through an intra-modal calibration effect.

The study of these effects was done through the development of different perceptual experiments, which were mainly done with remote participants recruited online.

Overall, we revealed advantages of early-to-late energy perception over the direct-to-reverberant sound energy ratio as an auditory distance reproduction criterion. Additionally, we researched temporal criteria defining how the auditory system perceives this early energy effectively to infer a distance percept. Additionally, we exhibited that the auditory distance perception of similar sound sources might vary significantly based on the listening position in the room, the shape of the room, and the presence of ambient sounds. This collection of results aims to demonstrate what objective requirements a sound signal generated by a distance rendering approach must meet in comparison to real sound sources and how it should adjust based on the a priori knowledge provided to the user regarding the environment.

RÉSUMÉ

La RAA est un terme qui désigne un ensemble de technologies visant à fusionner un contenu auditif généré par ordinateur dans l'environnement acoustique réel d'un utilisateur. Les systèmes de RAA ont des exigences fondamentales qui les distinguent des interfaces homme-machine conventionnelles : l'utilisateur doit avoir une perception de son environnement à travers toutes les modalités sensorielles et un système de reproduction audio doit permettre une intégration transparente des événements sonores virtuels dans l'environnement de l'utilisateur. Différents enjeux découlent de ces exigences critiques.

Parmi les différents défis techniques, l'un d'entre eux consiste à appliquer au son différent traitements de spatialisation afin de contrôler la position apparente des sources sonores virtuelles et reproduire les propriétés acoustiques de la salle. La précision requise pour la reproduction des indices acoustiques est étroitement liée à la capacité du système auditif humain d'inférer un percept de distance à partir de signaux sonores.

La première partie des objectifs de la thèse se concentre sur le rôle critique de la reproduction des indices acoustiques dans la perception de la distance auditive des sources sonores virtuelles dans le contexte de la réalité augmentée auditive. En comparaison avec d'autres dimensions, la perception auditive de la distance a peu été étudiée. Elle est basée sur une série d'indices catégorisés comme objectifs/acoustiques et cognitifs/subjectifs. Nous avons examiné quelles stratégies distinctes de pondération des indices acoustiques sont utilisées par le système auditif pour créer la perception de la distance sonore. Plus particulièrement, nous avons tenté de caractériser comment l'énergie précoce est perçue en relation avec la réverbération en considérant différentes segmentations spatiales et temporelles.

La seconde partie des objectifs de la thèse se concentre sur la façon dont, dans les applications de RAA, les indices liés à l'environnement peuvent avoir un impact sur la perception des sources sonores virtuelles. Dans les applications de RAA, la géométrie de l'environnement n'est pas toujours complètement prise en compte. Plus particulièrement, l'effet de calibration induit par la perception de l'environnement visuel sur la perception auditive est généralement négligé. Nous nous sommes également intéressés au cas particulier dans lequel des sources sonores réelles co-occurrentes dont l'emplacement est inconnu de l'utilisateur peuvent affecter la perception auditive de la distance des sources sonores virtuelles par un effet de calibration intra-modale.

L'étude de ces effets s'est faite par le développement de différentes expériences perceptives qui ont été réalisées principalement avec des participants à distance, recrutés en ligne.

Au cours de cette thèse, nous avons révélé les avantages significatifs du rapport entre énergie précoce et énergie tardive en comparaison du rapport entre l'énergie

du son direct et celle du son réverbéré comme critère de perception de la distance auditive. De plus, nous avons étudié les critères temporels par lesquels le système auditif perçoit efficacement cette énergie précoce pour en déduire un percept de distance. Enfin, nous avons montré que la perception de la distance auditive de sources sonores similaires pouvait varier de manière significative en fonction de la position d'écoute dans la pièce par rapport aux parois, de la forme de la pièce et de la présence de sons ambiants. Cet ensemble de résultats vise à démontrer quelles exigences objectives un signal sonore généré par une approche de rendu de distance doit satisfaire par rapport à des sources sonores réelles et comment il doit s'ajuster en fonction des informations a priori fournies à l'utilisateur par l'environnement.

ACKNOWLEDGMENTS

Je tiens tout d'abord à témoigner toute ma gratitude envers Norbert Kopčo et Nicolas Grimault pour avoir accepté d'être rapporteurs de cette thèse, sans oublier Etienne Hendrickx et Quentin Grimal, pour leur rôle d'examinateurs.

Je tiens à remercier Olivier Warusfel et Isabelle Viaud-delmon qui ont encadré ce travail de thèse. Leurs expertises scientifiques complémentaires, leurs conseils, leurs nombreuses corrections, ainsi que leur bienveillance ont été d'un secours sans lequel ce manuscrit n'aurait pas pu exister. Je leur suis reconnaissant pour ces 3 années (et demi) qui se sont déroulées dans d'excellentes conditions, malgré le contexte particulier de la crise sanitaire.

Un grand merci également à toute l'équipe Espace Acoustiques et Cognitifs de l'IRCAM, à commencer par Pierre Massé et Thibault Carpentier pour m'avoir apporté une aide précieuse dans la compréhension des outils de spatialisation sonores, et tout cela avec une patience à toute épreuve. Je remercie également Lise Hobeika et Marine Taffout pour m'avoir éclairé sur les grands mystères de l'ANOVA et leur aide pour les statistiques en général.

Je remercie aussi mes amis qui se sont tous soumis sans discuter à l'exercice des pré-tests à maintes reprises. J'ai une pensée toute particulière pour Hadrien, Mathieu, Charles, Virgile, Lou, Andrea et Tristan avec qui j'ai passé toutes ces années à l'IRCAM du master à la thèse. Un grand merci à Yacine, Merwan, Mehdi, Antoine et Lila pour leur aide sans faille à travers ces années, dans les bons comme les mauvais moments.

Enfin, bien évidemment, je souhaite remercier mes parents pour leur soutien et leur confiance permanente.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AR	Augmented Reality
AAR	Audio-only Augmented Reality
VR	Virtual Reality
HRTF	Head-Related Transfer Function
HRIR	Head-Related Impulse Response
RIR	Room Impulse Response
BRIR	Binaural Room Impulse Response
SRIR	Spatial Room Impulse Response
ITD	Inter-aural Time Difference
ILD	Inter-aural Level Difference
DRR	Direct-to-Reverberant energy Ratio
FDN	Feedback Delay Network
HOA	Higher Order Ambisonics
SHD	Spherical Harmonics Domain
SMA	Spherical Microphone Array
EDR	Energy Decay Relief
ASW	Apparent Source Width
PWD	Plane-Wave Decomposition
VAS	Visual Analogue Scale
JS	Javascript
GDPR	General Data Protection Regulation

1

INTRODUCTION

1.1 GENERAL CONTEXT AND MOTIVATIONS

Audio-only Augmented Reality (AAR) is part of the larger concept of Augmented Reality Augmented Reality (AR). Augmented reality refers to the process of reaching a flawless superimposition of reality and computer-generated elements. The general context of AR is applicable to visual perception (superimposition of virtual images on real images), auditory perception, and/or proprioceptive perceptions such as touch.

AAR limits the modality of its virtual objects to the auditory sensory modality. Its main goal is then to reach the seamless integration of virtual sound sources within the environment of a user. The generation and display of virtual sound sources necessitate a variety of technical challenges, including producing 3D audio, allowing real sound sources to remain unaltered by the device used for sound rendering, and developing technologies that enable motion tracking.

A more detailed breakdown of the different processes involved in the integration of virtual sound scenes into a user's environment in AAR applications is presented in Figure 1.1.

In the past few years, the technological development of spatial audio rendering, especially for binaural audio and headphones, has facilitated the delivery of AAR applications. The technology has been used in a variety of applications, including teleconferencing, location-based games, and education.

The research work in this thesis was motivated by an inherent question that led to the choice and development of spatial rendering methods for AAR applications: how precise should a spatial rendering process be to satisfy the seamless integration of virtual sound events ? More precisely, among the different dimensions concerned, we only focused on the rendering of sound source distance. The main motivation of this thesis is at the interface between two different topics:

1. Auditory distance reproduction: in spatial audio reproduction, an inherent compromise exists in the reproduction of an acoustic environment between the access to a priori information, the use of limited computational power, and still reaching a satisfying level of precision in the rendering of a virtual sound scene. In the case of AAR, the main goal is for the virtual sound source(s) generated to be perceived at the intended location.

2. Auditory distance perception: a large body of research has been conducted on auditory distance perception, demonstrating its reliance on a variety of

acoustic and subjective cues. Common patterns, such as inherent compression and variations in perceived distance, have been revealed. It has nonetheless been established that auditory distance perception is highly dependent on the precise context of perception: the presence of visual signals, the type of source employed, the amount of reverberation in the acoustic environment, and so on.

Auditory distance perception has rarely been evaluated in experimental contexts emulating generic AAR scenarios. Additionally, these scenarios could put into perspective the requirements needed to be met by distance rendering methods in the spatial audio modules of AAR technologies. This initial motivation led us to research multiple possible effects on auditory distance perception of virtual sound sources that could arise from AAR scenarios and to verify their relevance using perception experiments.

1.2 OBJECTIVES OF THE THESIS

The most obvious subject that has arisen was the role of acoustic-only cues for auditory distance perception, and how their reproduction was critical in AAR applications. The first objective was to evaluate the cognitive process behind the use of reverberation and if it could be summarized by a minimal quantity of objective criterion. We focused on how the combination of such a cue with sound level, another primary distance cue, produces an auditory distance judgement. We assessed to what extent the weighting attributed to each acoustic cues in the combination was individual or could be linked to the environmental context of the user.

The second part of the objectives concerned the influence of the environment on auditory distance perception. What partly defines AAR is the capacity for the user to have access to the real environment through all sensory modalities, vision included. Vision, when available, is well-known to have strong cross-modal interactions with audition in localization tasks. The condition of associating a visual source with an auditory signal through a so-called "ventriloquist effect" has been subject to a large body of research. However, how the visual modality can act as a calibration scale for auditory distance perception has been subject to relatively little research in comparison. This aspect, which seemed critical and that could largely modify the auditory distance perception depending on the visual environment, caught our attention. A part of this thesis was searched to appraise its significance.

Finally, one characteristic of a substantial proportion of AAR applications that drew us to a last research objective is that virtual sound sources can co-occur with real sound sources. We assessed the possible intra-modal calibration effect where real sound sources calibrate the cognitive process behind the interpretation

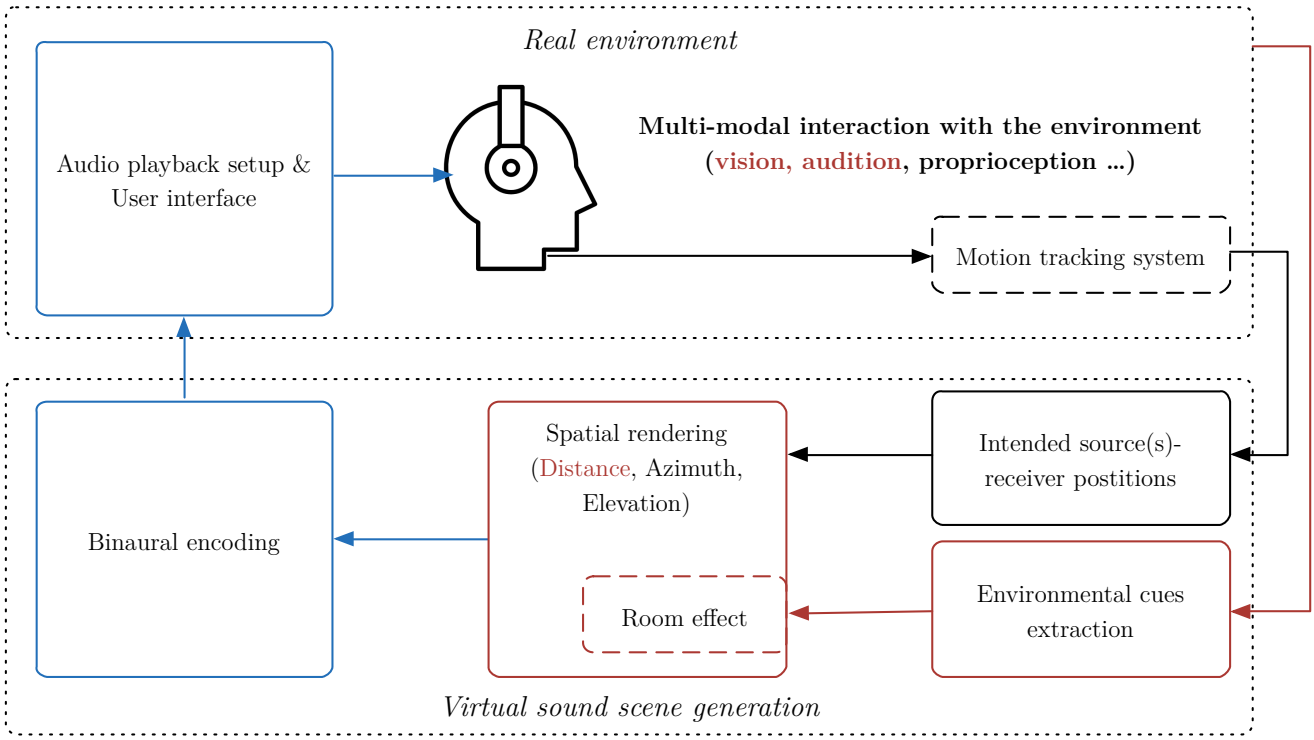


Figure 1.1: Schematic breakdown of the concept of AAR for a headphones display, all technologies used in the final process should be able to run in real time. All processes appearing in blue are part of the technical choices of the thesis framework described in Section 1.3. In red are the sub-processes in which are included the motivations and objectives described in Section 1.2.

of acoustic cues into distance judgements. Moreover, we explored which acoustic cue reproduction had a primordial role in this calibration effect.

1.3 KEY ASPECTS OF THE FRAMEWORK

The thesis consisted of the evaluation of the significance of the perceptual effects inherent to AAR scenarios on the auditory distance perception of virtual sound sources. The evaluation of the different effects was based on perceptual experiments, as we aimed to assess their significance across a broad range of applications. First, we had to keep the methods used for spatial audio rendering generic and easily reproducible to test different effects:

- Similarly to most AAR applications, we focused on binaural rendering for headphones and on the use of Spatial Room Impulse Responses (SRIRs). This technique can be used for spatial audio rendering and enables its decoding on different rendering setups as well as spatial transformations (e.g. rotation) in dynamic applications.
- The evaluation of acoustic cues were made with distance rendering methods based on the extrapolation of SRIRs from a single measurement.
- All virtual sound sources were spatialized in the far field (beyond 1 meter).

Second, we kept the experimental context of the auditory distance evaluations as generic as possible, so the results can be considered valid for more complex and specific situations:

- Only static situations were employed.
- Speech signals were used in all scenarios in order to keep the same degree of familiarity with the sound source for all participants.
- The effect of the vision of the room was explored. However, the perception of audio-visual sound sources was not studied. The distance perception of audio-visual sources depends on subjective factors and is mainly driven by the visual modality.

A longer explanation of these choices in the framework is presented, in the context of the previous research done in their related fields, in Chapter 2, Chapter 3 and Chapter 4.

1.4 CONTRIBUTIONS

The main contributions of the thesis are:

First, we successfully developed an online experimental protocol for frontal distance localization tasks. We also used it as an evaluation of the listener's environmental cues, which has never been done to our knowledge.

Second, we used and developed distance rendering methods to evaluate the use of reverberation and sound level for distance perception. We evaluated different criteria to define how listeners use the ratio between early energy and reverberation for distance perception and how it is relatively weighted with sound level to produce an auditory distance judgement. The results of the first of two experiments linked to this contribution has been published in the proceedings of *Forum Acousticum 2020* conference [109].

Third, we conducted online experiments to illustrate the importance of the calibration of the auditory space by vision. Findings from this study have been published in the special issue *Psychoacoustics for Extended Reality (XR)* of the journal *Applied Sciences* [108].

Finally, we evaluated, through other online experiments, the calibration of the auditory distance perception of virtual sound sources through the perception of co-occurring real sound sources. The presented experiment's results were presented at the *ASA Meeting 2021* [107].

1.5 THESIS STRUCTURE

The first part of the manuscript introduces the thesis setting within its connected research areas. In Chapter 2 we introduce the concept of AAR and we discuss how our objectives are related to the many challenges associated with enabling AAR applications. Chapter 3 focuses on an overview of the mechanisms and behaviors behind auditory distance perception. We address how perceived distance is considered to be tied to acoustic and non-acoustic cues. Additionally, we present how auditory distance judgments are usually predicted. Finally, the relationship between auditory distance perception and externalization is reviewed. The motivation for evaluating acoustic cues is discussed in relation to this section of the thesis's theoretical foundation. In Chapter 4 the reader is introduced to the role of vision in auditory distance perception. To begin, a review of the mechanisms and performance of visual distance perception is presented, emphasizing its higher reliability in measuring distances when compared to the auditory modality. Then, we discuss the different types of multisensory effects that could have an effect on auditory distance perception. To conclude the first part, we explain our choices for studying the auditory space calibration impact caused by vision in AAR applications.

The second part of the thesis describes the different methods used in the design of experimental protocols. Chapter 5 covers the technical approaches used throughout the thesis to display spatialized virtual sound sources over headphones. The first choice of using Spatial Room Impulse Responses (SRIRs) was made since it appears to be a promising technology for AAR. We detail the procedure by which we measured and converted initial measurements into usable Binaural Room Impulse Responses (BRIRs). Chapter 6 refers to specific methods used for the experimental protocol. The first section describes in detail the auditory distance report-

ing method used in all experiments. Its advantages and drawbacks are compared to other regularly used methods. The second section discusses the online protocol that was extensively implemented throughout the thesis studies. We go through the existing tools as well as the ones we used to build online experiments with.

The third part of the thesis presents the results of the different experiments achieved to measure the impact of acoustic or environment-related cues on auditory distance perception of virtual sound sources. Chapter 7 presents and discusses the results of experiments aiming to evaluate the use of acoustic cues for distance perception. A particular emphasis is put on the definition and evaluation of a reverberation-related distance cue. It contains results of a preliminary lab-based experiment including the perceptual evaluation of sound sources generated by different distance rendering methods. It is followed by the results of two online experiments designed to corroborate the first findings. Chapter 8 presents and discusses the results of an online experiment examining the impact of visual environmental cues on the auditory distance perception of virtual sound sources. Two hypotheses were tested, how a visual spatial boundary could act as a calibration of the auditory space and how room volume could calibrate the acoustic cues weighting strategies. Throughout this thesis an auditory-only discrepancy, due to the reproduction of the room effect will be referred to as "divergence". A discrepancy between auditory cues and visual environment will be mentioned to as an "incongruence". Chapter 9 presents and discusses the results of an online experiment measuring one possible impact of the presence of co-occurring real sound sources on the auditory distance perception of a virtual sound source. We assessed the impact of an acoustic divergence between real and virtual sound source. We notably investigated if an intra-modal calibration effect, caused by the presence of the real sound sources, could occur.

To conclude, Chapter 10 summarizes the work realized in this thesis. A broad overview of the contributions and findings is included, as well as some propositions for further research directions.

Part I

THEORETICAL CONTEXT

2

AUDIO-ONLY AUGMENTED REALITY

This chapter aims at giving an overview of the concept of audio only augmented reality. First, the definitions of Augmented Reality (AR) and Audio-only Augmented Reality (AAR) are presented in Section 2.1. Then, a review of different applications and devices used in AAR is given in Section 2.2. Finally, the main technical challenges emerging from the requirements enabling AAR are presented. To conclude, the positioning of the thesis among these different challenges is characterized.

2.1 INTRODUCTION

2.1.1 *Augmented Reality Definition*

Nowadays, Augmented reality is a well-known concept. Its first technological development goes back to the 1960s [160] and in 1994 Milgram and Kishino [122] formalized the "reality-virtuality continuum", a continuous scale defining all the forms of mixed space that aim to blend virtuality and reality. In this scale, AR and Virtual Reality (VR) are both parts of the general concept of mixed reality.

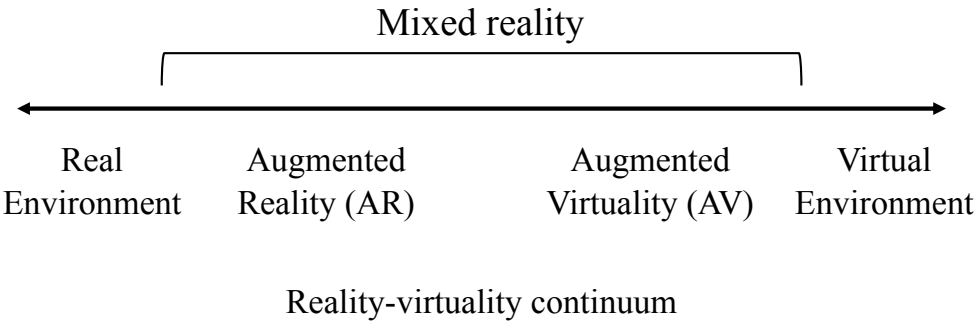


Figure 2.1: General principle of the reality-virtuality continuum as defined by Milgram and Kishino [122].

Whereas VR aims to replace reality by virtuality, augmented reality refers to a set of technologies and devices able to enhance the perception of the real environment through the superimposition of virtual objects on a physical environment. From the user's point of view, virtual elements are positioned and aligned in order to appear as if they were part of the real world. Thus, in all applications of AR, the users should keep contact with the real world. A specific definition of AR was introduced in 1997 by Azuma [13]:

- AR combines real and virtual content.
- AR is interactive and runs in real time.
- All virtual content is registered in the real world in 3 dimensions.

2.1.2 *Audio-Only Augmented Reality Definition*

The general context of AR encloses technologies linked to all sensory modalities and is not limited to a single one. AAR is considered as an outgrowth of the AR concept, it restrains the modality of the virtual content to audition. It does not imply that the user only has access to audition but that virtual events should only be auditory while the user has complete sensory access to the real-world [71]. A fourth property that can be added to the definition of AR introduced by Azuma to explain the concept of AAR is that all virtual content displayed to the user should be auditory.

The use of headphones or loudspeakers for AAR can be restricted to self-explanatory mono (0-dimensional), stereo (1-dimensional), or surround (2-dimensional). However, most applications developed nowadays focus on 3D auditory display on headphones. This chapter mainly focuses on the technological aspects and applications induced by this type of auditory display.

2.2 SOME AAR APPLICATIONS

The emergence of affordable consumer technologies for 3D audio listening, as listed earlier, has facilitated the delivery of AAR applications. Most of the applications were thought of as mobile (MAAR) or wearable audio augmented reality (WAAR) applications. Indeed, in mobility contexts, audio is an appealing alternative to vision as a display modality, avoiding the physically and cognitively demanding interaction with graphical user interfaces when on the go [60, 75]. In this section, we present a review of different AAR application scenarios. The existing applications can be categorized as: human-to-human interactions (Section 2.2.1) or location-based applications (Section 2.2.2).

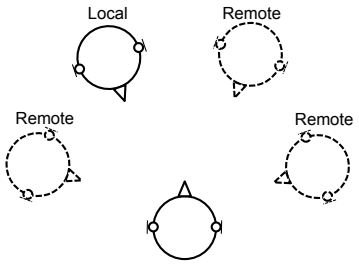
2.2.1 *Human-to-human interactions*

BINAURAL TELEPHONY A normal communication through mobile phones transmits through a mono signal with a limited sound bandwidth of 300Hz to 3400Hz, even with the use of headphones. Binaural telephony means that the signal transmission is done through a Head-Related Transfer Function (HRTF) (See Section 2.3.2.1). This type of signal is incompatible with normal telephone lines or GSM networks, and thus must be done through the Voice over IP standard. A standard of communication that allows no limit to the use of frequency bandwidth. In a telephony

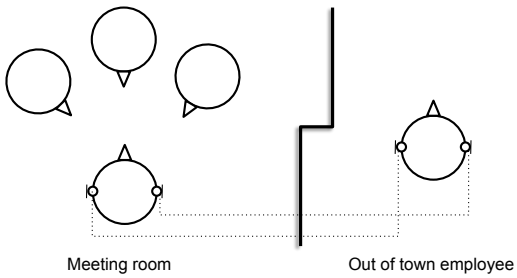
scenario, the use of binaural signals can allow one interlocutor to be spatially located from the perspective of the other. The spatialization of the interlocutor's voice in front of the other creates a more natural feeling of the conversation [103].

AUDIO MEETINGS This principle can be extended to audio meetings where multiple users are present. In recent years, audio meetings have become increasingly popular. Since the COVID-19 Pandeminc, face-to-face meetings have become increasingly difficult.Audio meetings are traditionally conducted via telephones and speakerphones. One of the issues is the lack of telepresence. Each voice of the participants in the meeting is displayed on a single mono channel. The use of spatialization of the different interlocutors can take virtual meetings to a new extent. Remote interlocutors can be panned around the user (see Figure 2.2) and blended into the user's acoustical environment. The key benefit is that it enhances the overall comprehension of the conversation [103].

Among the different possible use cases of this type of AAR, one instance is when a traditional meeting is scheduled and one team member is unable to attend because he or she is out of town. This member can virtually participate in the meeting if he or she has an AAR device and at least one speaker is present at the meeting (see Figure 2.2). One downside being the need of a speaker system at the other end.



(a) Multiple remote and local participants



(b) Single remote and multiple local participants

Figure 2.2: Exemple of audio meetings situations.

2.2.2 *Location-based applications*

AAR devices providing motion-tracking capabilities allow for location-based applications. The basic idea is that the AAR application knows the relative or absolute geographic position of the user and can display virtual sound sources associated with a specific location or a real object.

ART EXHIBITION AND MUSEUM TOURS One use-case being explored for location-based applications is museum tours [19, 39, 185]. The challenge tackled by these applications is to embed sound recordings into works of art. When the user moves in the physical space of the exhibition, he or she can experience the display of virtual sound sources placed in the environment, associated with a specific work of art. The idea is to increase the perceptual, emotional, and pedagogical effects of a museum exhibition. Moreover, the transparent aspect of AAR devices allows people to keep social contact with other visitors and interact as they experience the exhibition.

AUDIO TOURIST GUIDE The museum tour principle can be extended to audio tourist guide applications for navigation in outdoor areas. Urban tourists can be found exploring unfamiliar areas alone or with other tourists in a variety of ways, ranging from unstructured, spontaneous, and unorganized explorations to completely structured and fixed navigation. As a result, unlike systems for indoor application settings, AAR tour guide systems for outdoor, metropolitan areas must meet distinct and perhaps more complicated user demands, such as freedom of choice, open-space exploration, and social interaction [23, 113, 162].

It has the potential to replace tourist guides by allowing users to explore a city without following predetermined paths or schedules. With the application being aware of the user's location and direction of gaze, the user can then use the AAR headphones to wander around the city and automatically hear information on interesting places they pass by. Some applications tried to add the possibility for users to create and share their own recordings associated with specific locations[23, 113].

PERVASIVE GAMES Pervasive games represent an emerging game genre, wherein the gameplay is transferred from the virtual to the real world, thus spatially, temporally, and socially extending the way of playing games. Pervasive games are based on scenarios exploiting environmental information about the user. The use of spatialized sound is beginning to be exploited in prototypes of pervasive games[38, 53].

2.3 TECHNOLOGICAL CHALLENGES OF AAR

For an application to be defined as AAR, many issues must be overcome so the experience is satisfying. AAR aims to superimpose virtual events on a real acous-

tic environment [71], meaning that the listeners should perceive the surroundings as if they were not wearing headphones and should not be able to distinguish between real and virtual sound sources. This is still one of the most difficult technological challenges to overcome in the area [54].

Among the different processes presented in Figure 1.1, there are three main challenges arising in order to achieve the seamless merging of real and virtual sound sources. Firstly, real sounds must be unaltered by any devices implied in the AAR applications. Secondly, the audio system must create a virtual sound that closely resembles real sound and embed it within the real environment in which the user is located. Finally, to enable interactivity in AAR applications motion tracking is often integrated into the overall process.

We present here an overview of how these three issues are generally tackled in AAR applications.

2.3.1 *Treatment of real sounds*

The rendering of real sounds is closely linked to the playback system used for the AAR application. A loudspeaker array may be used to create a static AAR environment, but the true value of AAR is realized when it is mobile. That is why almost all actual applications in AAR implies the use of headphones (see Section 2.2). Moreover, headphone-based systems for spatial rendering provide the following advantages: they are portable and they have high channel separation, allowing precise control over the ears input signals [151].

The main problem with using headphones in AAR is the blockage of real sounds. Whether they are closed, open, or in-ear, the structure of the headphones attenuates sound coming from the acoustical environment of the user. This attenuation of the real sound can lead to an alteration of the user's natural hearing capabilities and localization accuracy.

Two types of solutions can be distinguished, so the headphones used are as transparent as possible [146]. The first possibility is using a hear-through equalization on headphones equipped with microphones in order to compensate for the attenuation due to the headphones' structure. A second one is to use bone conducting devices that are transparent by design and leave the ear canals open [71, 100].

2.3.1.1 *Hear-through equalization*

The wearing of conventional headphones significantly modifies the natural hearing process, regardless of its type [145]. This effect leads to a modification of the auditory cues conveyed by a sound source, and leads to an alteration of the perception of sound localization. Several attempts to playback spatial audio on conventional commercial headphones have been made in recent years. A first AAR headset was developed in 2004 using a pair of in-ear binaural microphones, and an individual equalization process to assist the user in hearing real acoustic scenes[71]. This

system was further developed with the addition of an AAR mixer [144, 162]. The mixer allows to superimpose virtual scenes to the rendering of the real sounds on the headphones. A basic system diagram of the functioning of the overall device is displayed in Figure 2.3. Works from Gupta [68] refined this system by considering and compensating possible sound leakage from the headphones to the binaural microphones.

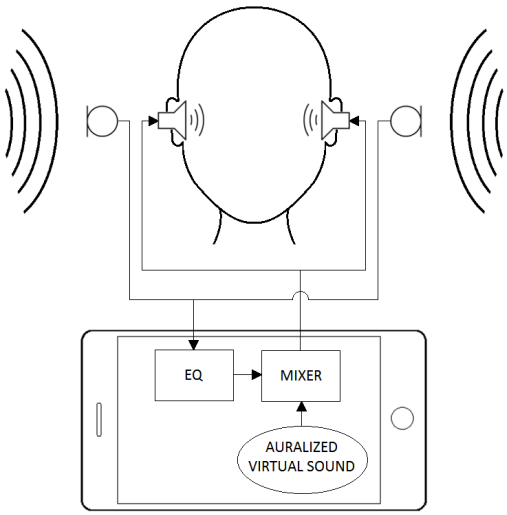


Figure 2.3: Diagram of a smartphone-based AAR system using Hear-through equalization. After Engel & Picinali [54].

2.3.1.2 *Transparent headphones*

BONE CONDUCTION One of the transparent headphone technology that was successfully used for sound spatialization in AAR scenarios is bone-conduction [100, 105].

Bone-conductive headsets, sometimes known as "bone-phones" [169], transfer sound to the cochlea by producing vibrations in the skull directly, bypassing the ear canals. Bone-phones have already been used to produce spatialized audio and "hear-through augmented reality" [105]. Different works by Lindeman and Barde [14, 15, 101] tested the performance of a bone conduction headset in various localization tasks (elevation, azimuth, and externalization) of virtual sound sources. Its performances were demonstrated as satisfying in terms of elevation and externalisation, but slightly poor for azimuth reproduction when compared to regular headphones or speaker arrays.

Most of these AAR experiments with bone conducting headphones, individualized HRTF were used to enable sound spatialization. A large part of the localization errors attributed to these works are explained by the incapacity of the headphones to reproduce faithfully the attributes of this transfer function. This is explained by the relatively narrow frequency range of bone conducting headphones (e.g

Audio-Bone used in [105] ranges from 50Hz to 4kHz) when compared to regular headphones.

Figure 2.4: Bone-conducting headphones used by Macdonald et al. [105]

OPEN-EAR HEADPHONES The use of transparent headphones has been introduced by Tappan in 1964 [161] suggested using "Nearphones", i.e small loudspeakers worn near the ears. The *Bose Frames* and the *Microsoft Hololens* are two open-ear headphones currently on the market. Due to the recent emergence of these devices, little to no performance evaluation of these devices has been made in AAR scenarios. Nagele et al.[129] noted that the high latency of the *Bose Frames* was a problem for the reproduction of virtual sound sources in their interactive AAR scenario.

Figure 2.5: *Bose Frames* designed for AAR application, based on speakers integrated to the frame of the glasses near the user's ears.

2.3.2 *Generating a spatialized virtual sound source*

An augmented reality system should be able to merge real and virtual sounds so that the virtual sounds seem embedded in the real environment. Harma [71] even suggests that ideally an augmented reality audio system should withstand a test close to the Turing test for artificial intelligence [163]. If a listener can't tell whether a sound source belongs to the real or the virtual audio environment, the software generates a subjectively flawless augmentation of the listener's auditory environment.

This requirement demands the application of appropriate spatialization processing of virtual auditory events in order to meet this criterion, so the virtual source seems to be emitted in the real acoustic environment. Moreover, the room effect is well known to contribute significantly to the perceived location of sound events [20, 90]. As a result, the room effect processing must be carefully designed to ensure that the perceived location of a virtual event corresponds to the intended position. In AAR applications, the idea is that the virtual event seems to originate from a precise location in the environment, which could be, for example, a specific real-world object or position, or perceived behind, next to, or in front of other real-world sound sources.

The challenge is two-fold: a) choosing an appropriate spatialization model to control the placement of virtual auditory events and the related room effect in order to meet room-related perceptual criteria, b) obtaining a priori knowledge about the acoustical or architectural features of the real environment to tune the model appropriately; and c) being able to run in real time, since it should be applied in an AAR scenario. The spatialization model chosen has a direct impact on the replication of auditory cues transmitted by the room effect. It will have an impact on the spatial perception of a sound source and, more broadly, on the perceptual representation of the overall virtual sound scene.

This current section focuses on the different methods used for the spatial rendering of virtual events in front of the listener and how to perform a binaural rendering for a headphone display.

2.3.2.1 *Binaural Rendering*

A normal-hearing listener can obtain all the auditory information about incoming sounds, such as distance and direction, with just two ears, depending on the time, level, and frequency content of the sound signals received at the two ears. Thus, if the two ears' signals can be replicated exactly as in direct listening, a flawless reproduction of the real auditory scene may be synthesized and, as a result, natural sound can be generated. The disparities in sound received at the two ears are caused by shadowing, reflections, and scattering of sound on the torso, head, and pinnae prior to reaching the ears [121]. Three so-called binaural cues emerging from these disparities can be interpreted by the auditory cognitive system to enable spatial localization Inter-aural Time Difference (ITD), Inter-aural Level Dif-

ference (ILD) and spectral cues [124, 130]. ITD and ILD are crucial for the perception of the sound azimuth and the near-field distance. Spectral cues are required for the perception of elevation.

The reproduction of the filtering due to the head, torso, and pinnae and, therefore, of the binaural cues can be achieved by the use of a Head-Related Impulse Response (HRIR) in the time domain or by the use of a Head-Related Transfer Function (HRTF) in the frequency domain. A HRTF is a set of two functions, that are defined as the filtering relationship between the sound pressure at a point inside the human ear canal and the sound pressure at the center of the head in the absence of the listener [22]. The profile of a HRTF is highly dependent on the different morphological structure of each individual. In practice, HRTFs are measured by placing a microphone at the entrance of the ear canal (blocked ear) of a participant and measuring impulse responses from desired directions in a free-field condition. The Fourier transform of these measured HRIRs constitutes the HRTF set of an individual. A large number of measurements is necessary to record HRTFs with an adequate resolution.

The rendering of HRTFs necessitates channel separation in order to precisely control the sound signal at each ear. Thus, headphones are particularly suited for their reproduction. In order to spatialize an auditory object for headphones reproduction, a filtering of a monophonic input signal by the HRTF is generated for each ear. The auditory object is then perceived to be positioned in the same direction as the measured HRTF. This process is referred to as "binaural synthesis" [84].

A common problem in headphone reproduction is the inside-the-head localization (IHL) [22]. It is a perceptual phenomenon characterized by the perception of a virtual sound source inside the head. Similarly to real sound sources, the listener must perceive spatialized virtual sound sources as coming from outside the head. This phenomenon is defined as "externalization" [72]. A review of the cues involved in externalization can be found in reviews by Durlach et al. [52], Blauert [22] and Best et al. [21]. Among the cues enabling externalization, binaural cues play a major role.

Reaching externalization with headphones reproduction is a challenging problem. One way to tackle it is to use personalized HRTF [88]. Using personalized HRTF allows for the correct reproduction of individual binaural cues, suiting the expectations of a listener concerning the content of externalized sound [21]. This method has been proven efficient for the externalization of virtual sound sources in the horizontal plane, but sound sources presented frontally have been shown to be difficult to externalize [141]. Sources located in front present a high correlation between the sound signals at each ear. This lack of differences, and thus of binaural properties in the sound signals, leads to internalization or front-back confusion effect (a source in front of the head perceived behind it) [72].

Personalized HRTF are measured most often in free-field conditions, and pairing them with a room effect is necessary to enable spatial sound reproduction. The combination of a personalized HRTF with a room effect (see Section 2.3.2.2)

increases the decorrelation of the sound signal at the two ears and is therefore beneficial for externalization [21]. The concept of externalization is more extensively reviewed in Section 3.3, notably its relation with the perception of sound source distance.

2.3.2.2 *Room effect reproduction*

The ability to reproduce the room effect is a critical prerequisite for the AAR applications in order to provide expected room-related perceptual cues to the user. The process of reproducing the room effect is referred to as "auralization." In this section, various existing auralization approaches are presented, including the straightforward Room Impulse Response (RIR) convolution approach, physically-motivated approaches for reproducing the sound field, and perceptually-motivated approaches with the goal of accurately reproducing room-related perceptual criteria. The benefits and drawbacks of these various techniques for AAR applications are highlighted.

RIR is defined as the transfer function of the sound between a source and a receiver in a room. It precisely characterizes the acoustic signature of an environment. A RIR between two places in a room is generally measured by producing a deterministic signal, such as a sine sweep, with a loudspeaker at one point and a microphone recording the sound pressure at the other [156].

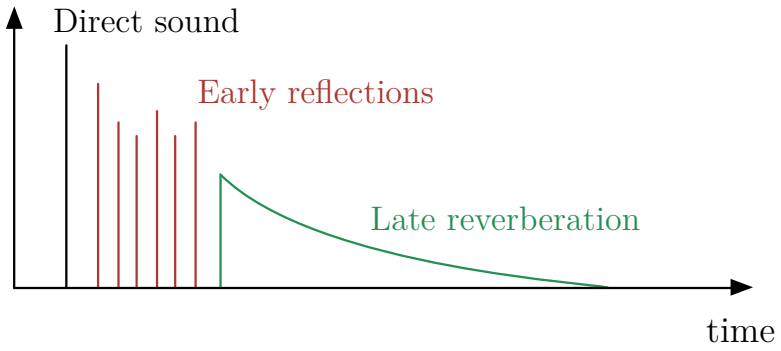


Figure 2.6: Standard Room Impulse Response (RIR) decomposition.

An anechoic input signal can be convoluted with a RIR to produce a reverberated signal corresponding to a single source-receiver position in the room. Because of the technique's precision, the convolution approach is particularly suited for comparison or predicted auralization. However, unless numerous microphones or a moving microphone are available, measuring RIRs over a wide range of positions and orientations can be a laborious and time-consuming task. Therefore, interactive auralization with a moving source or receiver is not permitted. Interactivity

can be partly enabled with the use of SRIR measurements. This type of multichannel recording can be measured with a spherical microphone array. These measurements enable rotation transformations and can be decoded with a personalized HRTFs for binaural rendering. A longer explanation of this technology and of the associated decoding process is presented in Chapter 5.

Because of the intrinsic constraints of the convolution approach, several room effect reproduction methods have been developed to replicate, at least partially, the performance of the convolution approach with actual RIRs.

A solution to the reproduction of a specific room effect is to use physically-motivated approaches to approximate a solution of the wave equation in the acoustic environment. Computational acoustics methods such as finite-difference time domain (FDTD) [94], finite element method (FEM) [154], and boundary element method (BEM) [89] are physcially-motivated [168]. This type of methods are based on the time and space discretized solutions of the wave equation. The accuracy of the overall solution is linked to the level of discretization employed and, thus, to the computational power allocated.

Another type of physically-motivated approach is geometric modelling methods. It designates the use of one or several methods, such as the image-source method (ISM) [6], ray tracing [95], and its variants to approximate the calculation of real RIRs.

In order to achieve a satisfying level of accuracy, these solutions come at a very high computational power cost. However, in the case of AAR, the spatial audio module used is governed by several constraints:

- To reach a satisfying level of precision in the reproduced localization
- To reach a satisfying level of realism in respect with the enhanced environment
- To be able to run in real time in order to enable interactivity and communication with a motion tracking system

Ultimately, the precision and realism are not only limited by the computational power but also by the necessity to run in real time to enable interactivity. Therefore, physically-motivated approaches, due to their cost in computing power, are not adapted to AAR applications.

It is then critical to identify which characteristics of the room effect must be reproduced, and what is not necessary when developing a spatial rendering module for AAR. Indeed, a full model that accounts for all the characteristics of spatial hearing has yet to be produced. Nonetheless, algorithms with reduced computing and hardware costs but perceptual performances comparable to more complex physically-motivated systems have been developed. These methods rely mainly on perceptual effects in order to generate the room effect, and can be referred to as "perceptually-motivated".

One method is based on the use of Feedback Delay Networks (FDNs) [83]. The general idea behind their use is to synthesize a generic room effect, reproducing different parts of a RIR. It consists of a recursive delay network that can generate the first reflections and the late reverberation for an input signal. Their design aims to reproduce a desired room effect by manipulating the energy and the time-frequency profile of the generated first reflections and late reverberation. The tuning of the FDN parameters is often made through measures of the acoustical characteristics of the reproduced room effect [37].

In order to enable interactivity, this type of method needs to be paired with a distance rendering method that tunes the FDN parameters as a function of the source-receiver distance. Moreover, another model must be added to modulate the room effect as a function of the source and receiver orientations.

Finally, RIR synthesis method by extrapolation of existing measurements can be mentioned [177] as perceptually-motivated. This method implies identifying the relevant cues and information in real measurements, in order to feed a model that extrapolates RIR corresponding to other source-receiver positions. This method is relatively unusual when compared to FDN applications. During this thesis, we chose to develop and apply an extrapolating method of this type. We notably aimed to define what objective criteria in terms of reproduction must be met in the reproduction of RIRs.

2.3.3 *Motion tracking*

To enable interactive AR applications, knowing the user's position and orientation is necessary. Therefore motion tracking modules are often implemented in AR applications. The exploitation of a geographic absolute location and the head orientation of the user allows overlaying information onto the physical environment. To do so, the module tracking device should aim to achieve the following characteristics:

- provides orientation and position information of the user.
- transmits tracking data with minimum latency and high update rate to avoid discrepancy between proprioceptive sensations of the user and the auditory feedback.

A complete overview of the different solutions generally used to generate motion tracking in the context of AR can be found in [60].

2.4 SUMMARY AND TECHNICAL CHOICES

We define here how the work done in this thesis is located among the different challenges inherent to AAR and presented in the previous sections. The different technical choices resulting from these challenges are also precised.

As precised in Chapter 1, the main motivation of this thesis is the understanding of the auditory distance perception of virtual sound sources for AAR. The goal is to investigate how AAR-specific contextual effects influence auditory distance perception and how prerequisites can be determined for the design of distance rendering methods.

First, among the main technical challenges existing in AAR research, the thesis mainly concentrates on issues linked to the production and perception of virtual sound sources. Among the different dimensions possible (azimuth, elevation, and distance), this thesis focuses on frontal auditory distance perception. and omit the treatment of real sounds (Section 2.3.1). Nevertheless, the impact of the presence of co-occurring real sound sources on the perception of acoustically divergent virtual sound sources was addressed in Chapter 9. We notably examined whether the presence of divergent and co-occurring sound sources has an effect on the calibration of the auditory space, and which acoustic cues ultimately drive this intra-modal calibration.

Second, like most of AAR applications developed (see Section 2.2), binaural rendering for headphones was used to display virtual sound sources (Section 2.3.2.1).

Third, all the acoustical information employed in the experiments was extracted from SRIRs measured in the room that is supposed to be reproduced by the distance rendering method. We have notably tested in Chapter 7 a distance rendering method based on the extrapolation of a single SRIR measured in a reference room. This choice was motivated by the limitation in AAR applications to have access to a priori acoustical information and the limited computational power inherent to these applications (see Section 2.3.2.2). This method seemed as an easy way to encapsulate the acoustic signature of a room. Moreover, manipulations of the initial measurement could easily been done in real time, in accordance with a key requirement of AAR applications. Therefore, the main challenge was to find out which prerequisites a distance rendering method based on SRIR extrapolation must fulfill.

Finally, due to the COVID-19 pandemic, most experiments were done with remote participants, preventing the use of motion tracking systems. In this way, issues related to motion tracking (Section 2.3.3) are not treated, and the experiments presented in this thesis are limited to static scenarios.

3

AUDITORY DISTANCE PERCEPTION

The mechanisms and performances of auditory distance perception are discussed in this chapter. In Section 3.1, the systematic biases linked to auditory distance perception are presented, as well as a common model used for estimating the tendency to compress perceived distances. Furthermore, in Section 3.2 the different cues involved in auditory distance perception are listed, and more specifically, for the frontal plane. The chapter is concluded by presenting how the literature has impacted the methods and motivations of the thesis work.

3.1 AUDITORY DISTANCE ESTIMATION

Auditory distance perception of a sound source is a complex process involving the perceptual system's interpretation of acoustic and non-acoustic information. Depending on the characteristics of the listening situation, the same source-receiver distance can be estimated very differently. The explanation of this behavior is addressed in Section 3.2. However, a pattern in the way a sound source distance is translated into a percept has been established. In this section, we first present the systematic biases observed on sound source distance judgements. We limit our review to sound sources in the far field, i.e. at distances superior to 1 meter and to studies relying mainly on absolute distance reports.

3.1.1 *Auditory distance perception accuracy: An inherent compression effect*

Perceived auditory distance is inherently compressed. Listeners tend to overestimate the distance of far sources and underestimate the distance of close sources [90]. What is meant by "far" and "close" is related to the concept of a "crossover point" [7]. It is the distance for which there is no bias in perceived distance; its value is considered to be around 1 meter but varies depending on acoustic environment characteristics. Contrarily to azimuth and elevation, which present limited absolute errors of localization [22], the error in auditory distance perception is virtually infinite for increasing distances. As a result, it is regarded as the most imprecise dimension in the perception of sound localisation.

In order to characterize auditory distance perception, Zahorik proposed a model [183] where the perceived distance D is related to the actual distance d through a

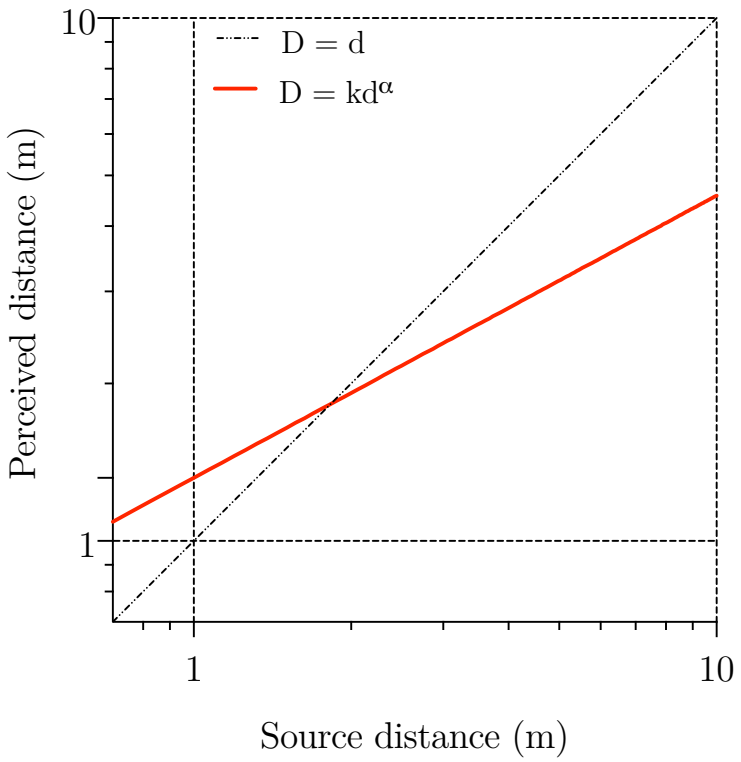


Figure 3.1: The perceived distance as a function of the sound source distance according to the compressive power function (in red) introduced by Zahorik [183] over a logarithmic scale.

compressive power function, which is a suitable approximation to the majority of psychophysical distance functions:

$$D = k * D^{a} \tag{1}$$

Where k and a are the fitting parameters of the function. They are respectively called the linear compression (when $k > 1$) or expansion ($k < 1$) coefficients, and the non-linear compression (when $a < 1$) or expansion ($a > 1$) coefficients. They are equivalent to the slope and intercept when represented on a logarithmic scale (see Figure 3.1).

This two-variable function offers a comprehensive representation of the compression effect on a set of reported distances and is used thoroughly in this thesis work. In [183], 84 data sets were fitted with the above compressive power function. A mean value of 1.32 was found for k, while a mean value of 0.54 was found for a. This result illustrates the systematic compression effect observed on auditory distance reports.

This compression effect has often been held responsible for the so-called "auditory horizon effect". This effect is defined as the limit from which an increasing distance is no longer perceived [27]. Because of the compression effect, all increasing distances are progressively absorbed by the perceptual blur of auditory distance perception. As a result, if the auditory horizon is detectable in a set of distance reports, its value can also be used to characterize a compression effect present in a set of distance reports.

3.1.2 *Auditory distance perception variability*

Auditory distance perception manifests a large intra-subject variability in auditory distance reports, often mentioned as a "perceptive blur". Zahorik [180] has shown that the variability in distance reports can reach 20 to 60% depending on the participant. Haustein [74] is more optimistic and reports an intra-subject variability of 5 to 25% for distances ranging from 0.9 to 9 meters. This intra-subject variablility, often qualified as a "perceptual blur", is also linked to the familiarity of the participants with the listening situation (nature and characteristics of the source and the environment) [22].

The large differences in terms of observed accuracy and variability across different studies may also be due to the different report methods used. Possible biases due to the report method are more extensively presented in Section 6.1.

3.2 AUDITORY DISTANCE PERCEPTION CUES

The perception of sound localization is based on the sound signal perceived through the two entrance points of the auditory system: the ears. The location of a sound source is processed by the cognitive system based on the objective and subjective content of the audio signals perceived at the ears. For an estimation of the sound source distance, a variety of cues are extracted from the perceived signals:

- **Acoustic (or objective) cues** are solely related to the signal. They include the four following cues [22, 90, 119, 183]: intensity, reverberation, spectral content and interaural differences.
- **Non-acoustic (or subjective) cues** are linked to the idiosynchratic processing of sound and vary from one person to another.

A distance percept results from the combined interpretation of these different cues. We first describe how each of the acoustic cues can be interpreted as an auditory distance. Then, we define how cognitive cues could also affect the final distance percept. Finally, we discuss the specificity of a dynamic situation in which the source and/or the listener are moving.

3.2.1 *Acoustic cues*

Among the acoustic cues, a distinction is made between absolute and relative cues. An absolute cue does not require multiple distance presentations of the same sound source, or a priori knowledge of the sound source from the user, to be interpreted into a reliable distance judgement. Usually, an acoustic cue becomes absolute when the user is completely familiar with the sound source and the surrounding acoustical environment (see Section 3.2.2). The acoustic cues and their significance for auditory distance perception are presented in ascending order of impact, with intensity and reverberation being considered as primary cues for auditory distance perception.

3.2.1.1 *Intensity*

As sound intensity decreases with distance, its level is informative of sound source distance [41, 59]. In the free field, an environment without any reflective surfaces, sound intensity is proportional to $1/r^2$, resulting in a sound level decrease of approximately 6dB per doubling distance. However, in the more common case of a reflective environment, the sound level decrease with distance is specific to the environment depending on its architectural characteristics. Thus, sound intensity requires not only the a priori knowledge of the sound source power but also of the environment in order to be absolute and thus be interpreted as an auditory distance [119].

Despite its relative nature, sound intensity is often considered as one, if not the most important cue for auditory distance perception [119], because of the capacity of the auditory system to detect small changes in sound level. Miller found that broadband noise changes could be detected if they were over 0.4dB [123], while sine waves' smallest perceptible changes were 1 to 2 dB [81]. For relative distance perception between two sound sources, the pressure-discrimination hypothesis states that just-noticeable differences are linked to the sound intensity sensitivity of the auditory system. According to this hypothesis, the auditory system should be capable of discriminating between two sound sources separated by a distance difference of 5%[12] to 25%[3].

3.2.1.2 *Reverberation*

Contrarily to the case of the free field (i.e., in a reverberant environment), part of the sound emitted by a source reaches the listener directly and the rest arrives after reflecting on surfaces. From an objective point of view, information relative to the distance between the source and the receiver could be extracted from the ratio of energies between the direct sound and the reflection. Direct sound decreases faster than reverberation when the distance to the sound source increases. It has been shown that an artificial augmentation of the reverberation in an acoustic environment leads to longer auditory distance judgements [31, 131]. More generally,

a room with a larger amount of reverberation tends to induce a longer impression of distance [117].

The Direct-to-Reverberant energy Ratio (DRR) is often considered as the cue that explains the role of reverberation in auditory distance perception. In theory, if the listener is familiar with the room, it is an absolute cue [119].

Warren demonstrated that the reproduction of this cue from a varying distance while maintaining a constant sound intensity still leads to coherent auditory distance judgements [173], even if those are more accurate if both cues are available to the listener [183].

The DRR could also be used as an explanation of the compression of perceived auditory distance (see Section 3.1), as the value of the ratio converges to a certain limit when the distance increases.

Despite a variety of results on DRR suggesting that it is a salient distance cue [31, 119, 179] [131, 180], its relevance for auditory distance perception has been questioned. Zahorik [181] indicated that the only discernible difference in the cue is around 6dB, which translates into surprisingly large differences in distance, a result that contradicts Warren's findings. He questioned how such an imprecise distance cue could be considered as salient. According to Zahorik[179], Bronkhorst [26] and later Kopčo and Shinn-Cunningham [93], it is very unlikely that the auditory system could effectively separate direct sound from reverberation, particularly for sounds with gradual onsets and offsets.

Several propositions have been made to find a more relevant reverberation-related distance cue. Larsen et al. [97] suggested that changes in acoustical parameters such as interaural cross-correlation or spectral cues parameters co-varying with the DRR could explain the role of reverberation for auditory distance perception. The early-to-late energy ratio has also been hypothesized to be a relevant reverberation-related distance cue, i.e. where the energy of the first reflections is aggregated to that of the direct sound [93]. Bronkhorst tested a predictive model in which he used an early-to-late energy power ratio to compute the perceived distance with success [27]. He also highlighted the role of the spatial distribution of early reflections. In [26] he showed that the early energy definition could be based on a lateralisation window (i.e. aggregation of the reflections sharing an ITD close to that of the direct sound) rather than on a time window.

3.2.1.3 *Spectral content*

Larsen et al. [97] stated that spectral changes could be related to the perception of the DRR. More precisely, he demonstrated that the perceptual sensitivity of DRR is correlated to spectral properties of the signal such as the spectral centroid and the spectral envelope. Apart from these findings, spectral cues are considered to influence auditory distance perception through two principles.

First, the high frequency components are progressively attenuated by air absorption when propagating over long distances. This effect was demonstrated to have a significant influence on auditory distance perception for sound sources at

distances superior to 15 meters [22]. More generally, sounds with a decreased content in the higher frequencies relative to the lower ones are perceived further away [31, 42].

Second, the sound scattering around the head varies with frequency and distance. For nearby sources (< 1m), these changes significantly influence auditory distance perception [28, 93].As the source approaches, a global low-pass filtering can be observed, associated with an increase of ILD at low frequencies.

In both cases the spectral content is considered as a relative cue, unless the listener is familiar with the spectrum of the sound source.

3.2.1.4 *Interaural differences*

When a sound comes laterally, it first arrives at the nearest (ipsilateral) ear and then at the opposite (contralateral) ear. This results in an interaural time delay ITD. Moreover, the presence of the head shadow reduces the level received at the contralateral ear and is responsible for an interaural level difference ILD.

Dichotic experiments where ITD and ILD may be manipulated independently, allow to reveal there respective roles and weights on the lateralisation of the sound percept[22].

For stationary pure tones, the ITD can be interpreted from the phase difference between the signals at the two ears. A complete lateralisation of the sound percept is only achievable for frequencies below 800Hz, i.e. where the maximum ITD corresponds to half-period. Interaural phase differences are only effective up to 1.6kHz. Above this limit, ITD is still effective for complex sounds (band noise, keyed signals...).

In the context of dichotic experiments, interaural level differences are effective throughout the full auditory frequency range. However, in a free sound field, the ILD depends strongly on frequency. At low frequencies, the shadowing effect of the head is significantly reduced. The ILD progressively vanishes and becomes ineffective.

In natural situations, ITD and ILD interact, and their relative weights depend on the sound signal, with increasing importance of the ILD above 1.6kHz and when the level is low [22].

Whereas both interaural cues play a major role in the perception of the sound source direction, their effectiveness for auditory distance perception is limited, except in the near field.

In the case of nearby sound sources (< 1m), and especially when presented laterally, these binaural cues have been shown to be beneficial to auditory distance perception. Indeed, since in the near field the wavefront cannot be assimilated to a plane wave [22], the binaural localization cues undergo significant changes with distance. In particular, the ILD changes significantly and increases progressively as the source moves closer to the listener's head [28]. In this case, the significant changes in ILD and ITD can be beneficial to auditory distance perception [153].

Another consequence is linked to the spectral cues originating from the scattering of the pinna [18], the head [50] and the torso [5] of the listener, which are represented by the HRTF. When the source is close enough to the listener, a substantial difference occurs between the angle of the source relative to both ears. This parallax induces different types of spectral filtering for both ears. Kim et al. [87] demonstrated the possibility to virtually elicit the sensation of nearby sound source distance from this auditory parallax. However, when primary cues such as intensity and reverberation are available, this effect is less significant. For instance, according to Zahorik [178, 180], the use of non-individualized HRTFs does not affect auditory distance perception.

3.2.1.5 *Acoustic cues combination*

Throughout this section, multiple acoustic cues relevant to auditory distance perception are listed. Their interpretation by the auditory system infers a distance percept.

In Section 3.2.1.2 we have notably seen that the perception of reverberation changes with distance can be linked to several cues: the DRR, the early-to-late energy ratio, spectral cues and spatial characteristics. Their classification does not imply that they are totally independent from each other. In a real listening situation all these cues are correlated to distance and, therefore, change simultaneously with it. Several studies have investigated the relative importance of these reverberation-related cues and of sound intensity for the perception of distance. Zahorik [179] showed that the perceptual system uses weighting strategies of intensity and of the DRR flexibly to produce a distance percept, depending on the characteristics of the listening situation: stimulus nature, angular position and source-receiver distance. Zahorik [183] also suggested that the weighting of cues may rely on the consistency associated with each cue. Cues that are either unavailable (e.g. reverberation in a free field environment) or unreliable (e.g. in a virtually created environment when a cue stays constant with distance) are given less perceptual weight in the combination process.

Recently, Prud'homme and Lavandier [141] showed in a study implying the perception of virtual sound sources, that when available, the primary mono-aural cues (sound intensity and reverberation-related), were sufficient for a majority of participants to judge the distance to the sound source. Spatial aspects had a limited influence, with exceptions made for some participants, notably due to the relation between externalization and distance perception. This relationship is more extensively reviewed in Section 3.3.

3.2.1.6 *Dynamic situations*

When the sound source and/or the listener are moving, additional dynamic cues contribute to the perceived source distance [90, 183]: the time-to-contact or acoustic tau, the absolute motion parallax and Doppler effect. They are mainly related

to changes of intensity, binaural and spectral cues, respectively.

The acoustic tau refers to the sound level variation occurring when the distance to the source changes [11]. It may be exploited by the perceptual system either for distance evaluation or for time-to-contact estimation when the source is looming or when the listener moves towards the source. Ashmead et al. [11] established that participants could benefit from increased auditory distance perception accuracy when they were able to move towards the source compared to situations where they stood still.

The absolute motion parallax refers to the case where the source and the listener are not moving exactly towards each other. In this case, the change in angular direction of the source creates dynamic changes in binaural information that can contribute to the distance estimation. Speigle and Loomis [155] notably illustrated the benefit of this effect in a situation where the sound source was displayed outside of the median plane of participants. Moving participants exhibited increased accuracy in judging the distance to a sound source compared to static situations.

A particular dynamic situation must be mentioned here. When a sound source moves towards a static listener (looming source), a systematic asymmetry in distance judgements is observed [70]. Indeed, the perceptual system tends to overestimate the change in intensity of looming sounds when compared to receding sound sources. This results in a systematic bias of underestimation of looming sound sources. This bias might be triggered by the perceived biological importance of looming sounds that could be potentially interpreted as a threat or an incoming collision [34].

Gardner [62] noticed that head movements could be very slightly helpful for auditory distance perception of speech signals in anechoic conditions. The main benefit that these movements could provide is the ability for the listener to hear the source laterally, which could enable the use of binaural cues for evaluating the distance of nearby sources [79].

3.2.2 *Non-acoustic cues*

The contribution of vision to auditory distance perception is reviewed in Chapter 4. However, other non-acoustic cues may influence auditory distance perception.

3.2.2.1 *Prior knowledge and expectation*

Without any prior knowledge of the sound source, most of the above described acoustic cues provide only relative distance judgements. In contrast, sounds familiar to the listener may enable the interpretation of acoustic cues as absolute distance judgements. Certain sounds, such as speech, are instantly recognizable

to all listeners, even more for languages with a prosody similar to their native language(s) [30]. Vocal signals also present particular characteristics that link their production mode (e.g. whispering to shouting) to an expected sound source power. Gardner [62] demonstrated that the estimated distance of a source playing back whispered speech is underestimated as a result of a low expected sound power. Conversely, the estimated distance to a source playing back shouted speech is overestimated due to its high expected sound power. Similar effects may be elicited by musical instruments or motor sounds.

3.2.2.2 *Learning*

Learning results from the repeated exposure to similar listening settings. Coleman [40] illustrated this phenomenon in a distance reporting experiment where participants were asked to assess the distance between an unfamiliar stimulus played back on loudspeakers distributed at various distances. Initially, listeners were unable to determine which loudspeaker was displaying the stimulus. As the session progressed, performances improved incrementally without any feedback from the researcher. Makous and Middlebrouks [106] and Kopčo et al. [92] also observed such a learning effect in similar experiments. Carlile [36] emphasized the importance of training sessions that enable participants to become acquainted with unfamiliar auditory environments and stimuli, as well as accustomed to the distance reporting method.

3.3 RELATIONSHIP WITH EXTERNALIZATION

In natural situations, sounds emitted from a physical source are perceived outside the head. This sensation is called "externalization". With the emergence of spatialization techniques and rendering over headphones, reaching and measuring externalization has been the subject of a large body of research. The contribution of different acoustic and non-acoustic cues to sound externalization has been reviewed by several authors [21, 22, 52].

The presence of reverberation and, more importantly, the presence of binaural cues enhances the externalization. Reproduction of spectral features related to the pinna is also important, which emphasizes the importance of using personalized HRTF for binaural rendering. Moreover, dynamic cues such as head movements and self-motion in general are also beneficial to externalization. Finally, the listener's familiarity with a specific listening situation and the presence of visual cues can contribute to a better externalization.

All these cues should fit the expectation of a listener about the spatial attributes of the acoustical environment. The lack of externalization, i.e. when the sound is perceived as originating from inside the head, results from a violation of this expectation [21]. Externalization was frequently linked to the plausibility and realism of the sound scene in studies with virtual spatial audio.

Another relationship that is often discussed and with conflicting conclusions is that with the auditory distance perception. As internalization is described as the perception of a sound inside the head, any signal that does not reach externalization could be qualified as perceived at a minimal distance. Thus, externalization could be considered as a prerequisite to a distance percept [141].

This assertion is still debatable. The primary justification for their depiction as distinct principles is that their perception is mostly based on distinct stimuli. While auditory distance perception is predominantly influenced by mono-aural acoustic cues (intensity and reverberation), externalization is driven by binaural cues. Additionally, the majority of studies examining auditory distance perception of virtual sound sources rarely take the externalization criterion into account. Frequently, participants are requested to judge the distance of frontal sound sources. Sound sources that are displayed on headphones through binaural rendering. A configuration that is well-known to produce poorly externalized stimuli. Furthermore, Zahorik [180] and Prud'homme and Lavandier [141] have demonstrated that the absence of individualized HRTFs had no effect on participants' auditory distance reports, despite the fact that it is a key factor in producing externalized stimuli.

3.4 SUMMARY & PERSPECTIVES ON THE THESIS FRAMEWORK

In this chapter, we reviewed the various cues that can affect auditory distance perception. Numerous topics and stated findings affected the formulation of this thesis's motivations and methods.

In Section 3.1, we have seen that auditory distance perception can be modeled with a two-variable compressive power function. In the experiments reported in **Part III**, this type of function is fitted to distance reports' data sets to get a comprehensive representation of the compression effect within them. It is applied to distance reports of all participants on a single condition to quantify the impact of the condition on the compression effect. It is also applied to individual results, in order to get an insight into individual acoustic cues weighting strategies.

In Section 3.2.1 we listed the different acoustic cues involved in auditory distance perception. As we mentioned in this section, the nature of reverberation-related distance cues and their relative perceptual weights for distance are still debated in the literature. It could be argued that it depends on a combination of an early-to-late energy ratio, several spectral cues and spatial characteristics.

In Chapter 7, we seek to determine the optimal time window depth for the early reflections, in order to define a perceptually relevant early-to-late energy ratio. Several values are tested, and the best fit for the observed data is researched. Moreover, weighting mechanisms with the sound level are examined in Chapter 7 and Chapter 8, as well as the possibility of an environmental or individual dependency.

Dynamic cues are frequently discussed in the literature as a supplement to the cognitive process of auditory distance perception. Their presence is frequently beneficial to the accuracy of auditory distance perception. As a result, we concentrated on static scenarios in this work. We assume that if the condition over the perceived distance of virtual sound sources is met in static scenarios, it should also be the case in dynamic scenarios. Moreover, practical reasons linked to the use of remote participants prevented us from using motion tracking systems.

However, as presented in Chapter 2 summary, we centered our work on tools (SRIR) and distance rendering techniques (impulse response extrapolation) that could be easily adaptable to real-time applications, as required by the context of AAR.

In Section 3.2.2 the importance of non-acoustic cues was stated. Prior knowledge or familiarity with the nature of a stimulus can significantly influence the perception of its auditory distance. As a result, all subsequent experiments reported in Part III use speech signals as stimuli, as they are presumed to be equally familiar to all individuals. Swiss-German was used as the speech signal language. An exclusion criterion concerned participants who understood German, so participants could not concentrate on the semantic content of the sentence but only on its acoustic aspects.

In Section 3.3, the question of the relationship between externalization and auditory distance perception was reported. It was reported that externalization could hardly be considered as a continuum of auditory distance perception. However, it can be considered as a prerequisite for auditory distance perception and total in-head localization could lead to important biases in terms of perceived distance. In all experiments reported in Part III, externalization reports were not collected, but a final questionnaire ensured that participants did not experience complete in-head localization during each experiment. Participants reporting this phenomenon were considered as outliers.

Hence, it was important for the production of auditory stimuli to keep spatial aspects, enabling partly externalization. Most of the auditory stimuli were constructed from SRIR measurements (here, in 4th order Higher Order Ambisonics (HOA)) converted to binaural signal. Chapter 5 describes the method used to generate the binaural auditory stimuli used throughout this thesis. The individual aspect of HRTFs on auditory distance perception was proven to be insignificant. Additionally, logistic reasons due to the recruitment of remote participants prevented us from using personalized HRTFs. Consequently, a generic set of HRTF linked to a dummy head, was used.

4

VISUAL CONTRIBUTION TO AUDITORY DISTANCE PERCEPTION

In most AAR applications the user has access to visual information about his environment. In Section 4.1 a short review of the mechanism behind visual distance perception and its performance in terms of accuracy and variability compared to the auditory modality is given. Then, in Section 4.2 the impact of vision on auditory distance perception of virtual sound sources in AAR applications is presented.

4.1 THE SUPERIOR SPATIAL RESOLUTION OF VISION

The human normal eye's monocular vision field extends horizontally to around 60 degrees toward the nose and 100 degrees away from it [170]. This field extends vertically for a range of 150 degrees. The binocular region, which is formed by the intersection of the two monocular vision fields, extends horizontally to around 120 degrees. The light coming from this field is focused on light-sensitive cells called photoreceptors, by the various components of the eye operating as an optical system. Incident light is converted to neuronal information and sent to the brain for processing via the optic nerve. Spatial information is finally derived from the data delivered by the optic nerves of both eyes.

This section solely focuses on the mechanisms underlying visual depth perception, and its accuracy and variability performances when compared to auditory distance perception.

4.1.1 *General mechanisms of visual distance perception*

Similarly to auditory distance perception, visual distance perception relies on cues that can be classified as relative or absolute. Absolute cues do not require external information to be computed into distance, while relative cues need a priori knowledge about the item or surroundings to be considered absolute.

Cutting and Vishton [46] have proposed a review of the different cues involved in visual distance perception, and have indicated the relative importance of each cue as a function of the distance between the observer and the visual object (see Figure 4.1).

The different monocular cues (requiring only one eye to be perceived) involved in visual distance perception are:

- **Occlusion** allows to position the objects relatively in depth. Indeed, if an object masks another object, it is then easy for the observer to know which

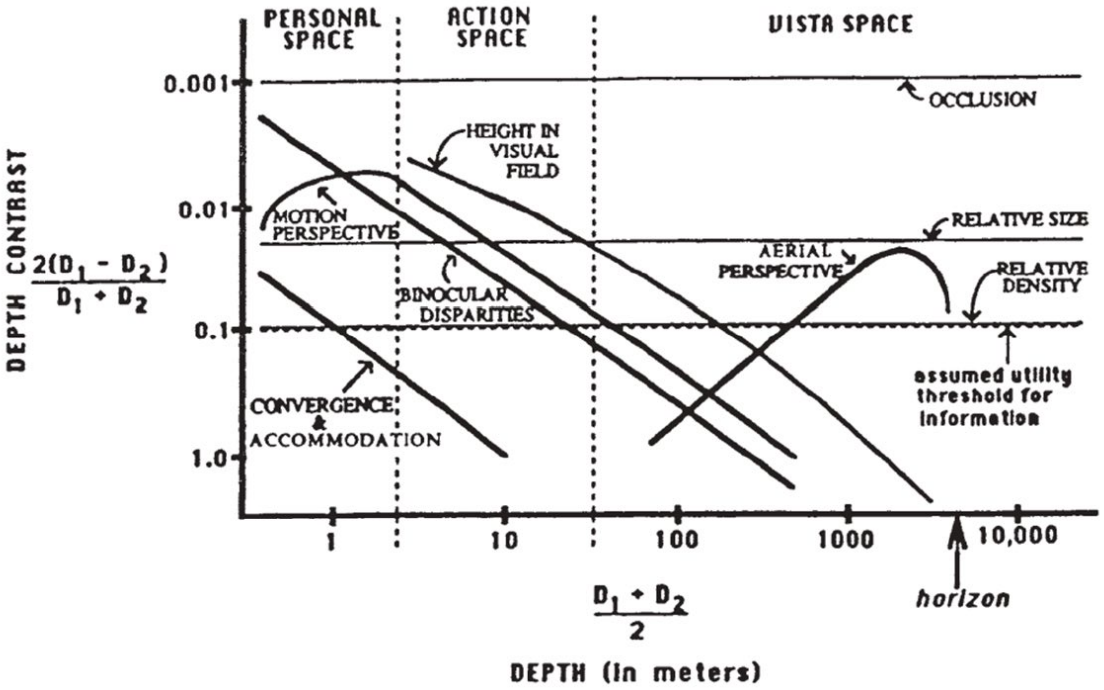


Figure 4.1: Just-discriminable depth thresholds as a function of the logarithmic distance from the observer. After Cutting and Vishton [46].

object is placed in front of the other. The occlusion is therefore a relative distance cue.

- **Height in the visual field** of an object relative to the horizon line allows one to estimate the distance in an absolute way. Indeed, the more an object is distant, the closer it gets to the horizon line.
- **Relative size and density** give information on the absolute distance of an object when its dimensions are known by the observer. In the opposite case, this monocular index only allows for relative distance judgments. The larger the retinal projection of a visual object, the closer the object is perceived.
- **Aerial perspective** can be seen as analogous to the role of air absorption for auditory distance perception. Here, aerial perspective refers to the filtering of the light due to the atmosphere. At large distances, visual objects' visibility and contrast decrease.
- **Accommodation** results from the deformation of the lens of the eyes. This adaptive phenomenon allows the eye to focus on near or distant objects and ensures the focus of the images projected on the retina. This monocular cue allows an absolute estimation of the distance of a visual object.

The binocular cues (requiring both eyes) involved in visual distance perception mainly result from:

- **Binocular disparities**, which are the differences between the projections of the light on both eyes. Each eye's fovea contains projections of the fixated object. When another object has projections that correspond to matching spots in the retina, the (angular) disparity is the angular distance between the projections and the fovea. It is self-evident that the difference increases with distance from the physical world's fixation point. Additionally, the sign of the discrepancy indicates whether an object is ahead of, or behind the fixation point.
- **Convergence** of the eyes' axis to the fixation point depends on the distance of the focused object. The closer the focused object gets, the more important the axis convergence becomes. This cue gives absolute distance information about the visual object.

Finally, when the observer or the object is moving relatively from one to another, dynamic cues are beneficial to visual distance perception:

- **Motion perspective** refers to the perception of changes in the projection due to motion. These changes can be interpreted as a relative movement of the object with respect to the observer. It gives relative information on the object's distance to the observer.

4.1.2 *Visual distance estimates*

In contrast with auditory distance perception, humans can evaluate visual distances in real environments accurately. The power function $D = k * d^a$ (see Section 3.1), can be used to link absolute perceived visual distances to real distances. For visual distance perception, in contrast to auditory distance perception, the a and k compression coefficients are close to a value of 1, being respectively slightly less than 1 and greater than 1. Reviews from Cook [43] and Da Silva [47] report a mean exponent $a = 0.9$ for visual distance perception, which approximates a fairly small compression effect.

Anderson and Zahorik [7] compared visual and auditory-only distance judgements. Auditory stimuli consisted in Gaussian noise reproduced on headphones through convolution with non-individualized BRIRs. Visual stimuli consisted of photographs of a loudspeaker displayed on a high-quality large screen. Despite the biases introduced by the reproduction method of the visual and auditory modality, their findings show that there are significant differences in terms of accuracy and variability between the visual and auditory distance perceptions.

4.2 AUDIO-VISUAL INTEGRATION

In the previous sections, we have listed the different mechanisms contributing to visual and auditory distance perception. However, the evaluation of the distance to a sound source in the real world or AAR scenarios does not rely only on the auditory sensory modality and is influenced by vision.

Two types of vision influence on audition can be distinguished. Either the sound source is associated with a visual object and therefore, stimuli from both sensory modalities (vision and audition) have to be integrated as a single percept. This "direct" influence of source-related visual cues is first discussed in Section 4.2.1.

In AAR scenarios, virtual sound sources are not necessarily linked to a specific visual object. In that case, vision does not directly affects the auditory distance perception of the source. However, the possibility for the user to see its environment can influence the auditory distance perception. This is discussed in Section 4.2.2.

4.2.1 *Ventriloquist effect*

One well-known consequence of the simultaneous presentation of auditory and visual stimuli is the "ventriloquist" effect. Even if the auditory stimulus is spatially incongruent with the visual stimulus, both stimuli can be perceived as a single percept.

Gardner [62] was one of the first to effectively assess the existence of a ventriloquist effect in distance. In his experiment, participants were placed in an anechoic room facing a line of 5 loudspeakers (from 1 to 9 meters) at eye-level, so they could only see the nearest speaker. When a speech signal was played on the furthest speaker at 9 meters, the participants always indicated the closest loudspeaker as the source of the sound. Here, the acoustic cues conveyed by the loudspeaker were particularly poor due to the anechoic situation and the arbitrary choice of the sound level. It is therefore not surprising to witness such a large integration in distance. Mershon et al. [118] conducted a similar experiment in an anechoic and reverberant environment and observed the same "image proximity effect". They observed, however, that in the case where the sound source was closer than the visual one, the depth of the integration window was reduced. More recently, studies have tried to define more precisely the limits of integration. Gorzel et al. [67] found a relatively large integration window in depth that tends to increase with the relative distance between the user and the source. Moreover, in this study, the quality of the sound reproduction tends to widen the window of multisensory integration. Both these effects can be explained by the maximum likelihood estimate model, in which the more variable the auditory distance perception is, the more likely it is to be integrated with a visual source as a unified percept [4].

4.2.1.1 *Maximum likelihood estimate model*

Welch and Warren [175] conceptualized a multisensory integration model. Among their different hypotheses, Welch and Warren proposed the sensory modality precision hypothesis, stating that a multisensorial percept is biased and spatially tends towards the most precise sensory modality. In that regard, Alais and Burr [4] later applied a maximum-likelihood estimate model to explain the ventriloquist effect. According to this model, the spatial localization m_{AV} of an audio-visual stimulus can be decomposed as:

$$m_{AV} = \frac{\sigma_V^2}{\sigma_V^2 + \sigma_A^2} m_A + \frac{\sigma_A^2}{\sigma_V^2 + \sigma_A^2} m_V \tag{2}$$

with m_V, the perceived distance of the visual-only stimulus, and m_A, the perceived distance of the auditory-only stimulus, as well as their associated variances, σ_A^2 and σ_V^2.

The resulting variance σ_{AV}^2 is defined as:

$$\sigma_{AV}^2 = \frac{\sigma_V^2 \sigma_A^2}{\sigma_V^2 + \sigma_A^2} \leqslant \min(\sigma_V^2, \sigma_A^2) \tag{3}$$

The model allows to predict, among other things, the distance perception of an audio-visual source. It illustrates that this prediction is linked to the reliability of each sensory modality. The model provides an explanation of the ventriloquist effect in distance, which is often translated as a bias of the overall perception towards the perceived visual location.

In the case of a more variable visual stimulus (e.g. smoke or haze degrading visual sensitivity), a reversed ventriloquist effect can occur. By artificially enhancing the noise of the visual modality, Alais and Burr [4] demonstrated that the overall audio-visual percept can significantly shift towards the auditory location. However, most often vision is significantly more accurate for spatial localization. For instance, when participants are required to make distance reports in auditory-only, visual-only, or audio-visual conditions, the audio-visual condition is very similar to the visual condition in terms of accuracy and variability, but significantly different from the auditory-only condition [7].

4.2.1.2 *Cognitive factors*

Cognitive factors can be beneficial to the integration of incongruent visual and auditory stimuli. Welch [174] argues that external information not linked to the structural differences (spatial and temporal) between visual and auditory stimuli can participate in their integration as a unified perception. External information could participate in the "unity assumption" driving the expectation of the listener that both stimuli come from the same source. The main factor that can strengthen the unity assumption is the familiarity with the audio-visual combination.

Stimuli that share a strong semantic link constitute a familiar combination. For example, a speech signal and a video of a person moving their lips coherently with the speech will induce a strong unity assumption (e.g. McGurk effect [114]). For example, a study demonstrated that auditory and visual information were successfully integrated despite a 200 milliseconds delay [167], far more than what was measured in previous experiments using noise and flash [184].

4.2.2 *Environment-related visual cues*

In AAR scenarios, the goal is not always to merge sound with a specific visual source, but to locate a sound source in the visual space at an intended position. In that case, vision and audition provide non-redundant information, preventing them from being directly merged as a unified perception. We intend here to make a review of the possible effects due to the vision of the environment that could influence the perception of virtual sound sources.

Vision can operate a calibration of auditory distance perception. Warren [171] first hypothesized the possibility implying that the vision of the environment could modify the way acoustic cues are interpreted into distance judgements. This sensory combination allows to collect non-redundant aspects of the observed environment to complement the auditory information.

This hypothesis was extended by Cabrera et al. [32], who argued the existence of a relationship between the visually perceived size of an environment and the auditory distance perception within the environment. According to them, if a sound source and a participant are in the same environment, then perceived distances are only possible within the limits of the visually perceived environment.

As seen in Section 4.1 visual distance perception is quite reliable for estimating distances, and thus has a great advantage over the auditory sensory modality for estimating the size of a closed environment. It could thus act as a reliable calibration of auditory distance perception if the assumption that the sound source is in the visual environment is respected.

Calcagno et al. [33] successfully illustrated the benefit of visual context information for auditory distance perception. They have demonstrated that participants who were given visual range information about the room were substantially more accurate in determining their distance to a hidden sound source. The greater the number of visual cues that were presented, the more accurate their reports were. The authors propose that the perceived distance of the sound source was calibrated by the visual information related to the size of the experimental room, calibration that was more reliable when the participants had complete vision of the room.

More broadly, the hypothesis that perceived room size calibrates the perceived distance agrees with the results over other combinations of sensory modalities. Etchemendy et al. [56] have registered a positive correlation between visual distance perception and the perceived size of the room through auditory cues. Kolarik [91] observed a correlation between auditory distance perception and audi-

tory room size perception.

In addition to the calibration, vision could also drive the expectation of the acoustical characteristics of a room. Calibration is not only influenced by the perception of the size of the room but also by the identification of certain absorbing materials and furniture, for example. An experiment by Sandvad [147] showed that participants could usually link photographs of rooms to reverberation time.

4.3 SUMMARY & PERSPECTIVES ON THE THESIS FRAMEWORK

In this chapter, we presented the potential impacts of visual information on auditory distance perception. In contrast with auditory distance perception, vision gives a very reliable estimation of distance.

This reliability can be beneficial for auditory distance perception of virtual sound sources in two ways: determining the distance of audio-visual sound sources and using the visual modality as a calibration tool for auditory distance perception. From a technical standpoint, linking a visual source with an audio signal enables a degree of relative tolerance in the replication of auditory distance. As the main motivation of this thesis is to evaluate the prerequisite for distance rendering methods linked to auditory distance perception, we chose to omit the investigation of auditory-visual integration objects in AAR scenarios .

However, visual calibration effects cannot be neglected for AAR. The auditory distance perception of virtual sound sources is not only affected by the reproduction of the acoustic cues but also by the visual environment of the user. In AAR scenarios, the acoustic environment may not be accurately characterized and/or faithfully reproduced, resulting in a possible incongruence between the visual and acoustic perception. Thus, this visual incongruence could modulate the auditory distance perception of virtual sound sources.

The experiment reported in Chapter 8, is dedicated to the influence of the visual environment on the auditory distance perception of virtual sound sources.

Part II

METHODS

PART II - INTRODUCTION

The second part of the thesis introduces the common methods used throughout the experiments of part III.

Chapter 5 focuses on the techniques employed to generate the different stimuli of the reported experiments. In all experiments, BRIRs are used as initial measurements for distance rendering methods or to create reference stimuli for perceptual performance comparisons. This chapter describes how BRIRs, usable for convolution, are converted from initial SRIR measurements. The technical solutions adopted for each of the sub-processes presented are referred among other existing alternatives.

Chapter 6 introduces two different aspects common to the experimental procedures applied throughout this thesis. First, the existing reporting methods commonly used in auditory distance studies are listed. The different advantages and drawbacks of these methods are discussed, and our choice among these methods is presented. Second, an overview of existing tools and the steps needed to design online experiments is given. This overview presents the specific tools we applied to conceptualize the online-based experiments reported in Part III. The different advantages and drawbacks linked to online approaches are discussed in relation to typical lab-based procedures.

5

BINAURAL RENDERING APPROACH OF VIRTUAL SOUND SOURCES

This chapter describes how the sound material evaluated in the thesis experiments was conceived, from the measurements procedure to denoised and usable BRIRs for convolution. The many stages of the process are reported in Figure 5.1.

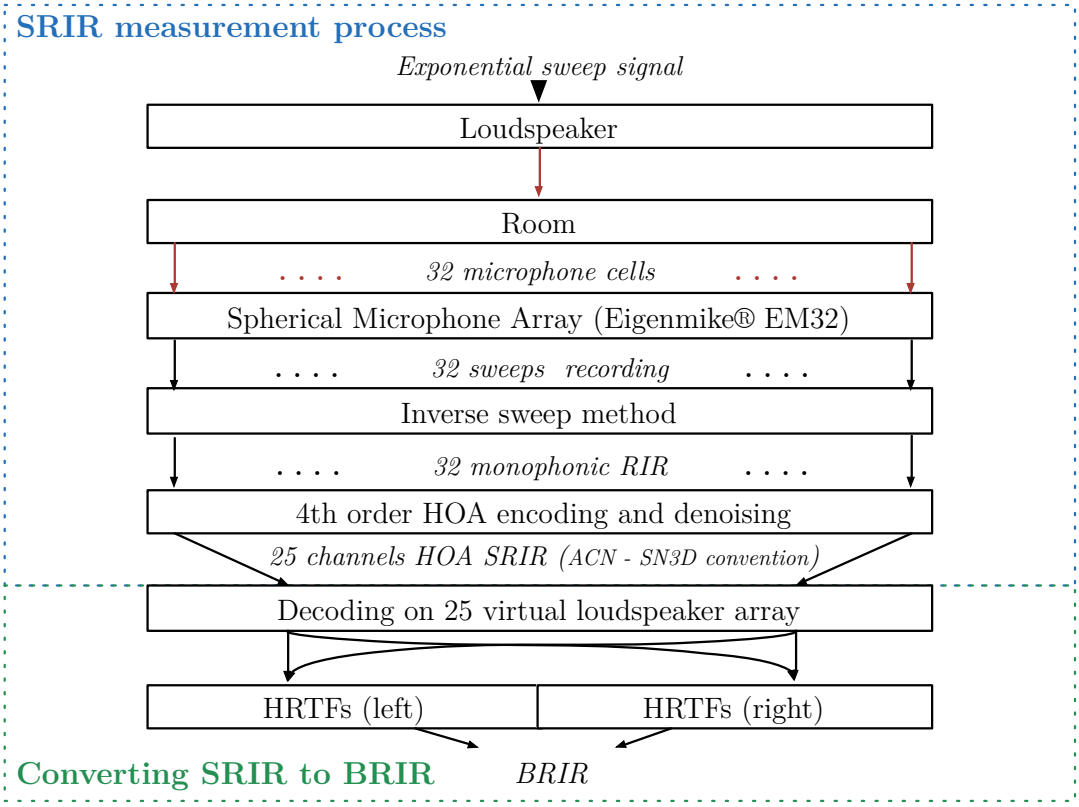


Figure 5.1: Schematic flow of the different processes involved in the measurements of SRIRs (framed in blue) and their conversion to BRIRs (framed in green).

Section 5.1 introduces the notion of SRIR and discusses why this type of measurement was chosen. It proceeds to describe the measurement technique and the tools used to carry it out. Then, the procedure used to convert SRIR to BRIR is described in Section 5.2. A presentation of the HOA concept is given, as well as a review of known decoding approaches. Finally, two specific treatments of the overall

method used here are presented: SRIR denoising and the diffuse-field equalization technique.

5.1 SPATIAL ROOM IMPULSE RESPONSES

As seen in Chapter 2, RIRs are commonly used to measure and reproduce the room effect for a given source-receiver position and orientation. It is a compact way of characterizing the acoustic properties of an environment. Typical methods use mono or stereo microphones to measure it. Therefore, it neglects spatial features of room acoustics, such as the directional of arrival of reflections. This limitation can be tackled by using multiple measurements [2] or microphone arrays for a more in-depth look into the acoustics of a room.

The use of Spherical Microphone Array (SMA) enables the measurement of Spatial Room Impulse Responses (SRIRs) on a single receiver point. The SMA spatially samples the sound field through each of its Q transducers. It consists in recording simultaneously the impulse responses on the Q transducers generally distributed on the surface of a rigid or open sphere [80]. The resulting recording consists of Q channels, each containing a monophonic RIR. Compared to conventional monophonic RIRs, such measurements extend considerably the possibilities of applications: spatial audio synthesis, spatial room acoustics parameters analysis and spatial modification of the resulting SRIRs.

Owing to the manipulation potential of impulse response's properties and of the information they contain, SRIRs were exploited to generate Binaural Room Impulse Responses (BRIRs) for multiple source-receiver positions with varying frontal distances.

5.1.1 *Measurement Procedure*

Figure 5.1 illustrates the general measurement procedure of SRIRs. The so-called swept-sine recording technique, and subsequent inverse-sweep convolution [58] were used to measure all SRIRs. It consists of sending a sine-sweep signal from 20 to 20000Hz to a loudspeaker through a sound interface, and recording the resulting signals from the different transducers of the SMAs. Generally, multiple successive sweeps are recorded for a given source-receiver position. As mentioned by Farina [58], averaging multiple repetitions of sweep allows reducing artifacts in the final response due to the occurrence of unwanted sound events and background noise during the recording. In order to avoid any time aliasing errors, the length of the silence after each sweep signal period must be longer than the room response. Finally, an inverse-sweep convolution is applied to each channel of the recorded signal to obtain a 32 monophonic RIRs. The swept-sine method is commonly used as it rejects the harmonic distortion artefacts prior to the causal part of the impulse response, which corresponds to the linear contribution. It provides an excellent signal to noise ratio [156] when compared to the Maximum

Length Sequence (MLS)[149], the Inverse Repeated Sequence (IRS)[51], or to the Time-Stretched Pulse[10] impulse response recording method.

5.1.2 *Used tools*

Throughout all the work of this thesis, the *mh acoustics Eigenmike©EM32*, a spherical microphone array with 32 transducers, was used as the SRIRs acquisition system. An *Amadeus PMX5* amplified speaker was used as the sound source. A *RME Fireface UC* was used as the sound interface. *IRCAM Spatialisateur* was the software used to generate the logarithmic sine sweep and the multi-channel recording. Apart from the recording of raw incoming signals, the software handles the deconvolution of recorded sine-sweeps into impulse responses.

5.2 CONVERTING DIRECTIONAL ROOM IMPULSE RESPONSES TO BINAURAL ROOM IMPULSE RESPONSES

5.2.1 *Encoding into Higher Order Ambisonics (HOA)*

In the 1970s, Gerzon [63] developed ambisonics, a sonic theory based on the spatial decomposition of a soundfield into a succession of spherical harmonics. The decomposition allows rendering 3D sound fields in a flexible manner based on knowledge of their first order directive information at one point: omnidirective (W) and 3D bidirective (X, Y, Z) components constitute the so-called B-Format. A narrow listening region is limited by its low spatial resolution, especially for high frequencies. An extension called "higher order ambisonics (HOA)" enables the decomposition to be extended to a greater resolution by using higher order spherical harmonics [142, 143, 186]. The use of a Spherical Microphone Array with multiple transducers provides access to HOA, and ultimately, the storage of a recorded 3D sound field in the Spherical Harmonics Domain (SHD).

Spherical harmonics consist of a complete set of orthogonal functions along a spherical surface. Thus, the Spherical Harmonics Domain(SHD) is a logical representation for a recording made on a spherical microphone array. The representation in the SHD can be written as follows, with $X_{l,m}$ the coefficients for each component of the spherical harmonic $Y_{l,m}(\Omega)$ of order $l \in \mathbb{Z}^*$ and degree $m \in [-l, l]$:

$$X_{l,m} = \int_{\Omega \in S^2} x(f, \Omega, t) Y_{l,m}(\Omega)\, d\Omega \tag{4}$$

Where, $\Omega(\Theta, \Phi)$ is the angular position on the sphere at a fixed radius $r = a$, in spherical coordinates. $x(f, \Omega, t)$ is a time-frequency domain representation of the sound field on the sphere. This equation describes the ambisonics transformation using ACN channel ordering, following the *Ambix* convention.[128] The SHD decomposition of Equation 4 has a theoretically infinite number of terms as order

can be infinite. In practice, the decomposition of a spherical microphone array signal is limited to a certain order L. The Ambisonics order L is determined by the number of transducers Q on the microphone array. Actually, we can estimate HOA signals up to a restricted order L so the total number $(L+1)^2$ of components does not exceed the number Q of sensors [48].

The integral of Equation 4 can be discretized and approximated by a weighted sum over the microphone position on the spherical array. The discretization can be written:

$$x_{SHD}(f,t) = Y * x(f,t) \tag{5}$$

Here $x(f,t)$ is the column vector containing a time-frequency domain representation of the signal measured at each of $\Omega(q)$ transducer's angular positions. Y is the encoding matrix of $(L+1)^2 * Q$ elements $y_{q,n} = \alpha_q * Y_{l,m}(\Omega(q))$, with $n = l^2 + l + m + 1$ and α_q the weight estimated from the approximation of the integral in Equation 4. $x_{SHD}(f,t)$ is the column vector of the SHD coefficients X_n. Finally, a correction for the SMA's so-called mode strengths (or holographic functions) must be added in order to get an array-independent representation of the observed sound field. According to Daniel and Moreau [48], this is the case with the higher-order ambisonics format, which uses the center of the sphere as a reference point and for which the correction filters are determined.

In our case, a spherical 32-microphone array was used, and the encoding of the recorded signal was of the 4th HOA order.

5.2.2 *Decoding HOA to the binaural format*

In order to decode the HOA signal into a binaural format, a convenient way is to use the so-called the so-called virtual speakers paradigm. It consists of decoding an HOA signal on an array of virtual loudspeakers. Corresponding to the position of each virtual speaker, associated HRIRs for both ears are computed. Each of them is convoluted with that speaker feed, and the convolution products for each of the ears are then summed up, giving the binaural signal for each ear. In this thesis, generic HRTFs measured on a dummy head *Neumann KU100* were employed.

Numerous techniques have been developed to transcode an HOA signal onto a loudspeaker array [187]: Direct-Sampling, also referred to as the Sampling Ambisonic Decoder, the Energy-Preserving Ambisonic decoder, Mode-Matching decoder...

Their goal is to transcode HOA signals on regular or irregular speaker arrays. Their advantages and drawbacks in different loudspeaker configurations are compared by Zotter et al. in [187]. In the case of the virtual speakers approach, most constraints linked to the configuration of the speakers array and the position of the ears where the signal must be reproduced are easily overcome. Therefore, a uniform layout of speakers, using equal area partitioning [99], was applied. In this

type of configuration, all methods are equivalent in terms of performance. In our case, the energy-preserving approach[188] was used.

For the decoding of HOA signal the minimal number of speakers is equal to $(L+1)^2$ with L the ambisonic order. In our case, as the Spherical Microphone Array limited the signal to the 4th order, an array of 25 virtual loudspeakers was employed.

5.3 SPECIFIC TREATMENTS

As mentioned earlier, the use of SRIRs enables the use of a large panel of manipulations and the extraction of room acoustics information. However, the method used to generate binaural signals comes with some drawbacks. Firstly, RIRs measurements made with a microphone present a non-negligible noise floor that can lead to a perceptible "infinite reverberation effect"[110]. Furthermore, it was intended to manipulate the energy of temporal segments of BRIRs to drive the perceived distance of convoluted stimuli. Such manipulations could lead to an amplification of the noise floor, and limit the usable dynamic range of the impulse responses. Therefore, a denoising process must be applied to the measured SRIRs. Secondly, audio processes can impact the spectral content of the impulse responses used to generate sound stimuli. The modification of the spectral content may lead to an additional bias in the auditory distance perception of sound sources created with these measurements. Thus, an equalization process must be integrated to compensate for these spectral differences.

5.3.1 *Denoising Spatial Room Impulse Responses*

A standard RIR can be divided into three parts: the direct sound, the early reflections, and the late reverberation. The latter is the part usually affected by the measurement's noise. As energy of the RIR decays with time it reaches at some point the "noise floor" induced by the measurements. The denoising process is thus based on a manipulation of the late reverberation.

RIRs were initially modeled as an exponentially decaying stochastic process [150]. This assumption has been demonstrated to be appropriate with a high enough echo density and modal overlap [139]. The echo density condition is satisfied beyond a certain time limit in the RIR and is qualified as the "mixing time". The modal condition density is fulfilled beyond a so-called Schroeder's frequency. The part of the RIR that is beyond both these limits can be qualified as the late reverberation field, and is considered as fully diffuse.

A diffuse sound field is a theoretical sound field in a volume V, that has a homogeneous energy density at all points of the volume with an equal probability of energy flow in all directions [49]. The diffuseness of a sound field is established through these three main characteristics: homogeneity, isotropy and incoherence. Homogeneity refers to the uniformity of the mean energy at all locations in the

volume. Isotropy signifies that, for a given point, the energy coming from all directions is equivalent. Incoherence implies that individual wavefronts are weakly correlated.

A zero-mean Gaussian noise filtered by an exponentially declining energy envelope can be used to create such a field [82]. Thus, zero-mean Gaussian noise filtered by a prolongation of the energy decay envelope can be used to replace the noise floor of measured RIR.

These different findings have driven the steps of the denoising process developed by Massé et al. [110]. Specific aspects of the method related to the denoising of anisotropic reverberation tail can be found in [111]. The implementation of the process with the software *MatLab* includes the encoding of the raw SMA measurements into SRIRs. The denoising is performed in the SHD. First, an EDR analysis is run to detect the envelope of the reverberation tail. Then, an analysis of the mixing time is realized to detect the beginning of what is considered the late reverberation field. Finally, the energy decay envelope of the SRIR is used to parameterize a synthesized incoherent reverberation tail replacing the noise floor. The specific changes in the method induced by this type of reverberation tail is mentioned in the description of the sub-processes.

EDR ANALYSIS EDR is an extension of the Energy Decay Curve (EDC) introduced by Schroeder, in which the energetic envelope is characterized for multiple frequency bins. The analysis of this EDR consists in each frequency bin undergoing three main steps: The noise floor is determined by fitting a theoretical profile of a reverse-integrated constant power noise, a segmentation of the EDR is performed to ensure the further fitting procedure; Then, the exponential decay region of the curve is determined by allowing for some headroom above this noise floor (delimited by the value $P_{noise}(f)$ and $t_{lim}(f)$) and avoiding non exponentially-deceasing noise; Finally, by fitting an exponential decay model to the given region (starting at the segmentation point t_{lim}), the decay parameters are found.

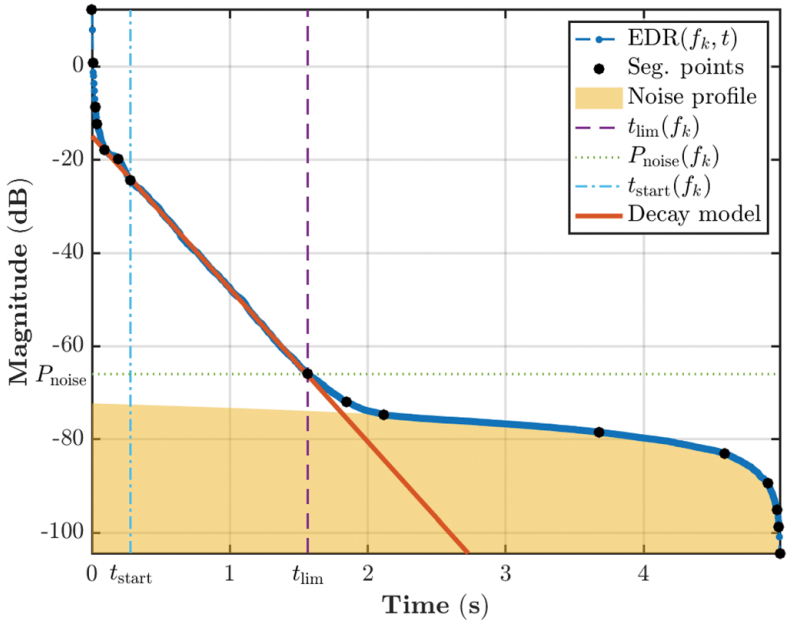
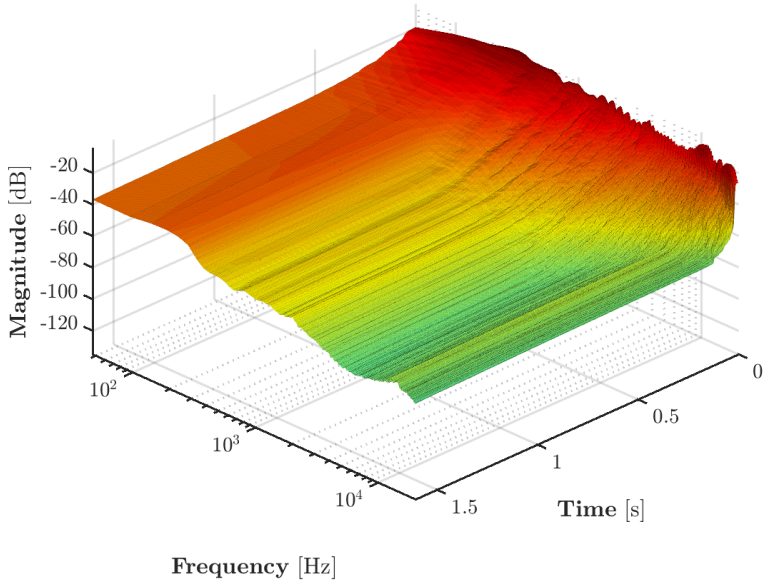


Figure 5.2: Schematic representation of an EDR analysis performed for a given frequency bin of an impulse response. The different segmentation points considered are represented (in black) with notably, the noise floor limit ($t_{lim}(f)$,$P_{noise}(f)$) and the beginning of the fitting of the decay model (in red) at $t_{start}(f)$. From Massé et al. [110].

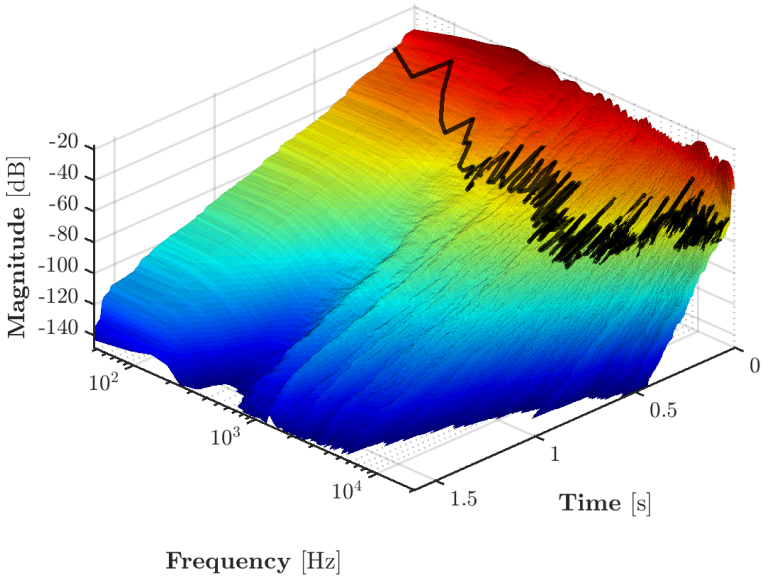
MIXING TIME ESTIMATION The goal of this sub-process is to detect the point at which the diffuseness of the impulse response is reached. The COMEDIE measure of diffuseness [55] is used in this approach, as it is well suited for SHD. It detects the time limit after which, in a given frequency bin, the evolution of the incoherence reaches its maximum value. It must be verified that this mixing time is located before the starting point of the noise floor. To this end, a weighted average value of the $t_{lim}(f)$ is defined as a global value for the noise floor limiting time. In the case of an anisotropic reverberation tail, the noise floor might be reached at different times depending on the direction. Therefore, this analysis is run in a direction-dependent manner through a Plane-Wave Decomposition (PWD) instead of running it in the SHD.

INCOHERENT TAIL SYNTHESIS As mentioned earlier, late reverberation can be synthesized as a zero-mean Gaussian noise. Here it is performed in the SHD. It is applied to each SHD component individually. Due to the orthogonality of the spherical harmonics and the spatial independence of plane waves, it can be demonstrated that the combination of a zero-mean Gaussian noise per SHD component produces a global incoherent reverberant field for the resulting SRIR. In the case of an anisotropic reverberation tail, the denosining process is run after a PWD and encoded back to the SHD. An example of the denoising process on a single SRIR is

displayed in Figure 5.3.



(a) EDR of the omnidirectional component of an SRIR before the denoising process.



(b) EDR of the omnidirectional component of the same SRIR after the denoising process.

Figure 5.3: EDRs of the omnidirectional component of an SRIR measured in the *Gallery* room at IRCAM (see Figure 7.8) before (up) and after (down) the denoising process. The black line shows the t_{lim} value for each frequency bin.

5.3.2 *Diffuse field equalization*

The nature of the microphone array used, HOA encoding and decoding, introduce spectral changes to the resulting binaural signal when compared to a direct BRIR measurement. In order to avoid perceptual biases caused by the spectral changes of the "binauralization" process, we applied a diffuse field equalization process. Contrarily, a direct sound (or free-field) equalization aims to compensate for the spectral coloration of the direct sound. As a result, understanding that it would have been more difficult to tackle the equalization of direct sound and diffuse sound separately, only a diffuse-field equalization process was applied to SRIRs.

During the denoising processes of the different SRIRs and BRIRs measurements used to create auditory stimuli, several parameters are calculated[110]. Among them, the initial power spectrum $P_0(f)$ of the impulse response late reverberation tail is calculated, where f is a frequency bin.

As explained in Section 5.3.1, in a stochastic model, the late reverberation field can be considered as diffuse, meaning it is isotropic and independent of the receiver and source positions. It is therefore only characterized by the initial power spectrum $P_0(f)$ of the late reverberation tail and a decay coefficients $\delta(f)$, related to the reverberation time as $T_{60}(f) = 3\ln(10)/\delta(f)$. The reverberation time is considered independent of the source and receiver positions, so the initial power spectrum must share the same properties as the late reverberation.

According to Jot et al. [82] the initial power spectrum can be expressed:

$$P_0(f) = \frac{\rho_0 c^2}{V} W(f) R_d^2(f) \tag{6}$$

where, ρ_0 is the density of the air, c the celerity of sound in the air, $W(f)$ the energy supplied by the sound source, and $R_d(f)$ the diffuse field sensitivity of the microphone's transducer. Therefore, if we consider measurements made in the same room at any given source-receiver position, with the same loudspeaker, but just differing by the microphone used, an equalization based on the respective value of $P_0(f)$ can be made. A filtering can be applied to a SRIR measurement so the resulting initial power spectrum $P_0(f)$ is equal to the one computed on a BRIR measurement. This so-called diffuse field equalization method enables the possibility to strip out the sound coloration of binauralized SRIRs and equalize it with BRIRs measured on a dummy head. A *Neumann KU100* dummy head was used for BRIRs measurements.

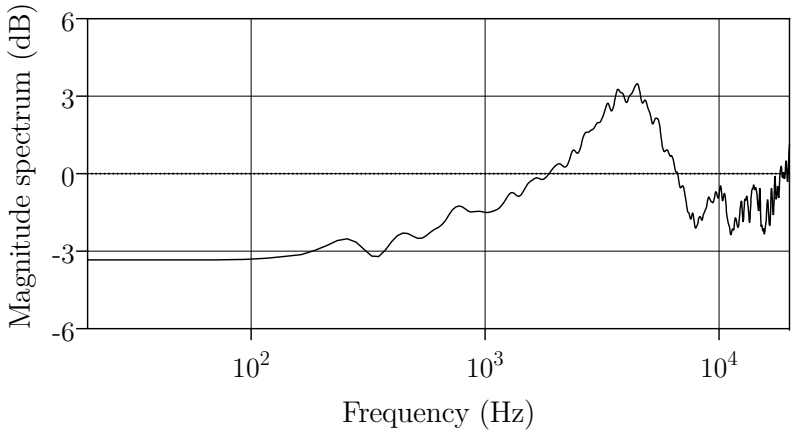


Figure 5.4: Power spectrum of the filter used for the diffuse field equalization process, applied to SRIRs measured with an *mh acoustics Eigenmike©EM32* and based on BRIRs measured with a *Neumann KU100* dummy head.

In our case, for the measured SRIRs, the value of the initial power spectrum was extracted from the omnidirectional component. While for the BRIRs, the mean value of $P_0(f)$ between the left and right channels was used. A single filter was averaged on SRIRs and BRIRs measurements. This equalization filter was systematically applied on every SRIRs employed in the conception of auditory stimuli used in following experiments. Its power spectrum is displayed in Figure 5.4

Ideally, the extraction of the $P_0(f)$ should have been made on the BRIRs converted from SRIRs, as the virtual speakers paradigm might also induce spectral changes. The performance of this process was first assessed in a preliminary perceptual experiment reported in Appendix a.

5.4 MEASUREMENTS USAGE IN THE EXPERIMENTS

Chapter 7 and Chapter 8 introduce three experiments using a common type of distance rendering method, illustrated in Figure 5.5. Initial SRIRs measurements are converted to BRIRs. One of these BRIR is used as input for a distance rendering method. In these different experiments, we seek to assess the importance of different acoustic cues, which are reproduced with varying degrees of accuracy by these methods. The perceptual performances are then compared to a reference method based on actual measurements in order to assess the influence of their differences in acoustic cues reproduction. The common objective of these experiments is to define the prerequisites for this type of distance rendering method in static AAR scenarios.

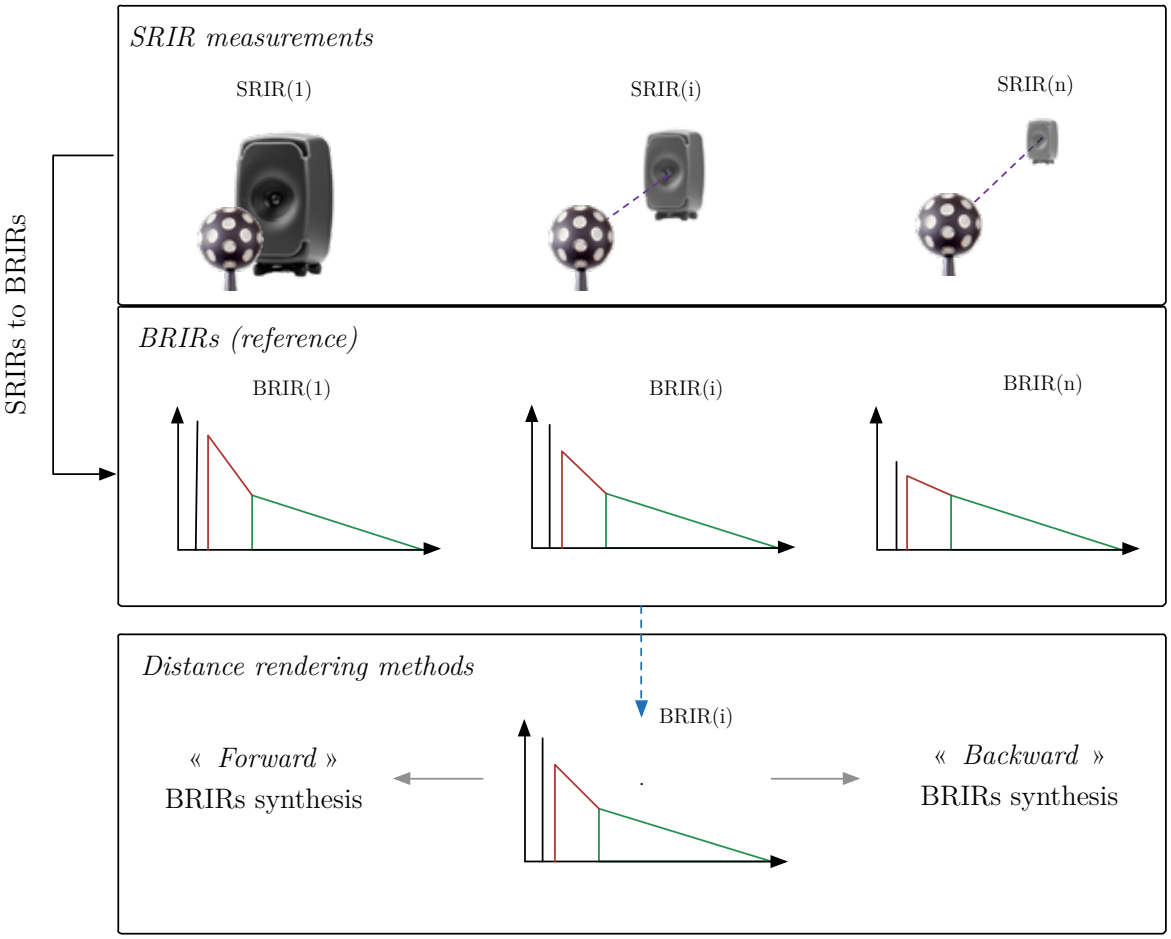


Figure 5.5: Schematic breakdown of the use of SRIRs measurements in Experiment I to Experiment III. Initial measured SRIRs are converted into BRIRs which will serve as reference. One of the BRIRs is then transformed by an extrapolation method-/model to synthesize several BRIRs corresponding to other distances. An extrapolation towards distances shorter than the initial measurement position is labeled as "*Forward*" synthesis in the manuscript. An extrapolation towards distances longer than the initial measurement position is labeled as "*Backward*" synthesis.

6

EXPERIMENTAL PROCEDURE

In this chapter, the different methodological choices that were made in the experiments are explained. A brief review of commonly employed report methods for auditory distance perception studies in the far field is presented. The report method used in the experiments listed from Chapter 7 to Chapter 9 is justified in the view of its benefits and drawbacks in Section 6.1.

In Section 6.2 the tools used for the design of online experiments are presented. The advantages and disadvantages of online compared to lab-based experiments and their impacts on the quality of collected data are discussed.

6.1 DISTANCE REPORT METHODS

In most of the studies investigating auditory distance perception, participants show large biases and high variability in their judgements. Close sound source distances (generally closer than 2 meters) tend to be overestimated, while far sound source distances are underestimated [2]. The compression of reported distances, nonetheless, greatly varies from one study to another. This variability can be partly interpreted as a consequence of the experimental conditions (type of stimuli, sound rendering method, visual context, etc.), but also by the methodological choices and, more particularly, the report methods employed[8, 57].We list different methods used in the literature, the biases they can create and how they may impact the results' analysis. These methods can be classified into three categories: verbal report, direct-location method, and spatial representation method.

Finally, we detail and justify the report method adopted in the experiments conducted during this thesis. Additionally, we discuss how this choice was driven by the benefits associated with this approach.

6.1.1 *Verbal report*

The most commonly used report method in auditory distance perception studies is the verbal report. It consists in asking the participant to report the distance of a sound source with an explicit distance scale (e.g. in meters or feet). This method is mainly used because it avoids practical limitations. The task is relatively easy to understand by participants compared to direct-location tasks and is not time-consuming. Hence, it allows a large number of reports within a fixed time session. Additionally, the verbal report method is easy to implement as it does not require any specific materials.

Verbal report has a great advantage when compared to other methods as it is not limited by the boundaries of the environment in which the task is performed. Hence, responses are not distorted or biased by a floor and ceiling effect. This advantage makes verbal report an appropriate choice for auditory distance perception investigations in virtual environments, since it allows virtual sources to be tested at great distances without requiring an equally large real environment. It has also been demonstrated to be a precise method, in comparison to the direct-location method, for estimating distances under certain experimental conditions, notably for sources at a distance further than 3 meters and closer than 5 meters[138]. The presence of visual cues is also beneficial for the verbal report method, as participants more easily associate visual distance with a scale in meters [33].

However, outside of these particular circumstances, verbal report of distances is generally less accurate than other direct-location methods[8, 57]. It also presents higher variability. The question of high inter-subject variability can also be raised as the spatial representation of an explicit scale may vary from a participant to another.

6.1.2 *Direct-location*

Direct-location methods designate reporting tasks in which the subject performs an action to indicate the perceived distance. These actions in far field auditory distance perception are often motor-associated (e.g. walking with eyes covered or uncovered) or vision-associated (e.g. pointing with a laser).

Comparative studies have demonstrated that responses relying on vision [57] and motor control [8] are generally more precise than verbal reports. Brungart et al. [29] [29] also compared these methods in the near field, obtaining similar results, with direct-location methods showing smaller bias and variability. They claimed that the direct-location approach appears to be a more natural response because no mental alteration of the target position is necessary, and individuals can determine the target's location using their own anatomical reference points. However, even if this type of report is considered more natural, one of the problems is that it calls for the use of another sensory modality (vision, proprioception...), introducing possible biases through cross-modal interactions (see Chapter 4).

Direct-location methods are rarely adopted for distance perception studies in the far-field. In the far field, the distances that need to be estimated are out of hand reach, thus the report is time-consuming and lessens the advantage of their consistencies. They also imply having access to a large space to perform them, raising important logistical challenges.

Etchemendy et al. [57] recently experimented with a method qualified as "Cross-modal direct location" bypassing some of these practical drawbacks. In a study comparing the latter report method to verbal report, distance judgements were reported using a visual marker whose position was piloted by a participant using a hand held control. This method quickened the response task, demonstrated fewer

biases, and increased accuracy when compared to verbal reports. Nonetheless, it still rests on specialized equipment, programming, and logistics.

6.1.3 *Selected method: the Visual Analogue Scale (VAS)*

The choice of a report method is often driven by practical reasons (duration and complexity for the participant) and by scientific reasons (stability, presence of cross-modal interactions, stability ...). Each method possesses advantages and drawbacks in both fields. The lack of consistency over the methodologies employed to gauge listeners' responses is a major flaw in auditory distance perception research. The outcomes of multiple studies are difficult to compare because of the methodological heterogeneity. Unifying the criteria used to evaluate auditory distance perception would be a significant step forward in the comparison of the findings. Our choice was motivated by a method that needs little logistical effort and is easily reproducible. In all three experiments, participants reported their auditory distance judgements on a Visual Analogue Scale (VAS) with no explicit scaling. This method has several advantages:

- Similarly to the verbal report method, the time needed between each response is short.
- It avoids any bias due to personal representation of explicit scales.
- It is easily reproducible with most Graphical User Interface builders

However, the main drawback of this type of method is the presence of a floor and ceiling effect. Moreover, likewise Likert scale, it may introduce a central tendency bias in the responses if no explicit instructions are given relative to the minimum and maximum position of the scale.

6.2 ONLINE EXPERIMENT METHODOLOGY

In March 2020, the situation due to the COVID-19 pandemic was a major setback for the development of lab-based experiments. Until July 2020, and during different epidemic outbreaks throughout the year, all experiments with participants in the laboratory were strictly excluded. Afterwards, the resumption of the research protocols was possible provided compliance with the health and safety instructions established for the management of the COVID-19 epidemic, and that inclusion was limited to only those persons who are not categorized as being "at risk".

During this thesis, the first lockdown (March 2020) happened two months after the analysis of the results of Experiment I. Instead of losing time in logistics in order to comply with the health instructions, we chose to start running experiments online. This had an impact on the research goal of the thesis and required the comprehension of tools specific to this type of experiment.

In this section we introduce which tools were used to develop and share online experiments. Then we discuss how this type of format might impact data quality when compared to lab-based experiments.

6.2.1 *Technical aspects of online experiments*

Online data collection has begun to become a large part of the behavioral sciences methodology. For example, in the last three years, the number of participants and researchers on the online recruitment platform *Prolific* has reportedly expanded by more than sevenfold, with a substantial portion of this increase occurring during the outbreak of the COVID-19 pandemic in early 2020. However, conducting carefully controlled behavioral online experiments introduces a number of new technical and scientific challenges, from experiment design to online compatibility to participant recruitment. The procedure for running an online experiment can be decomposed as follows: a) programming the experiment on an appropriate software b) uploading the experiment to a compatible host platform c) recruiting study participants. The first two steps must be compatible in terms of programming language, while the integration of the second and third should respect the European General Data Protection Regulation (General Data Protection Regulation (GDPR)). These requirements were necessary for the experiment to be approved by the Research Ethic Committee of *Sorbonne-Université*[1] (approval number CER-2020-080).

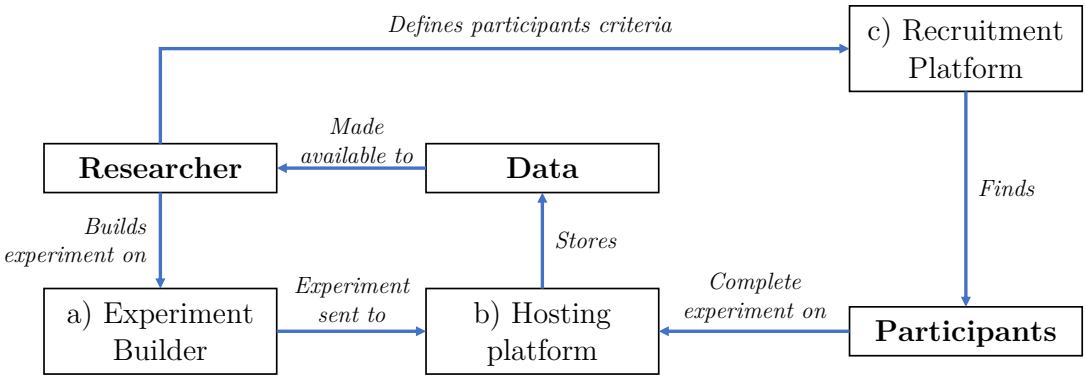


Figure 6.1: Schematic flow of the different processes involved in an online experiment as defined by Sauter et al. [148]. a) The experiment is conceived by a builder; b) The experiment is hosted on a platform's server; c) Participants are recruited following inclusion or exclusion criteria defined by the researcher. They run the experiment on the hosting platform, and their data are stored on the hosting platform servers and made available to the researcher only.

1 https://cer.sorbonne-universite.fr/

6.2.2 *Experiment Builder:* PsychoPy

As defined in the previous section, the first step in the development of an online experiment is the choice of a builder. Most of the softwares designed for online experiments are based on Javascript (JS), a well established language for interactive web page design. The main established open-source experiment builders are *lab.js*[2][76], *jsPsych*[3] *OpenSesame*[4][112], *PsychoPy*[5][136], and *PsyToolKit*[6][159]. Some of these softwares are more suited to be integrated with specific hosting platforms.

Other builders even offer the possibility to develop and host on a single ecosystem, enabling time savings and avoiding compatibility issues: *Gorilla.sc*[7][9], *Inquisit Web*[8], *Labvanced*[9], *Testable*[10]. Because it is generally not possible to change the code of the components, this type of solution lacks openness and flexibility. Due to this problem, this form of approach was ruled out.

Our choice of the software was also driven by the practical qualities offered for integrating audio in the experiment, and the possibility to integrate the experiment on a hosting service which does not require an institutional server. In the end, *PsychoPy* was the experiment builder that was kept. Moreover, *PsychoPy* offers the possibility to modify the base modules of its code with Python as it possesses a base-Python to JS translator.

6.2.3 *Hosting platform:* Pavlovia

As mentioned earlier, most of the experiment builders work with a JS backend. One possibility is to host the experiment on a lab-based specific server, which would require equipment that was not available in our case. Therefore, we directed our choice to centralized hosting providers. The goal of hosting experiments on these platforms is to make processes like user management, automatic data storage, and the generation of unique participation linkages easier to handle. On the other hand, many of these services come at a cost, usually in the form of an annual license or a per-participant fee.

Considering the choice of using PsychoPy, the most adapted platform for its integration was *Pavlovia*. In each experiment, a per-participant fee of 0.3€ was paid to *Pavlovia*. Its operation guarantees the anonymity of the data collected, in accordance with the General Data Protection Regulation (RGPD). The data was initially stored on the servers of *Pavlovia*, managed by *Open Science Tools Ltd*. located in Eng-

2 https://lab.js.org/
3 https://www.jspsych.org/
4 https://osdoc.cogsci.nl/
5 https://www.psychopy.org/
6 https://www.psytoolkit.org/
7 https://gorilla.sc/
8 https://www.millisecond.com/products/inquisit6/weboverview.aspx/
9 https://www.labvanced.com/
10 https://www.testable.org/

land, before being downloaded on the local hard disk of a computer and erased from *Pavlovia* servers

6.2.4 *Recruiting participants:* Prolific

Because of some constraints on participants' attention specific to online experiments (see Section 6.2.5), the time of the procedure must be minimized. Consequently, to obtain an equivalent number of responses when compared to a lab-based experiment, it is necessary to recruit a larger number of participants. The process of recruiting participants (e.g. through institutional mailing-lists) was considered, but due to the higher number of participants than in lab-based studies, we decided to employ a specialized provider. Two platforms were considered: *Prolific*[11] and *Amazon Mturk*[12].

Different aspects led us to choose the first option: *Prolific*. First, the global integration of the builder *PsychoPy* with the hosting platform *Pavlovia* is recognized as a compatible ecosystem with GDPR, with no need for an external anonymization system for data and payments. Other researchers also reported having an overall good experience with this ecosystem [148]. Moreover, *Prolific* is focused towards behavioral research studies and offers adapted targeting audience tools. Finally, a comparative study of Peer et al. [135] demonstrated the advantage of *Prolific* on data quality concerning behavioral research.

The recruitment process begins once the study is made available online after the pre-screening. For experiments II, III and IV it generally took a few hours to recruit 120 participants. The data of each participant had, however, to be carefully checked and approved before the participants could receive their compensation.

6.2.5 *Data quality concerns*

A considerable strength of online studies is the possibility to recruit and collect data of a large number of participants very quickly, especially if a recruitment platform is used. They can be easily scaled to large pools of participants, as recruiting larger samples does not require a higher workload. Moreover, it allows precise pre-screening of participants through different inclusion criteria. [134] For example, in the online experiments presented in this thesis, participants could be included based on the characteristics of their listening environments. Lab-based experiments, on the other hand, may suffer from the use of non-representative samples of the population, according to[77].

However, unlike lab-based experiments, many concerns on data quality have to be taken into account when preparing an online experiment. Four different aspects can be drastically different in online experiments and impact data quality: attention, comprehension, honesty, and reliability [135]. In the current section, we

11 http://prolific.co
12 https://www.mturk.com/

present how these aspects can affect data quality and what measures were taken to limit their impact.

Attention refers to the extent to which participants are committed to the experiment. Studies have demonstrated that online experiments can induce decreased attention because there is no direct interaction between the experimenter and the participant [85]. As a result, participants may be less likely to pay attention solely to assist the experimenter with their research. As a result, explicitly stating the research's relevance should be considered to increase participants' attention. Moreover, paying an appropriate fee to participants is also necessary as it can affect the motivation of participants[45]. In our case, an average hour rate of 9£ was used. Finally, attention tests placed at certain moments during the experiment procedure can ensure that participants still pay attention to the instructions.

As interactions between participants and researchers are limited, instructions must remain simple and, therefore, complex tasks must be avoided. In *Prolific*, an online chat system is present. During the period of time while participants were doing the experiment, we were ready to answer or tackle any problem encountered by a participant.

Finally, to assure reliability and honesty, a pre-screening of participants was conducted. The following inclusion criteria were applied:

- Participants had a high approval rate on *Prolific* studies($> 95\%$) and a number submissions to previous studies superior to 10.
- Participants were requested as normal-hearing and with no uncorrected visual impairment.
- Participants had an age inferior to 55 years old, to reduce large differences in memory capacities among participants.

The following exclusion criteria were applied:

- Participants who understood the German or the Swiss-German language. All stimuli were based on an anechoic recording of a Swiss-German sentence pronounced by a male speaker. The choice of the language, unknown to all participants, was made to avoid them focusing on the semantic content of the sentence.
- Participants who took part in one of the other experiments.

In both experiments, we also chose to hold a debriefing session following the experiment, with each participant, to verify that the answers given during the proposed questionnaires were consistent. Moreover, different criteria were applied to identify possible outliers:

- Participants with a mean response time inferior to 1 second.

- Participants showing larger deviations than 2 standard deviations from the global means.
- Participants reporting a total internalization of the presented stimuli.

Part III

EVALUATIONS OF ACOUSTIC AND NON-ACOUSTIC CUES FOR AUDITORY DISTANCE PERCEPTION

PART III - INTRODUCTION

This part presents four experiments conducted to address the role of acoustic and non acoustic cues for auditory distance perception in AAR. The studies and methods followed the tenets of the Declaration of Helsinki, and informed consent was obtained from participants prior to data collection and after the presentation of an information note. The collection of online data was in compliance with the GDPR.

Chapter 7 reports and discusses the findings from two experiments labeled "Experiment I" (lab-based) and "Experiment II" (online-based). The first experiment examines the perceptual performances of two distance rendering models that reproduce some of the auditory distance cues conveyed by real measurements. The application of these rendering models enables the exploration of 2 distinct objectives: gaining an insight into acoustic cues weighting strategies applied by the participants and investigating the relevance of the early-to-late energy ratio as a distance cue. The second experiment focused on the relevance of the early-to-late ratio in two different listening environments. The importance of reverberation-related spectral aspects for auditory distance perception is also discussed.

Experiment I results were published in the proceedings of *Forum Acousticum 2020* conference [109].

Chapter 8 presents and discusses the results of an online-based experiment labeled "Experiment III". It evaluates the influence of visually incongruent environmental cues on auditory distance perception. Its objectives are twofold: to study the impact of a visual spatial boundary and of the room volume on auditory distance perception.

Experiment III results have been published in the special issue *Psychoacoustics for Extended Reality (XR)* of the journal *Applied Sciences* [108].

Chapter 9 presents and discusses the results of the last online-based experiment labeled "Experiment IV". This experiment aims to reproduce an AAR scenario where the acoustic environment of the user is not correctly reproduced, thus inducing a room divergence between real and virtual sound sources. It investigates the possibility of an intra-modal calibration effect caused by the divergent acoustic cues conveyed by real co-occurring sound sources on the auditory distance perception of virtual sound sources. In particular, it examines the importance of intensity and reverberation in this calibration effect.

Experiment IV results have been presented at the *ASA Meeting 2021* [107].

7

EVALUATIONS OF THE IMPORTANCE OF INTENSITY AND REVERBERATION

This chapter presents two experiments labeled as "Experiment I" and "Experiment II". Their objective is to evaluate the importance of primary acoustic cues for auditory distance perception: intensity and reverberation-related cues. Two distance rendering methods, based on the manipulation of the energetic envelop of an initial SRIR are evaluated in Experiment I. Its results demonstrate the possibility to render sound source distance with this approach. It also illustrates that, despite a similar listening situation, the perceptual weights attributed to intensity and reverberation-related cues differ from one individual to another. Given the findings of Experiment I, the relevance of the early-to-late energy ratio as a reverberation-related distance cue is further evaluated in Experiment II (from Section 7.6). Additional distance rendering methods correctly reproducing this cue with distance are evaluated in this online-based experiment. Each of them synthesizes BRIRs with different temporal distributions of the early energy, while correctly reproducing the global energy and the late reverberation. It investigates if the reproduction of the early-to-late energy ratio could encapsulate the role of reverberation for auditory distance perception. Different temporal limits between early and late energies are evaluated through these methods.

7.1 INTRODUCTION

As discussed in Part I, reproducing room effects using perceptually motivated approaches is particularly well suited to AAR applications. These methods attempt to replicate cues relevant to the spatial perception of auditory events. In this regard, a review of the various acoustic cues relevant for auditory distance perception was provided in Chapter 3. Two of them are considered to have a prominent role: intensity and the DRR. The combination of intensity and the DRR is acknowledged as the main factor driving the auditory distance perception of stationary sources. Zahorik [179] showed that the auditory system uses weighting strategies of intensity and of the DRR flexibly to produce a distance percept, depending on the characteristics of the listening situation.

Concerning the DRR, several studies have demonstrated that it is very unlikely that the auditory system can effectively separate direct sound from reverberation and process a strict DRR. That is why different studies suggested the definition of perceptually more relevant reverberation-related distance cues correlated to the DRR [27, 93, 97, 141]. Three cues have already been evaluated, an estimation of the perceived distance with an early-to-late energy power ratio [27], the interaural

coherence [141], and monaural changes in the spectral centroid or in frequency-to-frequency variability in the signal [97]. Kopco and Shinn-cunningham [93] determined that an early-to-late energy ratio could be a good candidate. Concerning the temporal limit between what is considered as the early energy, which includes the direct sound as well as early reflections, and the late reverberation, no clear conclusions have been drawn.

The objective of the two experiments presented in this chapter is twofold: 1) to investigate the perceptual relevance of the early-to-late energy ratio as a distance cue through the perceptual evaluations of different distance rendering methods 2) to assess the acoustic cues weighting strategies applied by participants to infer a distance judgement.

A first lab-based experiment was conducted with online listening tests in which participants had to evaluate the distance of virtual sound sources produced by different rendering methods. The choice of the rendering methods was made to design stimuli categories in which the availability and reproduction quality of acoustic cues are different. Two different distance rendering models have been designed and their perceptual evaluation contrasted by a method based on actual measurements of SRIRs.

A first model labeled "envelope-based model" was used to investigate if discrepancies between the simplified energetic envelope of the model and that of the measurements would influence distance perception. Its evaluation enabled investigating whether or not a portion of the early energy of an impulse response can be considered as fused with the direct sound.

A second model, labeled "intensity-based model," was designed to reproduce accurately intensity while maintaining reverberation-related cues constant with distance. The perceptual evaluation of this model provided insight into participants' weighting strategies on acoustic cues, including intensity and reverberation-related cues.

These two models and the reference method based on actual measurements were tested in a first experiment (denominated Experiment I in the manuscript) on 20 participants, in a room from which the models take a priori information.

7.2 EXPERIMENT I: DEVELOPMENT OF DISTANCE RENDERING MODELS

Two models have been designed to reproduce the distance of a virtual sound source in an acoustic environment, from which the models exploit a priori information given in the form of a single RIR. In experiment I, these models are applied to a specific room: a classroom at IRCAM, labeled as *Classroom* in the following (semi-damped, dimensions: 8.7m $\times$ 4.7m $\times$ 3.5m – L $\times$ W $\times$ H, T_{60} at 1kHz of 0.55s).

7.2.1 *Reference measurements*

Different SRIRs were measured in the *Classroom* with a spherical microphone array *Eigenmike©EM32*. Nine SRIRs were measured for distances ranging from 1 to 7m (1, 1.5, 2, 2.5, 3, 4, 5, 6, 7m) by changing the speaker position and with a single microphone position. The different measurements were performed on an axis shifted 60cm from the median line of the room to avoid spatial symmetry of the first lateral reflections. This precaution was taken to favor decorrelation of the resulting binaural stimuli, which also contributes to their externalisation [21]. These SRIRs were converted from a 4th order ambisonic signal to BRIRs (see Chapter 5). This set of 9 BRIRs is referred to as the "reference". The BRIR corresponding to a distance of 1 meter was used as a single initial impulse response by the two following distance rendering models.

7.2.2 *Envelope-based model*

This model is based on a simplified representation of the energy envelope of the impulse response, here divided into two temporal segments: the early energy E_s comprising both the direct sound and early reflections, and the late reverberation energy E_{rev}. The energies of these two temporal segments are modified according to the desired distance.

Different perspectives can be considered to demarcate the time limit between these two segments. This transition time can be derived from perceptual considerations, regardless of the room geometry. For instance, a time limit of 50 ms has been considered in room acoustics for criteria describing the quality of speech perception, such as *definition* (D50) or *clarity* (C50). In order to delineate the 'useful sound' in opposition to the 'detrimental sound', Lochner and Burger [102] consider a weighting function equal to 1 until 35ms and then linearly decreasing up to 95ms.

Alternatively, this transition time can refer to the physical properties of the room and of its impulse responses. After reaching a sufficiently high echo density and modal overlap, the room reverberation exhibits an exponentially decaying stochastic behavior. This lower time limit is referred to as the "mixing time". Several estimators of the mixing time have been suggested in the literature. The estimation $t_m = \sqrt{V}$ (with t_m the mixing time in milliseconds and V the volume of the room in cubic meter) was proposed in [140]. Other methods rely on the evaluation of the diffuseness of the sound field from the statistics of the echoes observed in the impulse response. This estimation may either be conducted in the time domain [1, 157], or in the spatial domain, when a SRIR is available [111]. In our case, a temporal estimation of the mixing time was chosen and provided a value of 15ms (after the onset delay). This time defines the temporal limit between what is considered as the early part of the BRIR and the reverberation tail.

The design of this model was based on the hypothesis that the auditory system could not effectively separate the direct sound from the early reflections to infer a

distance percept [93]. A part of the early reflections could perceptually fuse with the direct sound. The mixing time, separating the pattern of discrete reflections from diffuse reverberation, was chosen as a first estimation of the length of the integration window of early energy by the auditory system.

The model applied to alter the initial BRIR to control its apparent source distance is inspired by previous work from Jot et al. described in [82]. In their proposed approach, the sound source distance is driven through the control of two temporal segments, the direct sound energy E_{dir} and the reverberated energy E_{rev}. The level of the direct sound according to the source distance d is expressed as follows:

$$E_{dir}(f, d) = \frac{S_\phi(f)^2 \mu(f)^d}{4\pi c d^2} \tag{7}$$

with c the sound celerity, f the frequency, $\mu(f)$ the frequency dependent sound absorption for a 1-meter propagation in the air, and $S_\phi(f)$ the free field transfer function of the source in the direction of the receiver.

The level of the diffused part of the reverberation after the mixing time τ in the impulse response is expressed as follows:

$$E_{rev}(\tau, f) = \frac{Tr(f) S_d(f)}{13.81 * V} \exp(-13.81 * \frac{\tau + \frac{d}{c}}{Tr(f)}) \tag{8}$$

with $Tr(f)$ the reverberation time, V the volume of the room, $S_d(f)$ the product of the diffuse-field transfer functions of the source and the microphone. The dependence of the reverberation energy with the distance d agrees with Barron's revised theory on energy relations in the room response [16]. In the present experiment, some further simplifications are made. Air absorption is neglected when considered distances are small (< 15m). The spatial dependence of the free field transfer function of the source is ignored as it is always heard from its frontal direction. Moreover, the attenuation law of the direct sound is extended to the whole early reflection pattern, thus introducing discrepancies between the measured and the simulated early reflections of the extrapolated source positions. Under these assumptions, the modifications that are applied to the early and late segments of the initial impulse response measured at a distance d_{ref} to derive the new impulse response at distance d can be written, respectively, as follows:

$$E_s(d) = E_s(d_{ref}) * \frac{d_{ref}^2}{d^2} \tag{9}$$

$$E_{rev}(d, f) = E_{rev}(d_{ref}, f) * \exp(-13.81 * \frac{d - d_{ref}}{c Tr(f)}) \tag{10}$$

This model is referred to as the "envelope-based model". Applying Equation 9 and Equation 10, nine BRIRs were generated upon modification of the initial impulse response measured at $d_{ref} = 1$m to simulate the different distances selected in the *Classroom*.

7.2.3 *Intensity-based model*

This simplified model extrapolates the BRIRs corresponding to different distances by applying a global gain to the initial reference BRIR measured at 1m. The gain used to extrapolate the BRIR for a given distance was tuned so that the loudness of the resulting stimulus corresponded to the loudness of the stimulus generated with the BRIR measured at the same distance. The loudness criterion used here is EBU R128. This model is used in contrast to the envelope-based model and the rendering method based on measurements as no modification of the reverberation-related cues is present when distance increases.

This second model is expected to give insights into the participants' auditory cues weighting strategies. The assumption is that, participants who primarily rely on intensity to infer a distance judgment, should show similarities in their distance reports for stimuli created with this model and for stimuli generated with the reference measurements. In contrast, participants who primarily rely on reverberation-related cues should exhibit significant differences between distance reports associated with the intensity-based model and the measurements.

7.2.4 *Objective comparisons*

The differences between the BRIRs generated with the two models and the measurements recorded in the *Classroom* are displayed within the scope of the energy contained in different temporal segments of the impulse responses. Figure 7.1 depicts the energy contained in the early part E_S, preceding the considered mixing time, and in the reverberation E_{rev} of each impulse response generated by the models and of the measured ones. When compared to the measured impulse responses, the envelope-based model mainly underestimates the early energy E_s, a difference that increases for longer distances. This behavior comes from the 6dB attenuation law being applied to the whole early energy considered in this model, instead of the direct sound only. Contrarily, the late energy tends to be slightly overestimated as distance increases, with a maximum difference not exceeding 1dB. In summary, we can state that the envelope-based model reproduces correctly the energy contained in the late reverberation, but strongly underestimates a part of the energy of early reflections.

For the intensity model, the attenuation law used to tune the global gain of impulse responses produces opposite differences in terms of energy contained in the different segments of the generated BRIRs. As a global gain is applied to the entire impulse responses, the late energy is highly underestimated as distance increases, whereas the early energy tends to be slightly overestimated as distance increases, with a maximum difference not exceeding 1dB.

Both models show a slight underestimation of the total energy, but the difference does not exceed 2dB at its maximum. A value just above the just-noticeable difference threshold of intensity perception [81]. The attenuation law used for the intensity-based model was based on a loudness criterion (EBU R128) over the re-

sulting stimuli, so their loudness was equivalent to the stimuli generated with the reference measurements.

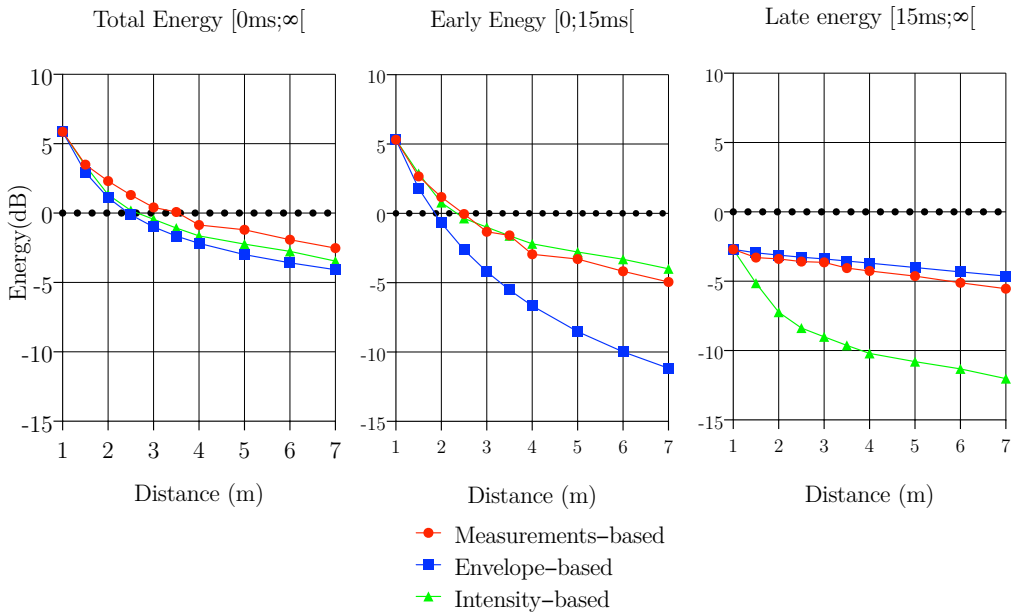


Figure 7.1: Evolution of the early energy E_s (left), reverberation E_{rev} (middle) and total energy E_{rev} (right) according to the source distance of the generated BRIRs for the two models and of the measured ones.

7.3 EXPERIMENT I: PERCEPTUAL PERFORMANCES OF THE MODELS IN A CONGRUENT SITUATION

The main goal of this first experiment is to have a first evaluation of the performance of the models in a simple AAR scenario. The listening environment is the same as the one used for the measurements. Therefore, the reproduction of the room effect was congruent with the visual environment. The envelope-based perceptual performances enable a preliminary assessment of the importance of accurately replicating the energy contained in the early part of BRIRs for the reproduction of sound source distance. The comparison of intensity-based model performances with reference measurements allows for the analysis of acoustic cues weighting strategies used by participants to infer distance judgements in this congruent situation.

7.3.1 *Material & Methods*

7.3.1.1 *Participants*

A total of 20 (8 women) participants, ages ranging from 19 to 35 (mean age: 26.05), took part in the experiment. Inclusion and exclusion criteria reported in Chapter 6

were applied. All the participants had little to no knowledge about audio processing or room acoustics. Informed written consent was signed by each participant before the experiment starts.

7.3.1.2 *Listening environment*

The listening tests were conducted in the *Classroom* used for the SRIRs measurements (see Figure 7.2). The participant was seated exactly at the place corresponding to the location of the SMA used for the measurements and was looking in the same direction. Twelve chairs were distributed every 45cm along a line facing the participant at 0° azimuth and starting 1.2m from the participant's position, in order to guide the visual distance perception. However, the participant was informed that the spatial distribution of the chairs did not correspond to the actual spatial distribution of the measured and modeled stimuli. A visual fixation cross was drawn at a height of 1.2m on the wall facing the participant who was asked to look at it during the playback of the stimuli.

Figure 7.2: Configuration of the *Classroom* during the experiment.

7.3.1.3 *Auditory Stimuli*

The stimuli were generated by convoluting a Swiss-German anechoic speech recording with each of the measured or modeled BRIRs. The playback level was set by

calibrating the level of the speech stimulus convoluted with the reference BRIR measured at 1m. For this BRIR, the stimulus was reproduced with headphones placed on the *Neumann KU100* dummy head, and the level was adjusted to match that of a standard male speaking standing at 1m in front of the dummy head (68dB - LAeq) . Stimuli were rendered through circumaural open headphones (*Sennheiser HD 650*), no head-tracking system was used. The participant's head was immobilised using a chin rest during the trials, to prevent inadvertent movements.

7.3.1.4 *Report method*

The participant reported the perceived sound source distance with a graphical slider presented on a touchscreen tablet (see Section 6.1.3). The software *MAX/MSP* was used for the rendering of the stimuli, the creation of the graphical interface, and the data collection.

7.3.2 *Procedure*

The participant was given the tablet and was introduced to the graphical interface used for reporting distance judgements. It was explained that the minimum of the slider corresponded to the participant's position and the maximum to the back wall. After an indication of the expected duration of the experiment (1 hour), the participant started a training session of 27 stimuli, composed of all the different possible conditions (9 distances $\times$ (2 models + reference)). The goal of the training session was to familiarize the participant with the distance reporting method and to ensure that the procedure was understood. After the training, the experiment was divided into three blocks, each of them containing 81 stimuli. Each stimulus was repeated 3 times within each block. The order of the stimuli within a block was randomized. During the trials, the participant could trigger the stimulus playback when she/he wanted, but it was played only once. The trial response was collected through the graphical interface given to the participant. A final questionnaire was filled out at the end of the 3 blocks to collect additional information related to the localization of the source (externalization, direction), realism, problems with the interface, global attention of the participant, and noticeable differences between the different stimuli apart from the distance.

7.4 EXPERIMENT I: RESULTS

Statistical analysis were performed using *TIBCO Statistica*© except for the power-function fittings, which were performed using *Mathworks Inc MATLAB*©.

Initial attention was focused on the normalization of the responses of each participant on every condition (9 distances $\times$ (2 models + reference)). A Jarque Bera-test indicated that one participant showed a non normal distribution of the responses for a majority of the tested conditions. This participant was excluded

from the pool of participants. The following analysis is based on the data of 19 participants.

7.4.1 *General results*

To analyze the performance of each model and reference, the geometric mean of the perceived distance of each participant in each condition was computed. For comparison purposes, the perceived distances considered here, result from a linear conversion of reports made by participants on the VAS (0% on the slider corresponding to 0m, and 100% to 7m).

$$D_g(d) = \prod_{k=1}^{n} \prod_{i=1}^{9} \sqrt[9n]{D_{k,i}(d)} \tag{11}$$

D_g the geometric mean perceived distance over all participants for a sound source at a distance d, n the number of participants, i the repetition of the stimulus (9 presentations of the same stimulus in total). This mean was used because it is admitted that distances are perceived following a power function [183].

Consequently, it is relevant to use a geometric mean over the perceived distance instead of a direct arithmetic mean. Thus, analysis is based on the logarithmic perceived distance. The logarithm of the geometric mean is equal to the arithmetic mean of the logarithm:

$$\log(D_g(d)) = \sum_{k=1}^{n} \sum_{i=1}^{9} \frac{\log(D_{k,i}(d))}{9n} \tag{12}$$

A repeated measures ANOVA applied to the geometric mean distances of each participant was carried out, with the within-subject factors DISTANCE (9 levels from 1 to 7m) and MODEL (3 levels: 2 models and the reference).

- The main effect DISTANCE was significant ($(F(1,8) = 283,08\ p < 0.01$, $Partial - \eta^2 = 0,9402$)
- as well as the MODEL ($F(1,2) = 12,87$, $p < 0.01$, $Partial - \eta^2 = 0.4168$)
- and the interaction DISTANCE $\times$ MODEL
  ($F(1,15) = 5,34$, $p < 0.01$, $Partial - \eta^2 = 0.2287$).

The analysis of the reference conditions confirms that the perceptual distance is globally overestimated for short distances, here from 1 to 5m and underestimated for longer distances. This behaviour is also observed for the envelope-based model. For the intensity model, the perceived distance is always underestimated, although it is close to the actual distance between 2 and 3m.

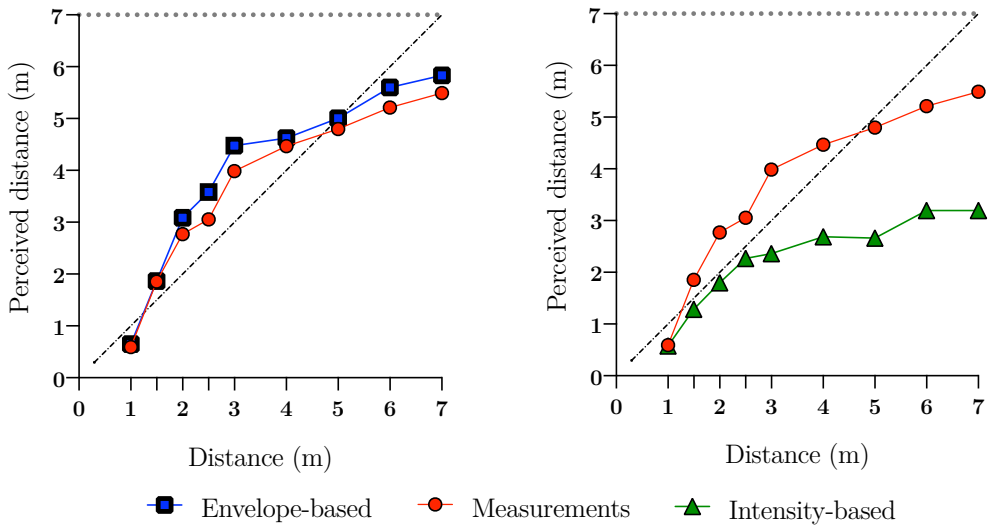


Figure 7.3: Experiment I: Geometric mean perceived distances according to the model used to generate the sound source: Reference based on measurements (red), Envelope-based model (blue) Intensity-based model (green).

The crossover point, the distance from which there is no bias in the perceived distance, is between 4 and 5 meters for the envelope-based model and reference condition, while it is between 3 and 4 meters for the intensity-based model.

The similarity between the enveloped-based model and the reference was further investigated using a post-hoc analysis (Fisher LSD). For each distance, no significant differences between the reference and the envelope-based model were found. Besides, this post-hoc analysis reveals the presence of a distance beyond which no significant effects of the distance are observed anymore. This auditory horizon appears at 5m for the envelope-based model and the reference and at 4m for the intensity-based model.

7.4.2 *Individual results*

The general results following the ANOVA revealed a perceptual similarity between the reference and the envelope-based model. The similarity between the envelope-based model and the reference can also be found at an individual level. The logarithm of the power function defined in Section 3.1 was fitted to the logarithm of the geometric mean perceived distance for each participant on each model (19 participants, 3 models) using a linear regression model (with k corresponding to the intercept and a corresponding to the slope). The values of the fit parameters, mentioned in the following as "compression coefficients" allow us to quantify the quantity of compression in the reported distances.

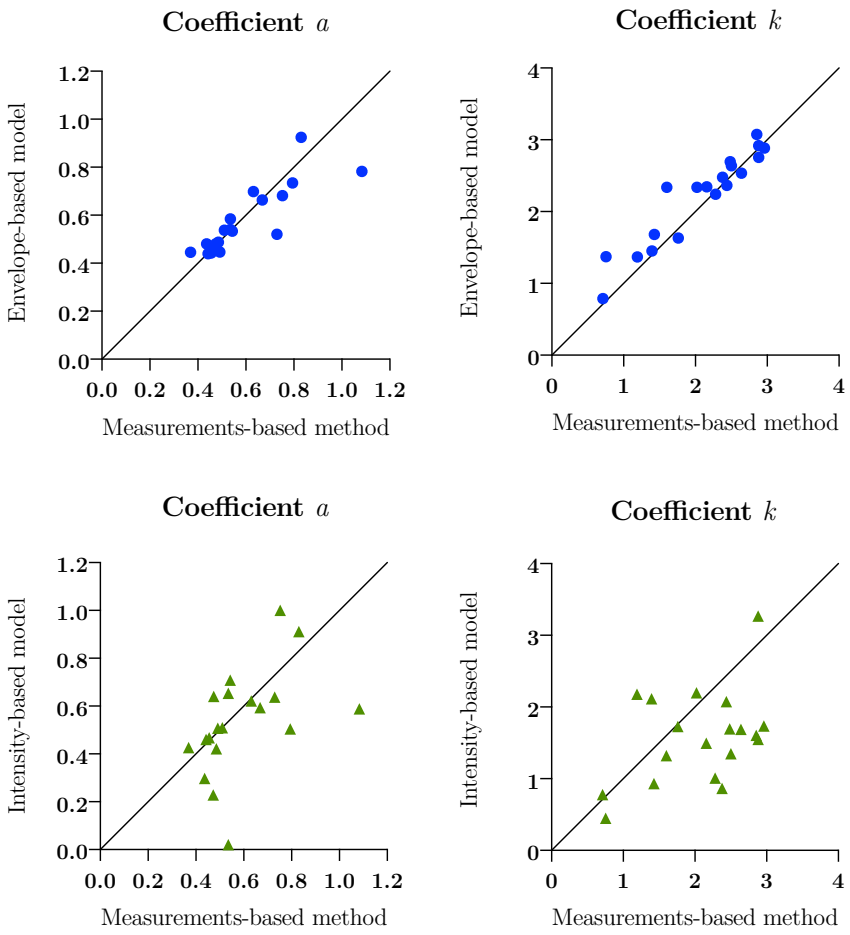


Figure 7.4: Comparison of the individual compression coefficients a (left) and k (right), between the reference and the models: Envelope-based (above,blue), Intensity-based (below,green).

The distribution of dots in Figure 7.4 confirms the presence of a similarity between the envelope-based model and the reference for each participant, in terms of auditory distance perception (the distribution of dots is close to the main diagonal). In contrast, the distribution of both coefficients for the intensity-based model is more dispersed, showing no correlation at an individual level. The result for the intensity-based model also shows that the majority of the non-linear compression coefficients a are lower than those obtained for the reference model. The graphical user interface had a limited range, which could introduce a bias in the collected perceived distances. Thus, for trials corresponding to distances from 5 to 7m, the normality of the responses was affected. Hence, no further analysis of the individual variability (intra-subject) could be conducted.

7.5 EXPERIMENT I: DISCUSSION

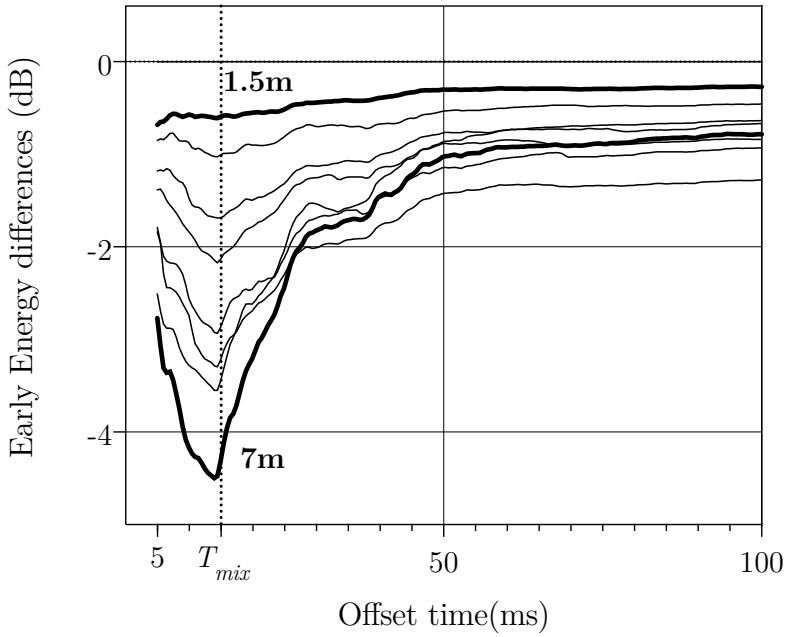
Experiment I evaluated the importance of different cues known to drive auditory distance perception. The procedure applied here aimed to be close to a simple AAR scenario, with participants having access to the vision of the room during the experiment, and the environmental context being congruent with the reproduced room effects. The roles of the global intensity and of the intensity contained in various temporal segments of the room impulse response were investigated through the comparison of three different rendering methods.

7.5.1 *Envelope-based model performances*

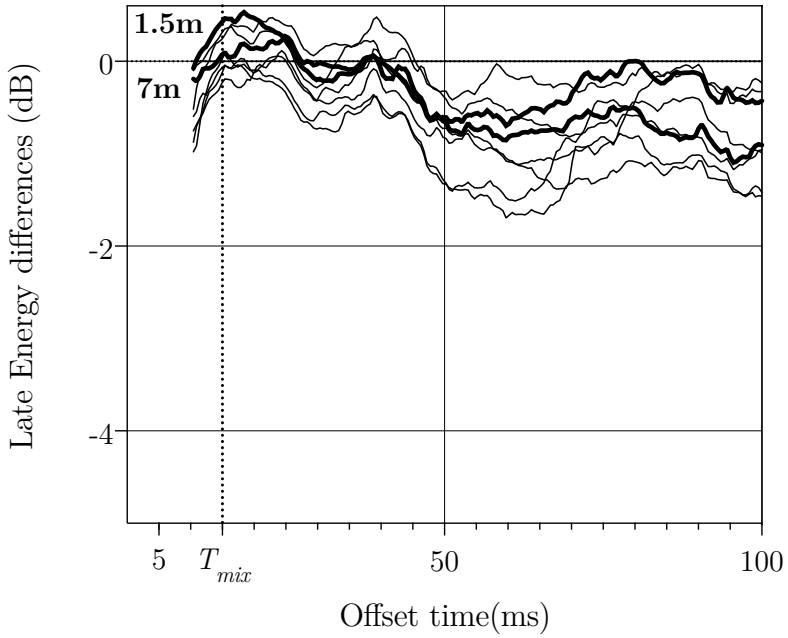
The envelope-based model and the rendering method based on measurements were shown to provide very similar auditory distance perceptions. This result was further confirmed by the post-hoc analysis applied to each tested distance. However, they differ significantly by design in terms of their early-to-late energy ratio. We intended to investigate if setting an adequate temporal limit (designated "offset time") on the evaluation of early energy may account for this perceptual similarity. To do so, the early and late energies of the BRIR associated with the envelope-based model and the reference were computed for different offset time values. For each distance, the early and late energies differences computed between a BRIR generated by the envelope based-model and converted from an actual measurement are displayed in Figure 7.5.

The envelope-based model underestimates the early energy for offset times under 50ms, and with increasing distances. In spite of this objective discrepancy, the observed perceptual similarity raises the question of the definition of a reverberation-related distance cue, and which definition could best explain the perceptual similarity between the model and the measurements. Previously, Kopčo and Shinn-Cunningham [93] have suggested that an early-to-late power ratio is the best candidate to overtake DRR as a relevant reverberation-related distance cue. However, past studies that explored the importance of this ratio as an auditory distance cue present contrasting results on the temporal limit that should be considered between the early energy and late energy [26, 27, 120, 125, 165].

The differences in terms of late energy between the measurements and the extrapolated responses by the envelope-based model are not significant for any offset time considered superior to 10ms (see Figure 7.5). Early energy differences tend to reach their minimum for a considered offset time of 50ms. When considering an early-to-late ratio for a limit of 50ms, consistent with the definition of speech clarity C50, the difference between measurements and generated impulse responses is around 1dB. For speech signals, the just-noticeable-differences of C50 are around 1.1 dB and a significant change value is considered to be around 3dB [24]. The early energy difference observed between the measured impulse responses and those generated by the envelope-based model is inferior to 3dB.



(a) Early energy differences as a function of the offset time considered as the end of the early energy since direct sound



(b) Late energy differences as a function of offset time

Figure 7.5: Early and late energy differences between measured BRIRs of the room and extrapolated BRIRs with the envelope-based model. Fifteen milliseconds corresponds to the limit considered in the envelope-based model between the early part of the impulse response and the late reverberation.

Thus, for a considered offset time of 50ms, the similarity in terms of early energy could explain the perceptual similarity observed in the current experiment, between the envelope-based model and the rendering method based on measurements. As the experiment results are only related to a single acoustical environment, it is not sufficient to draw a general conclusion about what an appropriate limit should be considered as an early-to-late energy ratio in order to compute a distance cue. It does, however, corroborate the assumption that early reflections are perceived as part of the direct sound [26].

7.5.2 *Intensity-based model performances*

The importance of reverberation-related distance cues for auditory distance perception is also put into evidence when comparing the reported distances obtained on the intensity-based model to those obtained for the two other rendering methods. Reported distances for the intensity-based model exhibit a stronger compression effect and a larger inter-subject variability than for the envelope-based model. Mean values and standard deviation of the compression coefficients a and k (see Table 1) confirm this observation. Both coefficients are closer to 1 for the envelope-based model and the rendering method based on measurements.

This result relative to the intensity-based model, is coherent with the objective differences illustrated in Figure 7.1. Stimuli generated by this model faithfully reproduce the loudness when compared to stimuli generated with actual measurements, but the energy contained in the late reverberation is strongly underestimated, and the early-to-late energy ratio is constant despite the varying distance of the source. The absence of variation when the stimulus distance increases is probably the main reason why the sound source distance produced by the intensity-based model is generally under-estimated when compared to the distances reported with the measurements. These results are consistent with auditory distance perception studies in anechoic environments, in which the early-to-late energy ratio is constant when distance changes [31].

7.5.3 *Acoustic cues weighting strategies*

Individual participant's compression coefficients for the envelope-based model and the rendering method based on measurements are homogeneous. Each participant presents a similar compression effect for these two rendering methods. In contrast, the difference between the compression coefficients associated with the rendering method based on measurements and the intensity-based model suggests that different strategies were used to judge the distance of sound sources. Participants for which the value of the compression coefficients estimated on the intensity model is comparable to the value obtained with the rendering method based on measurements mainly base their auditory distance judgements on intensity. Some participants obtain a smaller compression coefficient a for the intensity-

based model only, then do not only use intensity as a distance cue, and rely primarily on a reverberation-related distance cue. One participant received a coefficient a equal to 0, indicating that he did not consider intensity to infer distance judgments at all.

These results show that the acoustic cues weighting strategies are mainly an individual characteristic. This result is based only on 19 participants, therefore, it was further investigated in Experiment III (see Chapter 8). We notably examined if other participants only based their auditory distance judgements on intensity. The influence of environmental context on the weighting of acoustic cues is also examined to determine if it can be linked to characteristics of the environment or if it is only an idiosyncratic characteristic.

7.5.4 *Influence of the experimental context and comparison with past studies*

	Calcagno [33]	**Anderson [7] Audio**	**Anderson AV**	**Zahorik [183]**	**Ref.**	**Envelope**	**Intensity**
k	1.14 +- 0.12	2.22 +- 1.99	1.38 +- 0.91	1.32	1.24 +-0.50	1.16 +-0.49	1.75 +-1.22
a	0.89 +- 0.06	0.61 +- 0.30	0.87 +- 0.27	0.54	0.85 +-0.24	0.87 +-0.33	0.83 +-0.55
R^2	n/a	0.64 +- 0.22	0.84 +- 0.18	0.91	0.79 +-0.1	0.78 +-0.08	0.71 +-0.24
Sound rendering	Speaker	Binaural		n/a	Binaural		
Stimuli content	Noise Bursts	Gaussian Noise		n/a	Speech		
Room type	Semi-reverberant	Concert hall		n/a	Damped room		
Distance Range	2-6m	0.3-9.8m		n/a	1-7m		
Report method	Verbal	Verbal		Verbal	Visual analogue scale		

Table 1: Mean values of compression coefficients a and k with standard deviation and R^2 reported in different studies.

To evaluate the influence of the specific perceptual context (presence of room-related visual cues, speech stimuli) of the current experiment on accuracy and variability, the results are compared with previous auditory distance perception studies. Table 1 compares the mean value of the Zahorik's power function parameters obtained in several studies: a meta-analysis realized by Zahorik [183] over 81 different studies dedicated to the perception of auditory distance, a study from Anderson [7] comparing auditory distance judgements in audio-only and audio-visual condition and a study from Calcagno [33] studying auditory distance judgements in presence of visual cues. For the two models and the reference of

the current study, the non-linear compression coefficient a is shown to be closer to 1, compared to the values obtained in the meta-analysis realized by Zahorik. The coefficient k is also closer to 1 for the reference and envelope-based model. These values can be interpreted as a better global accuracy in distance judgements. The superior accuracy of auditory distance perception associated with the reference and the envelope-based model can be caused by the presence of visual cues and the use of speech, while the meta-analysis ran by Zahorik [183] is mainly based on studies implying blind auditory distance judgements and various types of stimuli (from noise bursts to speech signals). In order to confirm that the presence of visual cues and the use of speech signals enhanced the accuracy and possibly reduced the variability of auditory distance reports [104, 179], comparisons with auditory distance perception studies involving visual cues are made.

With both coefficients closer to 1, the results are similar to what can be found in the studies run by Anderson & Zahorik [7]. The results in terms of accuracy are also consistent with a study conducted by Calcagno [33]. One of their protocols is actually similar to the current experiment (see Table 1), as the participants could have access to the visual configuration of the room prior to the experiment. Although their experiment was using a real loudspeaker, in contrast with our experiment, which uses binaural reproduction on headphones, the mean fit parameters are comparable. However, the comparison in terms of variability must be done cautiously considering the difference in fitting methods (Calcagno uses fitting on raw data with a least square method instead of using linear fitting on logarithmic data) and number of trials per condition (3 instead of 9).

Compression coefficients are also consistent with those obtained in the audio-visual condition of the study by Anderson & Zahorik [7], although the nature of the visual cues' influence is different. In their study, each auditory stimulus condition is associated with a simultaneous projection of a loudspeaker image at the same location, whereas in our experiment, the distribution of the visual anchors (chairs) does not coincide with the auditory stimuli. The comparison of the standard deviation calculated on the compression coefficients a and k, indicates lower inter-subject variability in this experiment. However, Anderson & Zahorik used a verbal reporting method, and the inherent noise it induces could lead to an overestimation of the real perceptual noise. Contrarily to verbal reports, the use of a VAS induces the presence of a maximal reported value. Consequently, a ceiling effect was observed on distance reports for stimuli generated for distances beyond 5m. This lead to an underestimation of the intra and inter-subject variability.

7.6 EXPERIMENT II: EVALUATING THE RELEVANCE OF THE EARLY-TO-LATE ENERGY RATIO

The perceptual evaluation of the envelope-based model in Experiment I demonstrated its capacity, in the *Classroom*, to reproduce sound source distance when compared to stimuli convoluted with actual measurements. The model was directly inspired by the work of Barron [16] and Jot [82], which were initially introduced to model the objective impact of the source to receiver distance on late reverberation.

Compared to the actual measurements, the envelope-based model accurately reproduces the evolution of the late reverberation energy and of the total energy according to the distance (see Figure 7.1). However, the energy of the early segment, including the direct sound and the first reflections before the mixing time $t < T_{mix}$, becomes greatly underestimated as distance increases. Despite this energetic difference, the subjective distance reports were observed very similar between the envelope-based and the measurements rendering methods (see Figure 7.3).

In Figure 7.5 it was seen that, if a transition time $T_{trans} = 50$ms is considered between the early reflections and the late reverberation segments of the BRIRs extrapolated by the envelope-based model, then their respective energies are comparable to the ones of the measurements. This could be an indication that such a transition time should be considered rather than a strict DRR when defining an early-to-late energy ratio as a reverberation-based distance cue. However, although the model reproduces correctly the overall energy contained in the early segment up to T_{trans}, it induces a significantly different temporal distribution of the energy within it. This is illustrated by the schematic of the envelope-based model design displayed in Figure 7.6.

The purpose of this second experiment is to ascertain the perceptual relevance of the early-to-late energy ratio for auditory distance perception. In order to analyze the relevance of this criterion, BRIRs correctly reproduced this ratio with distance, but inducing various distributions of the energy within the early part, were synthesized. The stimuli convoluted by these BRIRs are compared to stimuli convoluted by actual measurements in an online-based experiment involving 120 participants.

Two different considerations of the transition time T_{trans} defining the temporal limit of the early energy were tested: 40ms and 80ms. For $T_{trans} = 40$ms, despite varying energy distributions in the early part of the impulse responses, we expected that no significant perceptual differences with measurements should be revealed.

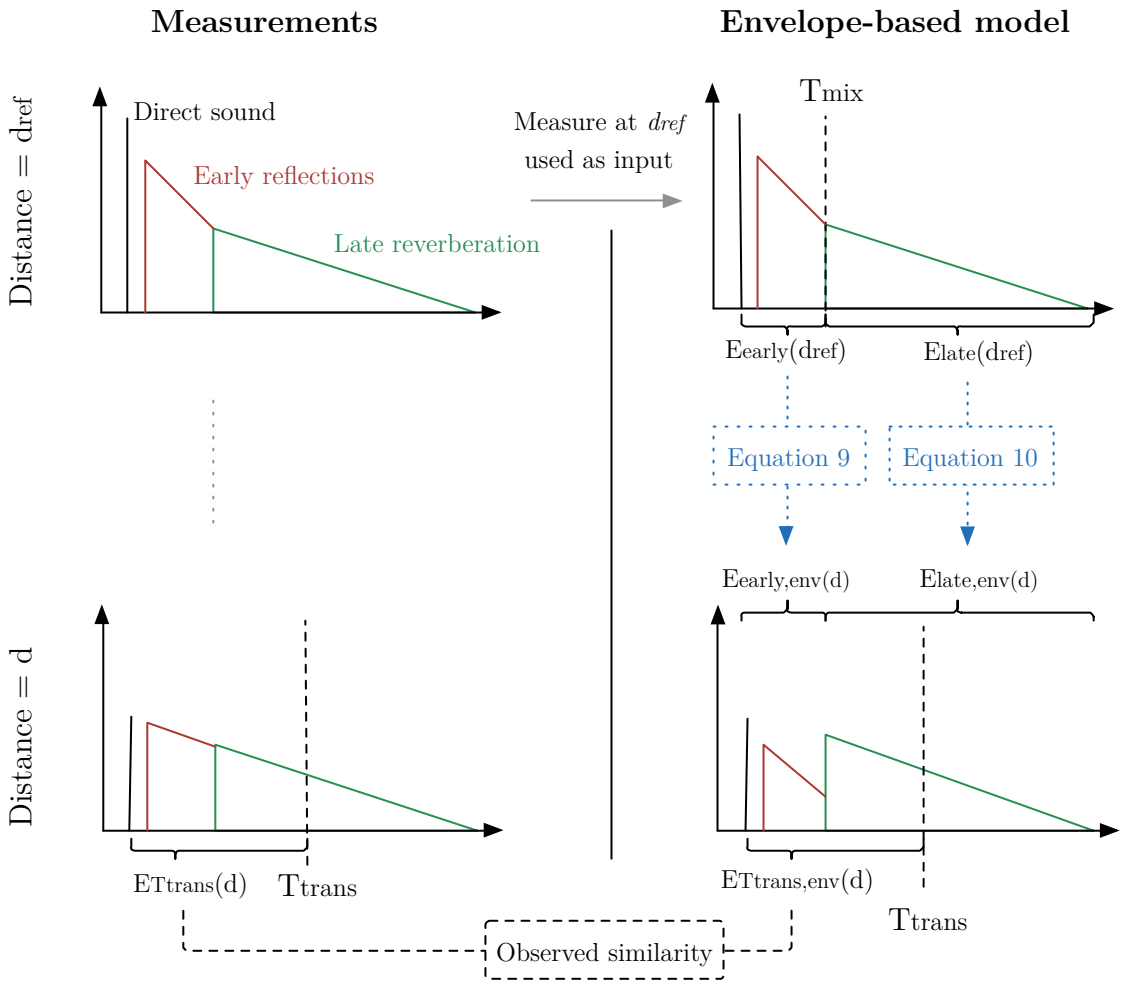


Figure 7.6: Schematic breakdown of the envelope-based model design (see Equation 9 and Equation 10). The energy in the first 50ms of the extrapolated BRIRs is comparable to one of the measurements. However, large energetic differences are induced in the early reflections distribution for $t < T_{mix}$.

7.6.1 *BRIRs synthesis method*

The following method, illustrated in Figure 7.7, is used to synthesize BRIRs with different temporal distributions of the early energy contained between their onsets and T_{trans}. For a given transition time T_{trans} and a set of SRIRs measured in a given environment:

1. The measured SRIRs are converted into BRIRs as described in Section 5.2.2.
2. The early segment (i.e. direct sound + first reflections) of an initial BRIR is extracted, from its onset time $t = 0$ to the transition time $t = T_{trans}$.
3. This initial early segment is then substituted to the early segment of the different BRIRs measured at other distances after applying a correction gain in order to match their original early energy.

This method allows to create synthetic BRIRs which differ from the original measured ones by the distribution of the direct sound and first reflections prior to T_{trans}. The synthetic BRIRs presents a late reverberation tail (for $t > T_{trans}$) fully identical to the original ones, as well as their total energy and the global early energy contained in the interval $[0, T_{trans}]$. However, although the global gain of the synthetic early segments matches that of the original BRIRs, their time, frequency and spatial distributions are modified.

Depending on the chosen initial measurement and the reproduced distances, different temporal distributions of the energy in the early segment $[0\text{ms}, T_{trans}]$, will occur. Here, the closest and furthest measurements available in the environment were used as initial measurements. This was done so that differences in terms of early energy distribution within the synthetic BRIRs were maximized.

When the initial measurement is chosen as the closest available, the synthesis consists of extrapolating further distances. This BRIRs synthesis is labeled as "*Backward*". The stimuli generated through convolution with such BRIRs are labeled as "*Backward* stimuli".

Contrarily, when the initial measurement is chosen as the furthest available, the synthesis consists of extrapolating closer distances. This BRIRs synthesis is labeled as "*Forward*". The stimuli generated through convolution with such BRIRs are labeled as "*Forward* stimuli".

For the "*Backward*" synthesis, extrapolating distances further than the initial measurement position implies an excess of the energy conveyed by the direct sound compared to the reference BRIRs. This is explained by the fact that sound level decreases following a 6dB drop per doubling distance, while early and late reflections decrease more slowly. Hence, applying a global gain to the whole part of the initial measurement creates an energy excess of the direct sound and possibly of the earliest reflections. This phenomenon is illustrated by the schematic of the method design displayed in Figure 7.7.

For the "*Forward*" synthesis, extrapolating distances closer than the initial measurement implies a lack of the energy conveyed by the direct sound.

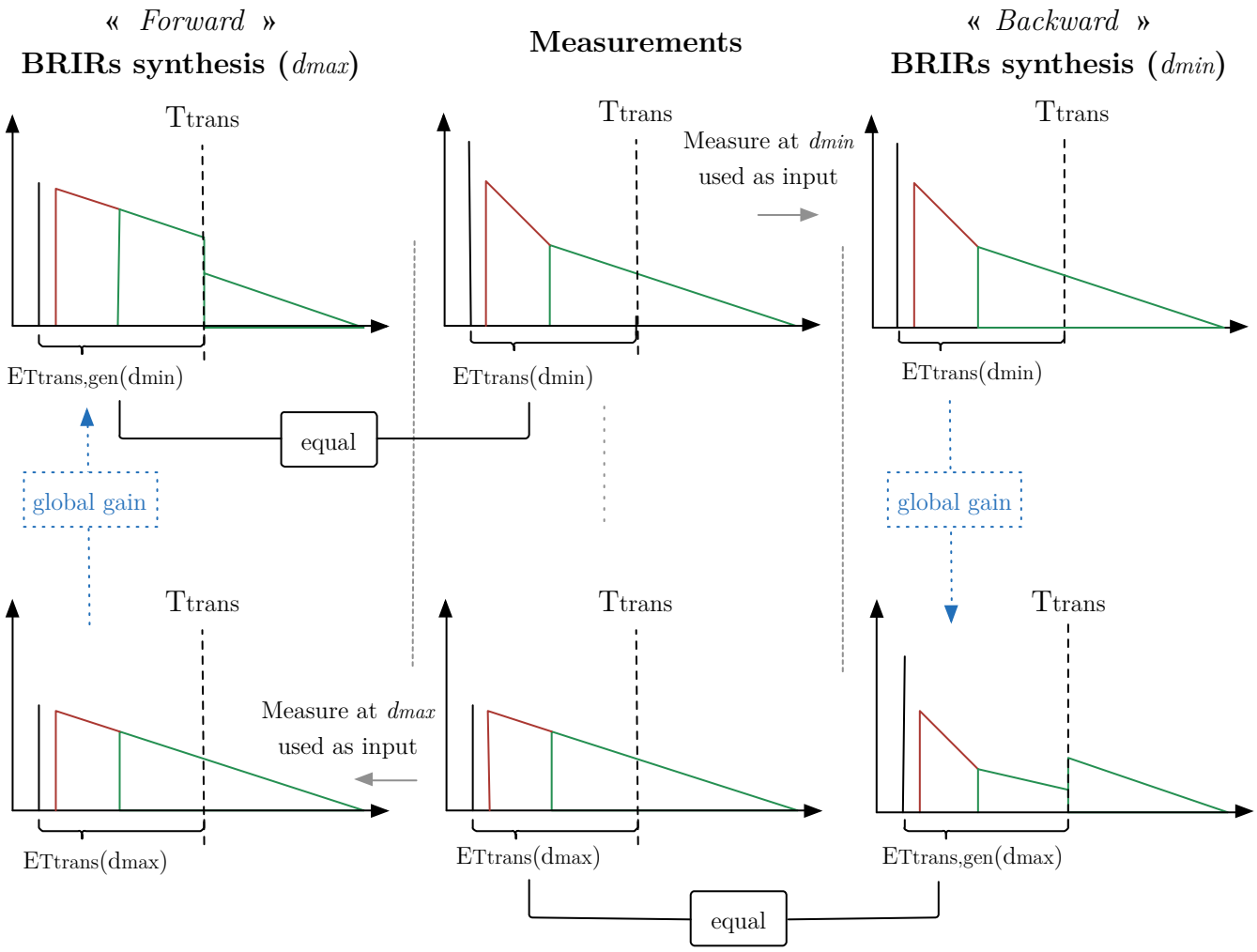


Figure 7.7: Schematic breakdown of the method used to synthesize BRIRs in Experiment II. The early segment of an initial BRIR is extracted, from its onset time $t = 0$ to the transition time $t = T_{trans}$. This initial early segment is then substituted to the early segment of the different BRIRs measured at other distances after applying a correction gain in order to match their original early energy. The late segments (for $t > T_{trans}$) of the original BRIRs remain unchanged.

In Experiment II, these methods were applied to two different rooms with different geometrical and acoustical characteristics.

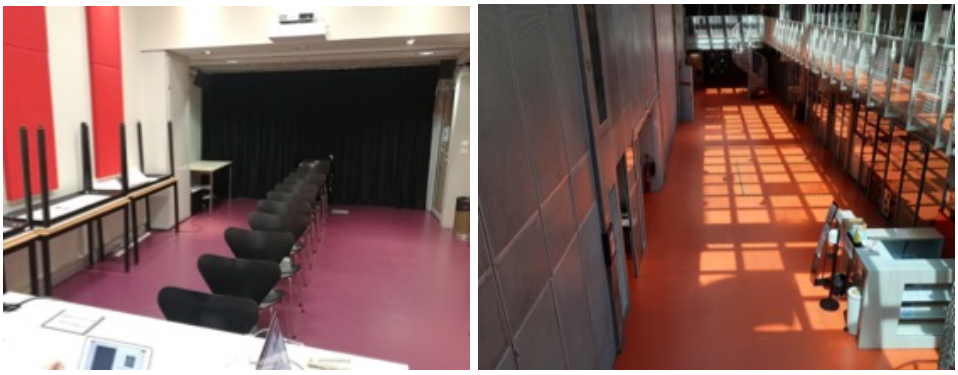
7.6.2 *Material & Methods*

The experiment was carried out using the previously stated online procedure (see Section 6.2). The experiment consisted of two separate groups, each evaluating one of the two room conditions. Indeed, gathering both room conditions within the same session could introduce biases as a result of the sequencing of successively contrasted acoustic conditions. As different room conditions with various level and duration of reverberation and for different ranges of distances are evaluated, merging all these conditions in the same evaluation test could lead to important compression or expansion biases in the distance reports. For each room condition, the performances of participants on the *Backward* and *Forward* stimuli were evaluated in contrast to reference stimuli based on measurements made in each room.

Additionally, for each room, two different considerations of the transition time were evaluated : 40ms and 80ms. These diverse situations (in terms of transition times T_{trans}, acoustic environments, and reproduced distances) permitted us to further assess the perceptual relevance of the early-to-late energy ratio by comparing participants' distance perceptions of stimuli convoluted with synthetic and measured responses.

7.6.2.1 *Tested conditions*

The two environments used for the measurements are illustrated in Figure 7.8.



(a) Classroom at IRCAM mentioned as *Classroom*
Volume = 144m^3
T_{60}(1kHz)= 0.55s

(b) Patio at IRCAM mentioned as *Gallery*
Volume > 1700m^3
T_{60}(1kHz)= 0.9s

Figure 7.8: The two different acoustic environments used in Experiment II.

CLASSROOM Six different distances were evaluated: 1, 1.5, 2, 3, 4, 6 meters. Measurements at 1 and 7 meters were respectively used as the initial measurements of the *Backward* and *Forward* direction syntheses. A total of 30 stimuli: 6 distances $\times$ 5 rendering methods (2 direction syntheses $\times$ 2 transition times + 1 reference) were evaluated in this room condition.

GALLERY Six different distances were tested for all rendering methods: 3, 4, 5, 6, 7, 9 meters. Measurements at 1 and 14 meters were respectively used as the initial measurements of the *Backward* and *Forward* direction syntheses. As for the Classroom, a total of 30 stimuli were evaluated in this room condition: 6 distances $\times$ 5 rendering methods (2 direction syntheses $\times$ 2 transition times + 1 reference).

Figure 7.9 and Figure 7.10 illustrate the differences, in terms of energy distribution within the part prior to T_{trans}, between the synthesized BRIRs and the measurements. In order to highlight these differences, a direct-to-reflections energy ratio $D/Ref_{T_{trans}}$ was computed for each BRIR used to generate a stimulus:

$$D/Ref_{T_{trans}}(d) = E_{dir}(d) - E_{ref,T_{trans}}(d) \qquad \textit{(in dB)} \tag{13}$$

E_{dir} is the energy of the direct sound and of the first reflections contained in the temporal segment $[0, 5ms[$ of the BRIR associated with the distance d. $E_{Ref,T_{trans}}$ is the energy contained in the temporal segment $[5ms, T_{trans}]$ of the same BRIR.

For each room condition, the range of tested distances was chosen to maximize the differences in terms of the $D/Ref_{T_{trans}}$ criterion.

To summarize, the late energy of each synthetic BRIR is identical to the corresponding BRIR measurement at a given distance. The *Backward* method synthesizes BRIRs with an excess of direct sound. The *Forward* method synthesizes BRIRs with a lack of direct sound.

Another representation of these differences can be found in Appendix b with a criterion similar to the one displayed in Figure 7.5.

7.6.2.2 *Participants*

A total of 120 online participants were recruited via the recruitment platform *Prolific* (see Section 6.2). Two groups of 60 participants performed each a separate version of the experiment, each version containing the evaluation of a single room condition. Inclusion and exclusion criteria reported in Chapter 6 were applied.

7.6.2.3 *Auditory Stimuli*

A 3-second recording truncated from the same 5-second speech anechoic recording used in the first experiment, has been pre-convoluted with each of the measured and synthesized BRIRs.

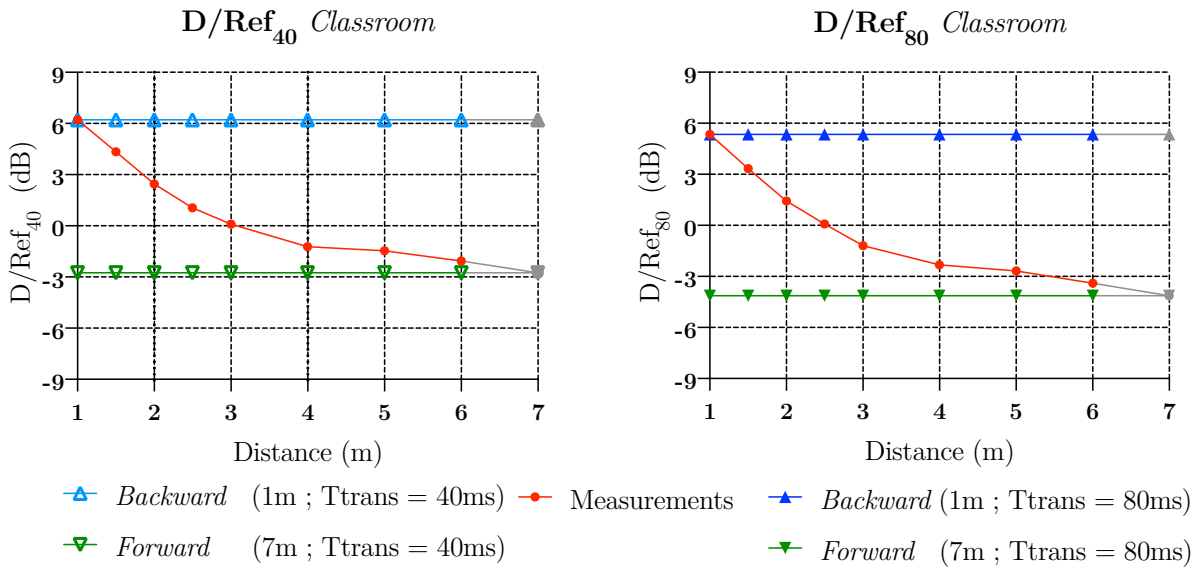


Figure 7.9: Direct-to-reflections energy ratio ($D/Ref_{T_{trans}}$) of the synthetic and measured BRIRs of the *Classroom*, for T_{trans} equal to 40ms (left) and 80ms (right). An initial impulse response at 1m (blue) and 7m (green) were exploited for the BRIRs synthesis. Grey symbols represent BRIRs available but not used in the experiment

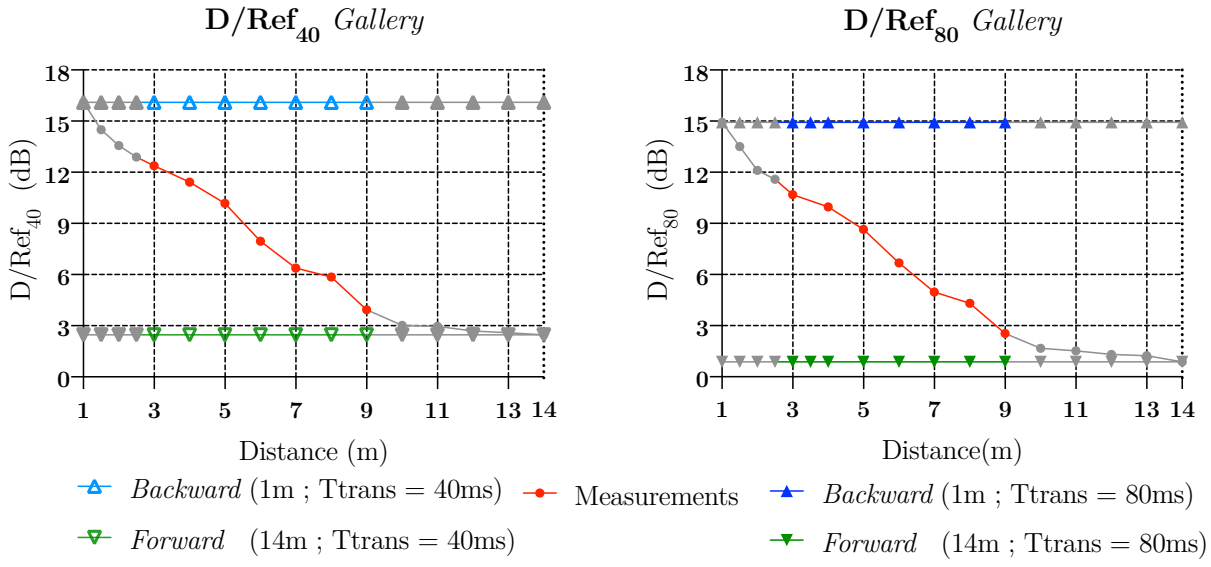


Figure 7.10: Direct-to-reflections energy ratio ($D/Ref_{T_{trans}}$) of the synthetic and measured BRIRs of the *Gallery*, for T_{trans} equal to 40ms (left) and 80ms (right). An initial impulse response at 1m (blue) and 14m (green) were exploited for the BRIRs synthesis. Grey symbols represent BRIRs available but not included in the experiment.

An additional equalization of the loudness of the generated stimuli was run following the EBU 128R criterion. The process was done so that all stimuli associated with a given distance were of the same loudness regardless of the rendering method used to generate it. The *Backward* and *Forward* syntheses induced a mismatch in the spectral content of the synthesized BRIRs when compared to actual measurements (which impact is discussed in the following Section 7.8.3). Then, the use of a speech signal resulted in differences in the loudness of the resulting stimuli, despite the fact that the total energy of the BRIRs used to generate them was identical.

7.6.2.4 *Report method*

A vertical VAS (continuous slider) was used to report the perceived distance of a sound stimulus (see Section 6.1.3). The top of the slider corresponded to a distant impression to the sound source while the bottom corresponded to an impression of proximity.

7.6.3 *Procedure*

After reading the information note and approving the consent form, participants were directed to the experiment. Participants had to fill out a questionnaire in which they were invited to give their age, gender, and the type of headphones used. They were requested to use circum-aural headphones if possible.

A screening process was launched to ensure the correct use of headphones: two audio clips corresponding to the weakest and loudest stimuli were played successively. Participants had to adjust the sound level for the quietest stimulus to be heard clearly. The loudest stimulus was then played, to ensure that it was heard at a comfortable level. Finally, broadband noise was displayed laterally first to the left channel of the headphones and then to the right one to ensure that participants were correctly wearing them.

Afterwards, a training session began. Participants were instructed to judge the apparent perceived distance to the male speaker using the VAS. Every stimuli used in the experiment were presented once during the training in a random order.

After the training session, the experiment began. Participants were informed that the procedure was the same as for the training session. Participants had to evaluate 120 stimuli (5 rendering methods $\times$ 6 distances $\times$ 4 presentations), randomized within a single block. During the trials, participants triggered the stimulus playback, but it could only be played once.

At the end of the experiment, participants were invited to fill out a final questionnaire to collect feedback on the experiment. More specifically, participants had to evaluate the general perceived externalization, the number of distinct distances perceived, and the maximum absolute distance perceived (in meters or feet). The mean duration of the procedure was 15 minutes.

7.7 EXPERIMENT II: RESULTS

The nature of the analysis used here is the same as the one employed in Experiment I. A focus is put on the comparison between reported distances associated with the 4 categories of synthetic BRIRs (two directions 'Forward' and 'Backward' and two transition times T_{trans} 40ms and 80ms) and reported distances associated with the actual measurements. The results of the two room conditions are presented independently.

OUTLIERS Six participants in the *Classroom* experiment and three participants from the *Gallery* experiment were excluded from the analysis. Among these nine participants, four of them were excluded because the mean and standard deviation of their responses in most conditions exceeded two standard deviations from the mean of the population. Two of them completed the experiment in an unrealistic short time ($<$ 5 minutes) which casts a doubt on the relevance of their reports. Three of them reported a front-back confusion effect or a lack of externalization.

DATA SCALING Initial attention was focused on the scaling of the responses in order to create comparable ratings between participants. Some participants exhibited a central tendency bias [65] in their distance reports, as they were only using the middle range of the slider. A min-max feature scaling was performed on the responses of participants using less than 95% of the total slider:

$$Y_{i,scaled} = \frac{Y_i - \min(Y_i)}{\max(Y_i) - \min(Y_i)} \tag{14}$$

with $Y_{i,scaled}$ the scaled response ranging from 0 to 1 and Y_i the raw response.

7.7.1 *Classroom*

With the intention of estimating compression coefficients (see Section 3.1), the perceived distances considered here, result from a linear conversion of reports made by participants on the VAS (0% on the slider corresponding to 0m, and 100% to 7m).

7.7.1.1 *General Results*

To determine the significance of differences between rendering methods, a repeated measures ANOVA ($\alpha = 0.05$) has been conducted on geometric means of reported distances of each participant per rendering method, with RENDERING (5 levels) and DISTANCE (6 levels ranging from 1 to 6m) as within-subject factors. The analysis revealed:

- a main effect of DISTANCE ($F(1,5) = 398.6$, $p < 0.001$)

- a main effect of RENDERING ($F(1,4) = 73.1$, $p < 0.001$)
- as well as a DISTANCE $\times$ RENDERING interaction ($F(1,20) = 7.23$; $p < 0.001$)

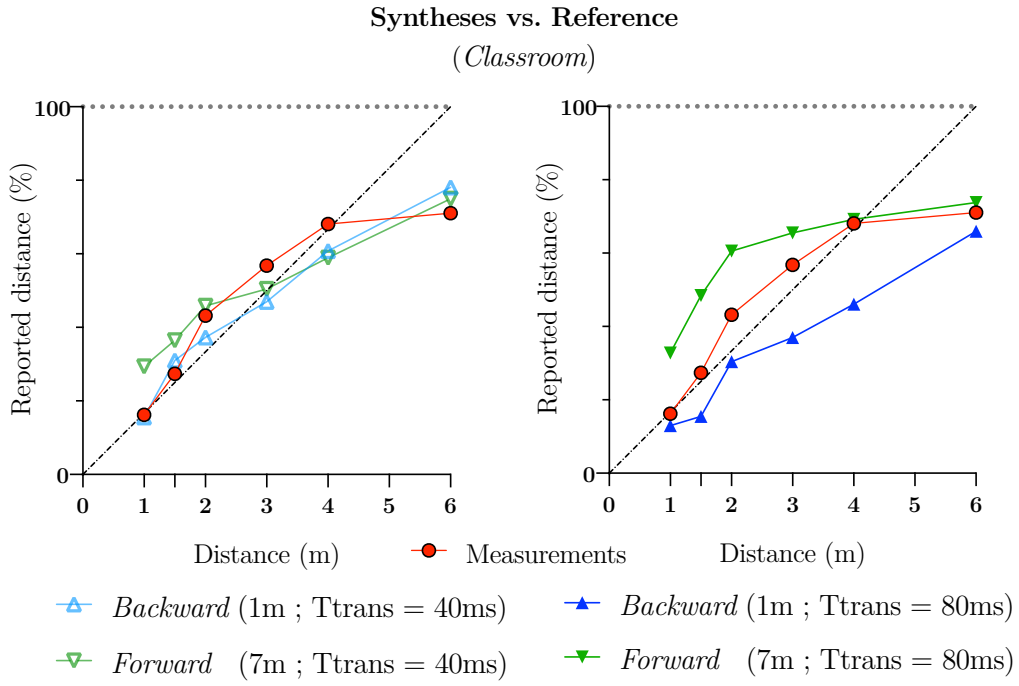


Figure 7.11: Experiment II (*Classroom*): Geometric mean perceived distances according to the method used to generate the stimuli. The reference (red) based on measurements of the *Classroom*, *Backward* (blue) and *Forward* (green) directions. With T_{trans} equal to 40ms (open symbols) or 80ms (solid symbols).

To assess the observations of Figure 7.11 a post-hoc analysis was run to verify if for each distance, the differences in terms of reported distances between one rendering method and the reference, are significant.

This analysis demonstrates that *Backward* stimuli rendered with $T_{trans} = 80$ms, induced a significant underestimation of the reported distances, for distances ranging from 2 to 4 meters. For $T_{trans} = 40$ms, no significant differences are revealed.

Concerning *Forward* stimuli rendered with $T_{trans} = 80$ms, an overestimation of the reported distances for 1 to 3 meters was exposed. For $T_{trans} = 40$ms, a significant overestimation of the reported distances for a distance of 1 meter only was revealed.

7.7.1.2 *Individual compression coefficients across rendering methods*

In order to get another insight into the perceptual differences induced by the *Backward* and *Forward* stimuli, the individual compression coefficients α and k were

computed by fitting the power function reported in Section 3.1. Individual fittings were made to the reports of each participant, for *Backward* and the *Forward* stimuli rendered with $T_{trans} = 80$ms, and for the reference. Only for this transition time were the coefficients a and k computed as these stimuli exhibit significant differences with the reference.

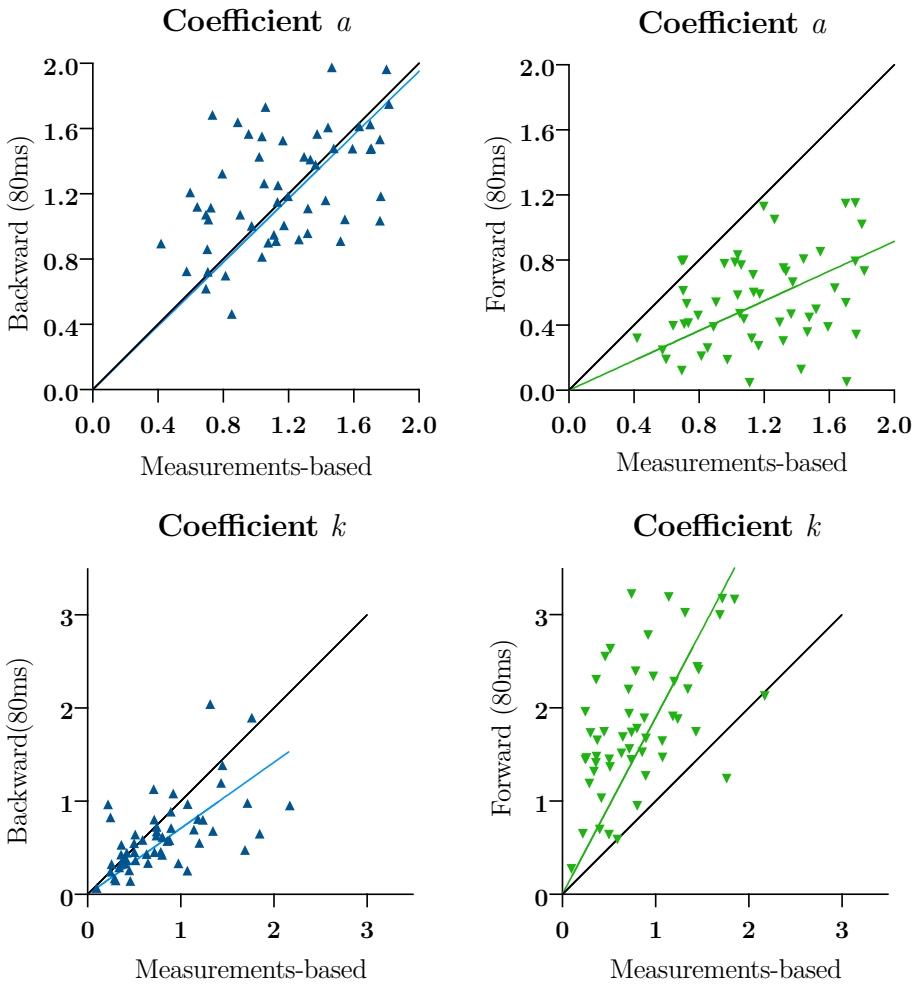


Figure 7.12: Experiment II (*Classroom*): Comparison of the individual compression coefficients a (above) and k (below) between the rendering method based on measurements and the *Backward*(blue,left) and *Forward*(green,right) syntheses.

The distribution of compression coefficients in Figure 7.12 confirms the presence of a higher compression in distance reports associated with *Forward* stimuli. All coefficients a and k, except one, are respectively inferior and superior, to those obtained by the same participants with the measurements.

7.7.2 *Gallery*

Results from participants who evaluated the *Gallery* condition are presented in this section. For comparison purposes, the perceived distances considered here, result from a linear conversion of reports made by participants on the VAS (0% on the slider corresponding to 0m, and 100% to 9m).

7.7.2.1 *General Results*

A repeated measures ANOVA ($\alpha = 0.05$) has been conducted on geometric mean of reported distances of each participant, with RENDERING (5 levels) and DISTANCE (6 levels ranging from 3 to 9m) as within-subject factors. The analysis revealed:

- a main effect of DISTANCE ($F(1,5) = 211.9$, $p < 0.001$)
- a main effect of RENDERING ($F(1,4) = 74.6$, $p < 0.001$)
- a DISTANCE $\times$ RENDERING interaction ($F(1,20) = 8.12$; $p < 0.001$)

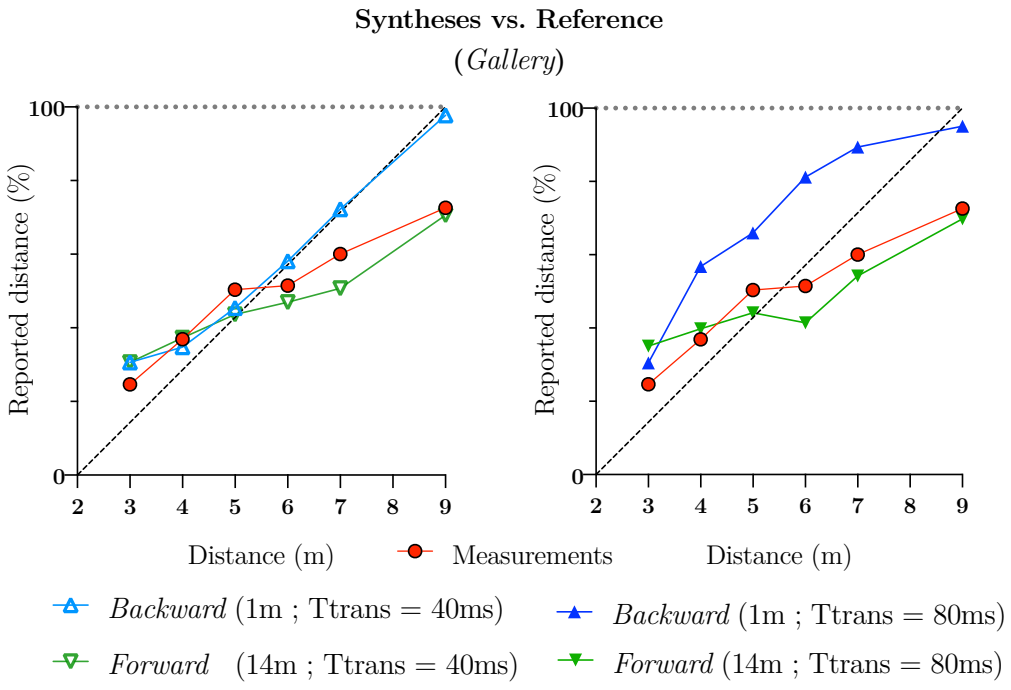


Figure 7.13: Experiment II (*Gallery*): Geometric mean perceived distances according to the model used to generate the sound source. The reference (red) based on measurements of the *Gallery*,*Backward* (blue) and *Forward* (green) directions. With T_{trans} equal to 40ms (left, open symbols) or 80ms (right, solid symbols).

To assess the observations of Figure 7.11 a post-hoc analysis was run.

For *Backward* stimuli, opposite differences with the reference such as those revealed in the *Classroom* were acknowledged. For $T_{trans} = 40$ms, distances of stimuli at 7 and 9 meters were significantly ($p > 0.05$) overestimated. For $T_{trans} = 80$ms, stimuli associated with distances superior or equal to 3 meters were significantly overestimated.

For *Forward* stimuli, no significant differences with the reference were revealed by the post-hoc analysis. Except for $T_{trans} = 80$ms, at a distance of 1m, the distance was overestimated.

7.7.2.2 *Compression coefficients across rendering methods*

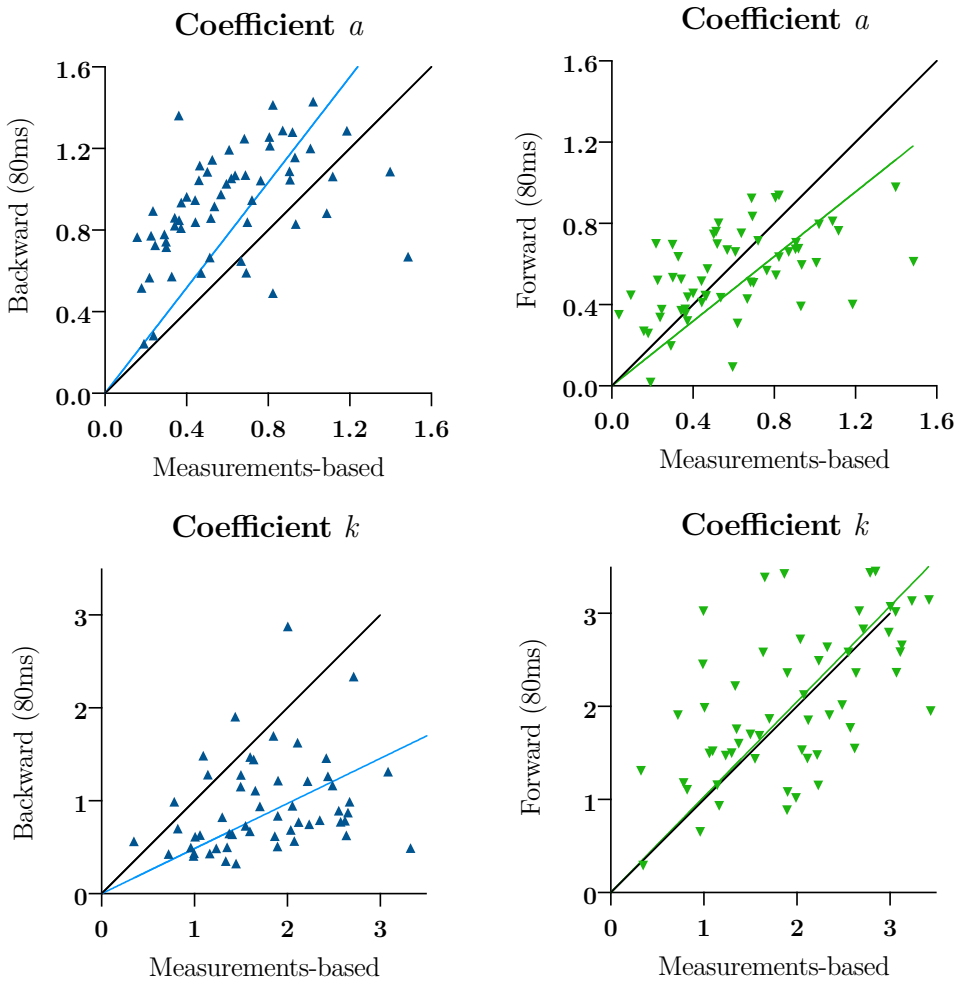


Figure 7.14: Experiment II (*Gallery*): Comparison of the individual compression coefficients a (above) and k (below) between the rendering method based on measurements and the *Backward*(blue,left) and *Forward*(green,right) syntheses.

Individual compression coefficients associated with *Backward* and *Forward* stimuli (rendered with $T_{trans} = 80$ms), and with reference stimuli were computed.

The distribution of compression coefficients in Figure 7.12 confirms the presence of a smaller compression in distance reports associated with the *Backward* stimuli. Most coefficients a and k are respectively superior and inferior, to those obtained by the same participants on reference stimuli.

Concerning the *Forward* stimuli, compression coefficients a are slightly smaller, indicating a stronger compression effect.

7.8 DISCUSSION

The goal of this second experiment was to assess the perceptual relevance of an early-to-late energy ratio for auditory distance perception. To do so, we used methods synthesizing BRIRs correctly reproducing the early energy (prior to 40 or 80ms) and the late reverberation with distance, but inducing different temporal distributions of the energy within the early part.

In respect with the results of Experiment I, we expected that for $T_{trans} = 40$ms, the *Backward* and *Forward* method could not be differentiated in terms of perceived distance from reference stimuli.

We expected that for $T_{trans} = 80$ms, an excess or a lack of the energy contained in the direct sound and the first reflections, would respectively induce an underestimation or an overestimation of the reported distances when compared to actual measurements. However, the results of this experiment do not show this symmetry. Thus, the evaluations of distance reports of the *Backward* and *Forward* stimuli are first discussed independently, in comparison to distance reports collected for the reference stimuli based on measurements.

Apart from creating a difference in terms of early energy distribution, spectral and spatial aspects of the synthesized BRIRs were inherited from the initial measurement. The possible effects of these spectral and spatial differences with the reference measurements are discussed afterwards.

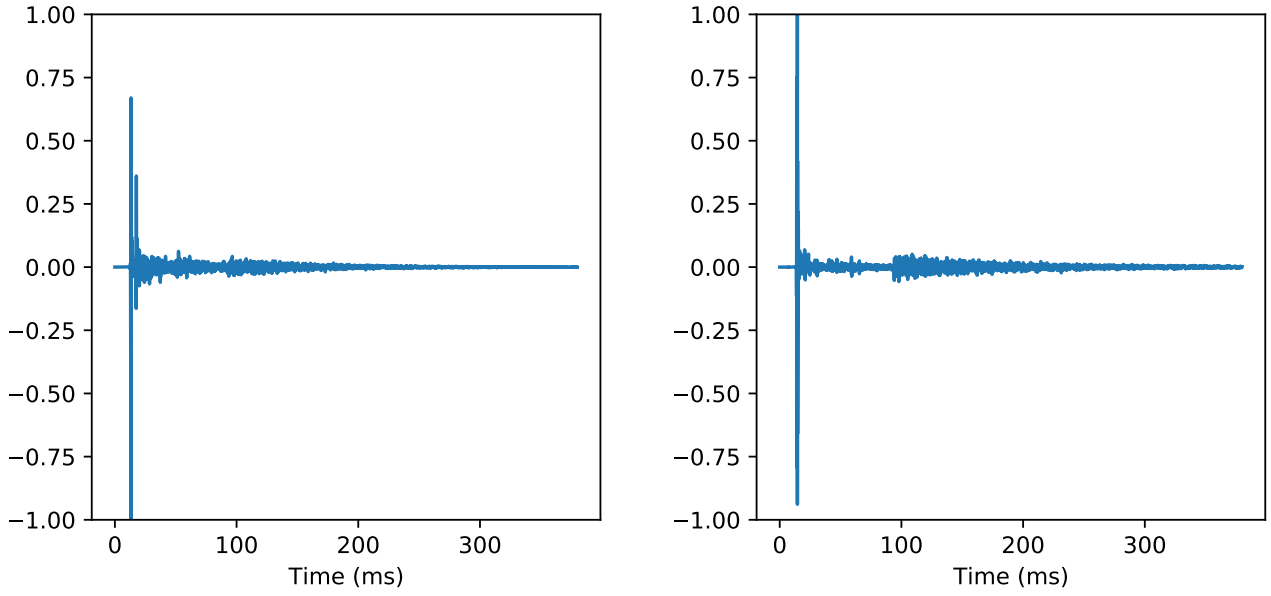
7.8.1 Backward *stimuli*

In the *Classroom*, for $T_{trans} = 40$ms, no significant differences are observed. For $T_{trans} = 80$ms, we effectively acknowledged an underestimation of the reported distances for stimuli from 2 to 4 meters when compared to the reference. This underestimation of the perceived distance can be attributed to the differences in the distribution of the early energy prior to 80ms. The excess of direct sound and early reflections induced by this BRIRs synthesis method incorrectly reproduced the early-to-late energy ratio.

In the *Gallery*, distances of stimuli are overestimated (over 7m for $T_{trans} = 40$ms; over 3m for $T_{trans} = 80$ms). Greater differences with increasing distance and transition time were effectively observed. However, due to the excess of direct sound,

we assumed that it should have induced an underestimation of the perceived distances.

A specific phenomenon linked to the *Backward* synthesis might be the cause. The time-energy envelopes of the early part of the synthesized impulse responses are inherited from the initial measurement. Thus, BRIRs synthesized from a measurement at 1m present a limited amount of energy in their early reflections when compared to actual measurements (see Figure 7.10).



(a) Waveform of the left channel of the synthesized BRIR at 6m with the *Backward* method in the *Classroom*

(b) Waveform of the left channel of the synthesized BRIR at 9m with the *Backward* method in the *Gallery*

Figure 7.15: Waveforms of *Backward* impulse responses synthesized in both rooms with $T_{trans} = 80$ms. An important lack of early reflections energy is noticeable in both room conditions and especially for the *Gallery*

The lack of early reflections might have provoked a reduced masking of the late reverberation. Temporal masking is a suppressing mechanism of the human auditory system that masks, both in frequency and time, lower level sounds after the onset of a stronger sound. This temporal masking is referred to as the "precedence effect" [61] or the "Haas effect" (for a single echo masking the perception of speech [69]). In a room impulse response, the situation is more complex than a single masker and a masked sound. As a result of the precedence effect, both the direct sound and reflections can be considered as maskers of the late reverberation [73]. Hence, the combination of direct and reflected sounds is heard as a single entity, and the perceived location corresponds to the direction of the direct sound. It is generally admitted that the temporal limit of the Haas effect for speech is 50ms [73]. Another study by Meesawat and Hammershoi [115] investigating this effect

within RIRs reported that the masking of late reflection by early reflections could extend to 40ms. Here, particularly for T_{trans} = 80ms, the lack of energy in the early reflections has reduced the Haas effect. Then, the late reverberation seemed perceptually more salient in the synthesized responses of the *Gallery* when compared to measurements, despite their late reverberations being strictly identical. As the perceived reverberation seemed stronger, the rendered stimuli could have been associated with a larger environment [147]. Moreover, as mentioned by several studies [32, 91] the perceived room size affects auditory distance judgements, a larger room size implies an auditory distance perceived further.

As a result of the method's design, this phenomenon had the opposite effect on perceived distances as anticipated by the energetic differences. For the *Classroom* as the reverberation time is lower, the masking effect is less important. Thus, for T_{trans} = 80ms, we still witness the effect of the excess of direct sound and first reflections, revealed by an underestimation of the perceived distances at 3 and 4 meters.

The presence of this "unmasking" effect prevents us from concluding clearly on the effect of the *Backward* synthesis in the *Gallery*. However, in the *Classroom* the effect seems negligible. Indeed, the maximal difference in terms of D/Ref with the measurements is much lower for BRIRs synthesized in the *Classroom* (9dB, see Figure 7.9), when compared to the BRIRs synthesized in the *Gallery* (18dB, see Figure 7.10). As a result, the energy distribution disparities produced by syntheses are larger in the *Gallery* than in the *Classroom*.

These results seem to corroborate that a transition time of 40ms, can be considered as an appropriate time limit for defining an early-to-late energy ratio as a relevant reverberation-related distance cue, as long as the disparities with actual measurements in terms of D/Ref are less than 9dB.

7.8.2 Forward *stimuli*

In the *Classroom*, the expected effect of energy differences on reported distances is observed. For T_{trans} = 80ms, close distances ranging from 1 to 3 meters, are overestimated. For T_{trans} = 40ms, the effect is weak, but still significant as reported distances for 1 meter are overestimated.

In the *Gallery*, the effect is almost insignificant. For a distance of 1m, and T_{trans} = 80ms, the distance is overestimated. For T_{trans} = 40ms, no significant differences with reference stimuli in terms of perceived distance are present, independently of the internal distribution of reflections within the early part.

These results corroborate that an early-to-late ratio with a temporal limit of 40ms is probably relevant as a reverberation-related distance cue. Its correct reproduction induced a single significant difference in distance reports when compared to reports of the reference stimuli.

The incorrect reproduction of additional reverberation-related cues could explain this single difference, and the differences between *Backward* and reference stimuli. The following sections discuss how observed differences in terms of per-

ceived distance can be connected to the incorrect reproduction of spectral or spatial characteristics.

7.8.3 *Spectral aspects*

Larsen et al. [97] showed that spectral cues account for the role of reverberation in auditory distance perception. They have proposed two different parameters to measure these spectral changes with distance: the spectral centroid and the spectral envelope. Prud'homme and Lavandier [141], have proposed the spectral balance as a parameter integrating a part of the content of both spectral cues. It is defined as the difference between the high-frequency ($>$ 1250Hz) and the low-frequency ($<$ 400Hz) sound levels contained in the stimuli. Spectral balance is negatively correlated with distance. A similar parameter was used here to quantify reverberation-related changes in the spectral content of the used stimuli.

In this Experiment, the *Backward* and *Forward* methods produced respectively an excess or a deficiency of direct sound when compared to measurements. This segment of the BRIR carries a large proportion of high frequency in comparison to early reflections. As seen in Figure 7.9 and Figure 7.10 the D/Ref in synthetic BRIRs is constant, contrarily to actual measurements. Thus, when distance changes, *Backward* and *Forward* stimuli exhibit fewer spectral changes than measurements. For all distances, *Backward* BRIRs yield more direct sound than measurements and thus carry more high frequencies. Conversely, *Forward* BRIRs carry less high frequencies.

In order to illustrate the lack of spectral changes in speech stimuli generated with the *Backward* and *Forward* syntheses, spectral balances of the stimuli were computed. The values are displayed in Figure 7.16. Because the frequency limits used by Prud'homme and Lavandier (respectively 400Hz and 1250Hz) produced spectral balances weakly correlated with increasing distances, a higher frequency limit of 2000Hz was used to compute the spectral balance of each stimulus. This could be explained by the difference in the nature of the stimuli employed in this experiment. While Prud'homme and Lavandier convoluted pink noise bursts, a speech signal with a narrower spectral envelope was applied here.

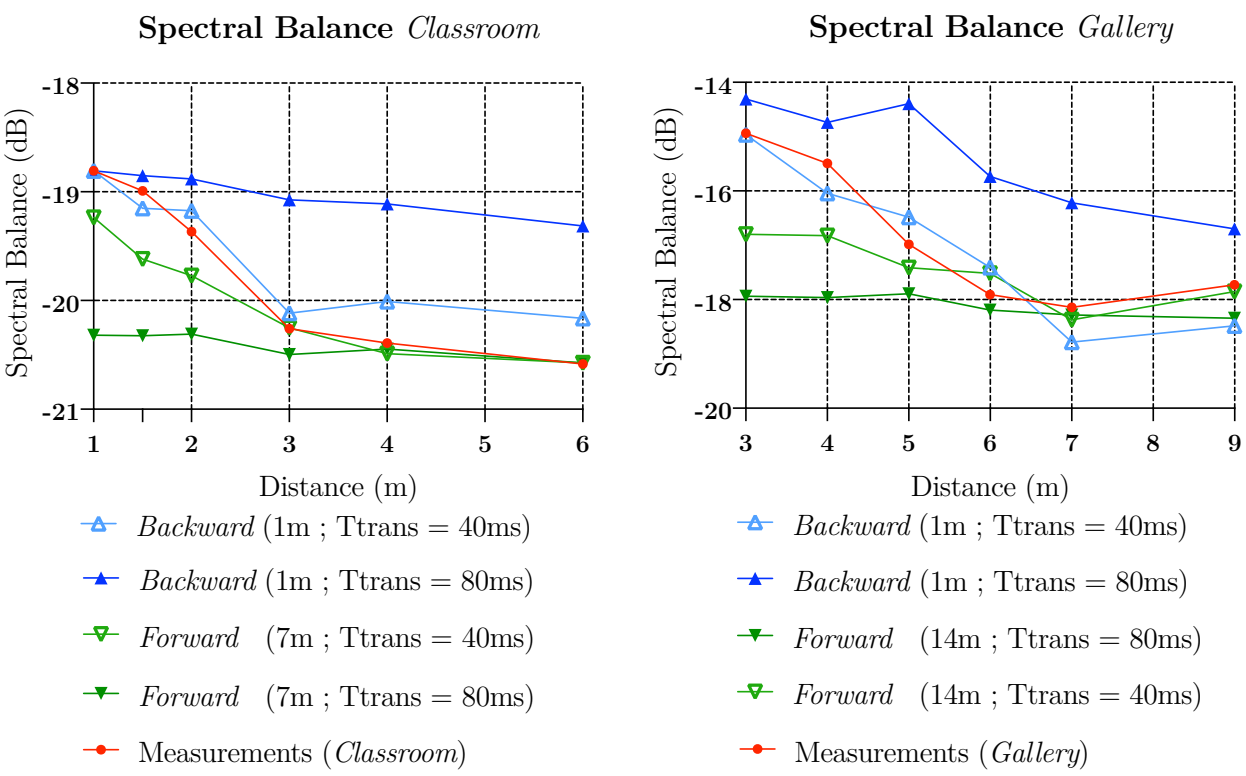


Figure 7.16: Spectral Balance of the stimuli used in Experiment II computed with the difference between the high frequency sound level (> 2000Hz) and the low-frequency sound level (< 400Hz) contained in each stimulus. The value of both channel was averaged to compute the spectral balance of the resulting stimulus.

In both rooms, for *Backward* and *Forward* stimuli rendered with $T_{trans} = 40$ms, the spectral balance is correctly reproduced with distance, except *Forward* stimuli rendered for distances < 2m. This could explain the single difference observed for the *Forward* stimulus at 1m in the *Classroom*.

For $T_{trans} = 80$ms, the spectral balances of stimuli are almost constant with distance when compared to measurements. Thus, reverberation-related spectral cues become unreliable for inferring a distance judgment. The incorrect reproduction of these spectral aspects might have participated in the observed perceptual differences for these stimuli.

These spectral differences could also explain the absence of significant perceptual differences observed in Experiment I between stimuli generated by the envelope-based model and those produced with real measurements. The spectral balances of the stimuli of Experiment I are displayed in Figure 7.17. Contrarily to the *Backward* and *Forward* methods, the envelope-based model correctly reproduces the spectral balance with increasing distances. Indeed, the design of the model implied a correct reproduction of the direct sound decrease with distance (see Equation 7). High frequencies are mainly conveyed by this part of the BRIRs,

correctly reproducing the direct sound limited spectral differences with the measurements.

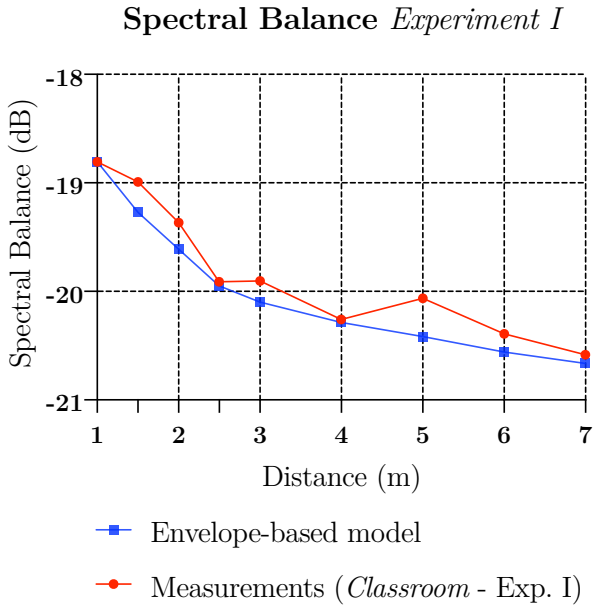


Figure 7.17: Spectral Balance of the stimuli used in Experiment I (envelope-based model and measurements) computed with the difference between the high frequency sound level (> 2000Hz) and the low-frequency sound level (< 400Hz) contained in each stimulus. The values of both channels were averaged to compute the spectral balance of the stimuli.

These spectral considerations suggest that summarizing the perception of reverberation solely on the basis of time-energy considerations is incorrect. Depending on the nature of the stimuli and the spectral content of the BRIR, different temporal segments can be made more perceptually salient than others. Reverberation-related spectral changes are therefore important for auditory distance perception. The importance of the spectral cues for each participant is further discussed in Section 7.8.5.

7.8.4 *Spatial aspects*

Each set of synthetic BRIRs partly retained their related initial measurement's spectral characteristics. Spatial aspects of early reflections were also inherited from the measurement : the directions of arrival of the early reflections remain identical with changing distances.

The main perceptual characteristic influenced by this modification is the Apparent Source Width (ASW). It can be referred to as "the perceived width of a sound image fused temporally and spatially with the direct sound image"[127]. There is no clear consensus on how ASW should be quantified. However, it is assumed to be

negatively correlated to the early Interaural cross-correlation $IACC_E$ [78] as well as to the early Lateral energy Fraction LF_E, and positively to early sound strength G_E [126, 133]. These parameters are generally computed on the first 80ms of an impulse response [25]).

The strength of sound (G) is defined as the energy of the impulse response measured using an omnidirectional microphone, relative to the energy of the same source measured at a 10m distance in a free field. Due to the design of the method, the early strength of each of the synthesized BRIR was identical to that of the corresponding measurements.

Interaural cross-correlation is a measure of the similarity between binaural signals. It was calculated on the first 80ms of the BRIRs used in this experiment. The values $(1 - IACC_E)$ (that could be considered as a quantification of ASW) are displayed in Figure 7.18. The just-noticeable differences in terms of $IACC_E$ have an average value of 0.075 [44] or 0.065 [132] for music stimuli.

In regard to these values, $IACC_E$ stayed constant with changing distances for T_{trans} = 80ms, and can be considered almost constant for T_{trans} = 40ms. For extreme distances significant differences in terms of $IACC_E$ between *Backward*, *Forward* and reference stimuli are observed.

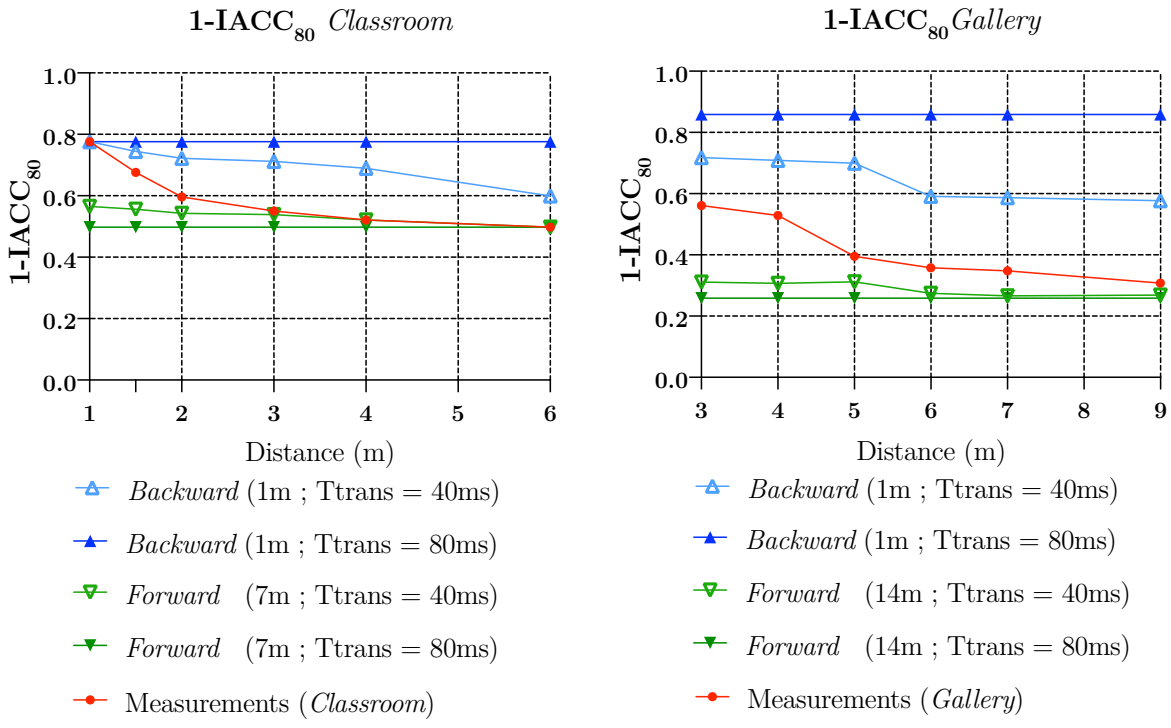


Figure 7.18: Interaural cross-correlation of the BRIRs used in experiment II, computed on the early part [0; 80ms] of each responses.

Early Lateral energy Fraction LF_E is defined as the linear ratio of the lateral early energy to the total early energy [17], generally computed on the first 80ms [133]:

$$LF_E = \frac{\int_{5ms}^{80ms} p_L^2(t)dt}{\int_0^{80ms} p^2(t)dt} \tag{15}$$

With $p_L(t)$ the impulse response measured laterally with a figure-of-8 microphone (corresponding to the Y component of an SRIR) and $p(t)$ the impulse response measured with an omnidirectional microphone. Here the LF_E was computed on *Backward* and *Forward* synthesized SRIRs as well as on the actual measurements. The values are displayed in Figure 7.19. Similarly to what was observed on the $IACC_E$ values, LF_E values of synthesized SRIRs are almost constant with changing distances. Cox et al. [44] determined that the just-noticeable difference of LF_E is equal to 0.06. Thus, significant differences are present between synthesized and measured responses for extreme distances.

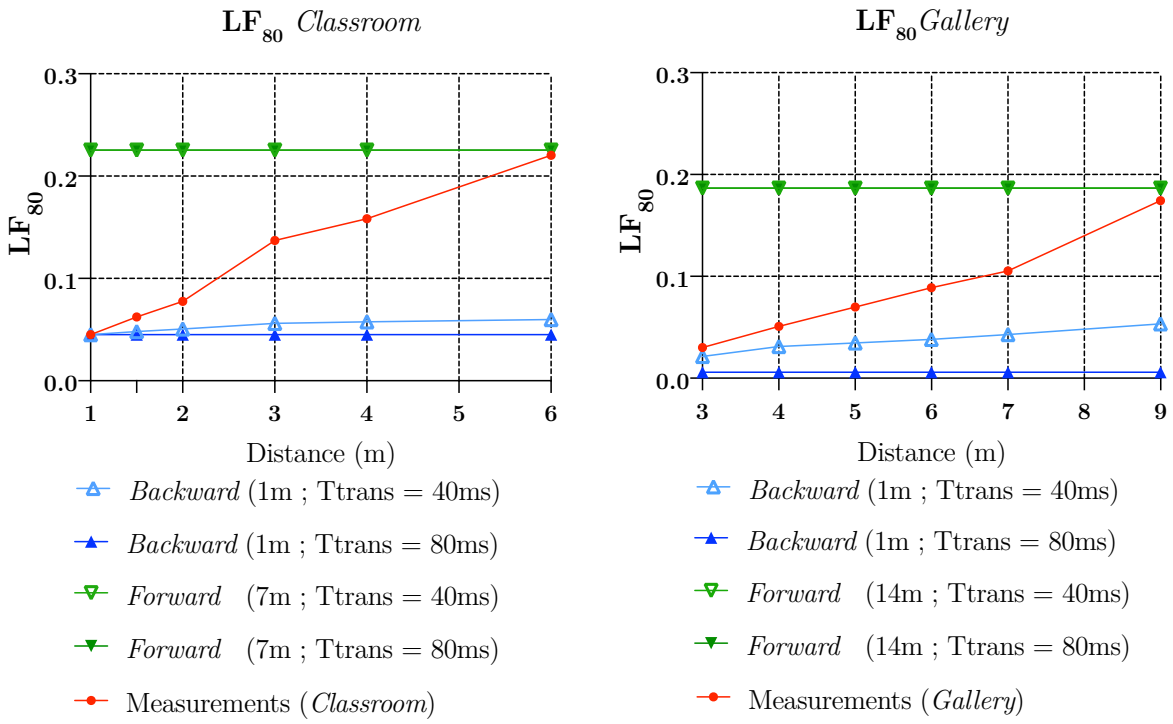


Figure 7.19: Early Lateral energy Fraction LF_E of *Backward* and *Forward* synthesized SRIRs and measured SRIRs, computed on the early part $[0; 80ms]$ of each responses.

The absence of variation for both parameters, induced constant values of ASW with distance for the *Backward* and *Forward* stimuli. While it is well known that this parameter decreases with distance [98], the effect of ASW on distance is still unexplored. The literature reports contradictory findings about the impact of reverberation's spatial aspects on auditory distance perception.

Bronkhorst [26] have demonstrated that the degree of lateralization of the early reflections could influence their fusion with the direct sound for auditory distance perception. In this regard, the incorrect reproduction of these spatial aspects could have participated in the perceptual differences observed between the syntheses and the reference. For the *Forward* stimuli at short distances, the increased lateralization of the early reflections could have induced a stronger integration of early reflections into the late reverberation. As a result, it could explain why these stimuli distances are overestimated in both rooms. Conversely, for *Backward* stimuli at large distances, the lack of early reflections' lateralization could have contributed to the underestimation of distances observed in the *Classroom*.

However, studies by Larsen et al. [97] and Prud'homme & Lavandier [141] did not observe significant effects of spatial aspects on auditory distance perception. Both these studies have addressed the influence of spatial aspects by comparing monaural and binaural RIRs. As a result, spatial cues have not been recognized as directly influencing auditory distance perception, but rather externalization. However, the lack of externalization could prevent some participants from accurately judging the distance of a stimulus [141]. Participants experiencing in-head localization are most likely to be affected by spatial aspects, which were considered outliers and not taken into account in the current analysis. Therefore, observed differences in distance reports are most likely not related to spatial cues. Nonetheless, this assumption needs to be tested in a study that particularly examines the impact of early reflections spatialization on auditory distance perception, as well as its relationship with externalization.

7.8.5 *Reverberation-related cues weighting strategies*

Similarly to Experiment I analysis, the distributions of dots displayed in Figure 7.12 and Figure 7.14 indicate that some participants present perceptual similarities across distance reports of the different categories of stimuli. Participants that demonstrate little to no differences probably highly weight the early-to-late energy ratio as a reverberation-related distance cue, and weakly weight spectral and potential spatial cues. This result suggests that perceptual weighting strategies for reverberation-related distance cues are mainly idiosyncratic.

7.9 CONCLUSION

In this chapter, we presented two experiments investigating the role of early energy in auditory distance perception. Under the hypothesis that the DRR can be overtaken by an early-to-late energy ratio for defining a relevant reverberation-related distance cue, we examined if an appropriate temporal limit could fit our initial hypothesis. Results associated with the *Classroom* show that for a transition time of 40ms, both syntheses correctly reproduced sound source distance in the *Classroom*. In the *Gallery* only the *Forward* method achieved the correct reproduction

of distance for this transition time. These results suggest that when considering a transition time of 40ms, the distance judgements are close to the measurements and nearly independent of the internal distribution of reflections. It is the case as long as the D/Ref ratio does not differ by more than 9dB, as in the *Classroom*. In contrast, for $T_{trans} = 80$ms, the distance judgments are strongly dependent on the internal distribution of reflections and differ significantly from the distances reported with the measurements. Further studies should be conducted to determine more precisely, what the optimal transition time is.

The different results also suggest that reducing the role of reverberation for distance by an early-to-late energy ratio only is oversimplifying. Notably, the lack of reflections before the transition time in Experiment II has considerably biased the responses associated with *Backward* stimuli in the *Gallery*. The reduced Haas effect on the late reverberation has considerably changed the perception of the room volume in which the stimulus is displayed. This effect illustrates that the energetic content of early reflections can modulate auditory room size perception. Room size perception and auditory distance perception are two intricate percepts that can hardly be considered independent.

Moreover, the incorrect reproduction of spatial and spectral cues can also explain the differences between distance reports associated with syntheses and the reference. Both type of cues are incorrectly reproduced by the sytheses, when considering a transition time of 80ms. The results suggest that the correct reproduction of spectral changes in the early part of BRIRs with distance is mandatory for auditory distance rendering methods. The individual results of participants illustrate that the perceptual weights attributed to each spectral and energetic reverberation-related cues to infer a distance percept are an idiosynchratic characteristic. Further studies isolating changes in spectral cues should be done to confirm these assumptions.

The question of early reflections' spatial characteristics could also be considered. Our initial hypothesis was based on the fact that direct sound could not be effectively separated from the early reflections by the auditory process. It can be argued that the direction of arrival of early reflections could participate in the fusion of early reflections with direct sound. It can be hypothesized that if the direction of arrival is close to the direction of the direct sound, the spatial proximity might enable a stronger fusion process. The use of SRIR spatial manipulation such as the "warping" [96] could enable the study of this effect.

8

EVALUATION OF THE INFLUENCE OF ENVIRONMENT-RELATED CUES

This chapter presents and discusses the results of Experiment III that aims to study the influence of environment-related cues on the auditory distance perception of virtual sound sources. The influence of the visual spatial boundary and of the room volume are studied through an online protocol. The stimuli used in this experiment and in Experiment I are the same, allowing a comparison between acoustic cues weighting strategies of participants in both experiments.

8.1 INTRODUCTION

An aspect that must be considered for auditory distance perception in an AAR scenario is the influence of the visual environment. Indeed, in most cases, users of an AAR scenario see their own environment while it is enhanced with virtual sound sources. The geometry of the room is not necessarily totally known and implemented in AAR applications, and several studies have suggested that the visual context influences distance perception of the sound sources through mechanisms related to multisensory integration [33, 166]. Incongruent visual context was also demonstrated to have an effect on the externalization of virtual sound sources [176].

It has long been argued that vision could calibrate auditory distance perception [171, 172]. Calcagno et al. [33] demonstrated that a priori visual information about an environment can be beneficial to the accuracy of auditory distance perception of real sound sources.

The authors hypothesized that the representation of the visual space can serve as a spatial reference into which distance cues are integrated to create an auditory distance judgment. Consequently, the reference space conditioned by visual information scales auditory distance perception and can compress or expand the auditory distance responses, depending on the congruence of the acoustic information with visual spatial cues. This hypothesis about calibration of space perception has been tested with stimulation of a single or multiple sensory modalities. Previous studies have notably demonstrated that distance perception responses can be calibrated by information collected by the same or a different sensory modality. For instance, Etchemendy et al. [56] demonstrated how sound reverberation conditions affect visual distance perception, with highly reverberant rooms implying an overestimation of visual distance perception.

Given these results about spatial calibration of a sensory modality, the specific impact of visual spatial boundaries on the perception of virtual sound sources is addressed in this experiment.

In Experiment I, we have shown that despite a similar listening situation, different acoustic cues weighting strategies were employed by participants. This finding suggests that the weight attributed to each cue is mainly idiosynchratic. In this experiment, environmental characteristics influence on the acoustic cues weighting strategies used by participants to infer an auditory distance judgement is investigated.

8.2 EXPERIMENT III: EVALUATING THE INFLUENCE OF INCONGRUENT VISUAL CUES

8.2.1 *Objective of the experiment*

The aim of this experiment is twofold: 1) to investigate the influence of two environmental characteristics, the visual spatial boundary and the volume of the environment, on the auditory distance perception of a virtual sound source 2) to evaluate the impact of these characteristics on acoustic cues weighting strategies.

8.2.2 *Material & Methods*

The experiment was carried out using the previously stated online procedure (see Section 6.2).

The evaluation of the acoustic cues consisted of online listening tests using headphones in which participants had to evaluate the distance of virtual sound sources produced by the same stimuli employed in the previous Experiment I (see Section 7.2).

An experiment gathering two pools of 60 participants each, were ran with two different inclusion criteria on the distance the user should be to the wall she or he is facing. In the *Close Wall* group (CW) of participants, the seating distance from the wall they were facing was required to be less than 3m. Participants who were seating at a distance of more than 5m from the wall were included in a *Far Wall* group (FW).

The analysis of the differences between both groups' results was made to investigate the influence of this visual spatial boundary condition on auditory distance perception. Participants of both groups had also to report the dimensions of the room in which they were during the experiment. The volume estimated from this report was considered as a primary descriptor of the quantity of reverberation in the room. As participants were each accustomed to the specific acoustic properties of the room where they were running the experiment, we decided to study through each participant's reported room volume, the influence of a possible expectation on auditory distance perception.

8.2.2.1 *Participants*

Online participants were recruited via the Prolific recruitment platform. A total of 120 participants were recruited and paid to complete the task. Initial inclusion and exclusion criteria reported in Chapter 6 were applied. Additionally, participants were screened on the basis of the self-reported distance from the wall they were facing. When the experiment had reached the targeted number of participants, data were manually accepted into the final results pool after inspection of the participants' individual data. After identification of statistical outliers, fifty-five out of sixty participants (age range 18 to 42 ; $M = 25$; $SD = 6.5$; 23 females, 32 males) were included in the CW group. Fifty-three out of the sixty participants were included in the FW group (age range 18 to 46 ; $M = 25$; $SD = 6$; 23 females, 30 males).

8.2.2.2 *Auditory Stimuli*

The rendering methods, speech stimuli, simulated acoustic environment and source to receiver distances are exactly the same as those used for Experiment I (see Section 7.2). A 5-second anechoic speech recording was pre-convoluted with each of the BRIRs, either measured or generated by the models. A total of 27 stimuli, simulating 9 different sound source distances for both rendering models and 9 stimuli based on the measures for the same distances, were used. Each stimulus apparent distance was evaluated 4 times throughout the test, for a total of 108 stimuli presentations.

8.2.3 *Procedure*

After reading the information note and approving the consent form, participants were directed to the experiment.

Participants had to fill a questionnaire in which they were invited to report the main geometric features of the room where they were running the experiment: surface (in square meters), height of the room (in meters), as well as the distance to the wall (in meters) they were facing. They were requested to use circum-aural headphones.

A screening process was launched to ensure the correct use of headphones: two audio clips corresponding to the weakest and loudest stimuli were played successively. Participants had to adjust the sound level for the quietest stimulus to be heard clearly. The loudest stimulus was then played, to ensure that it was heard at a comfortable level. Finally, broadband noise was displayed laterally to ensure that participants were correctly wearing their headphones.

Afterwards, a training session began. Participants were instructed to look at the facing wall while listening to a stimulus, and to judge the distance of the stimulus using the slider. Every 27 stimuli available (27: 9 distances $\times$ 3 rendering methods) were presented in a random order in this training session.

After the training session, the experiment began. The order of the stimuli within the experiment was randomized. During the trials, participants triggered the stimulus playback but the stimulus was only played once.

At the end of the experiment, participants were invited to fill out a final questionnaire to collect feedback on the procedure. The total duration of the procedure was 15 minutes. Two weeks later, a second debriefing was conducted through Prolific.co electronic mail service, to ensure the validity of the answers about the room characteristics.

8.3 EXPERIMENT III: RESULTS

Statistical analyses were performed using *TIBCO Statistica*© except for the power-function fittings, which were performed using *Mathworks Inc MATLAB*©.

OUTLIERS Five participants (1 female, 4 males) from the CW group and 7 participants from the FW group (2 females, 5 males) were excluded from the analysis. Six of them were excluded because the mean and standard deviation of their responses in most conditions exceeded more than two standard deviations from the mean of the population. Furthermore, their mean response time was substantially short and without variation, giving us reason to suspect that they did not correctly perform the task. Six of them were excluded because they did not participate in the final debriefing. The analyses were conducted on the remaining 55 participants (21 females, 34 males) for the CW group and the remaining 53 participants (23 females, 30 males) for the FW group.

For comparison purposes, the reported distances that will be considered in the fitting, result from a linear conversion of scaled reports made by participants on the visual analogue scale. The maximum of the visual scale (0% on the slider corresponds to the minimal distance of 0 meter and 100% to the maximum possible distance of 7 meters).

8.3.1 *General Results*

In the following statistical analysis, the logarithmic value of the geometric mean of each participant's responses over a type of stimulus was considered as a dependent variable (27 different stimuli: 9 distances $\times$ 3 rendering methods). For both groups separately, a focus was put on ensuring the normality of the dependent variables associated with a specific stimulus. A Jarque-Bera test indicated that in all cases for both groups, the null hypothesis "the data were normally distributed" was not rejected.

A repeated measures ANOVA ($\alpha = 0.05$) has been conducted on geometric mean of reported distances of each subject with RENDERING (3 rendering methods) and DISTANCE (9 levels ranging from 1 to 7m) as within subjects factors and GROUP (2 Groups) as inter-subject factor.

The analysis revealed:

- a main effect of DISTANCE ($F(1,8) = 580.8$, $p < 0.001$)
- a main effect of RENDERING ($F(1,2) = 68.0$, $p < 0.001$)
- a DISTANCE $\times$ RENDERING interaction ($F(1,16) = 11.7$; $p < 0.001$)
- a GROUP $\times$ DISTANCE interaction ($F(1,8) = 2.05$; $p = 0.039$)
- and no significant main effect of GROUP was revealed ($F(1,1) = 0.74$, $p > 0.05$).

8.3.2 *Effect of room volume*

Participants had to answer questions about the boundaries and dimensions of the room they were in. Their answers were used to estimate the volume of the room in which they performed the task (see Figure 8.2). In the FW group, the estimated room volumes were larger on average and more dispersed than in the CW group. A non-parametric statistical test confirmed that both distributions were significantly different (Kruskal-Wallis: $H = 16.85 p < 0.001$). Consequently, in the following analysis, the influence of volume on auditory distance perception will be investigated separately in each group. For each group, separate linear regressions were applied to responses associated with each specific stimulus (27 different stimuli: 9 distance $\times$ 3 rendering methods) with the room volume as a factor. In the CW group, significant correlations between the room volume and distances longer than 4 meters generated with the intensity-based model were found ($D = 4m$ $F = 4.16$ $p = 0.034$; $D = 5m$ $F = 6.05$ $p = 0.017$; $D = 6m$ $F = 5.06$ $p = 0.028$; $D = 7m$ $F = 6.40$ $p = 0.014$). For these specific stimuli only, larger volumes were significantly correlated to a larger compression of the reported distances. No significant correlations were found for shorter source distances ($p > 0.05$) and stimuli generated by the other rendering methods, nor for all stimuli in the FW group.

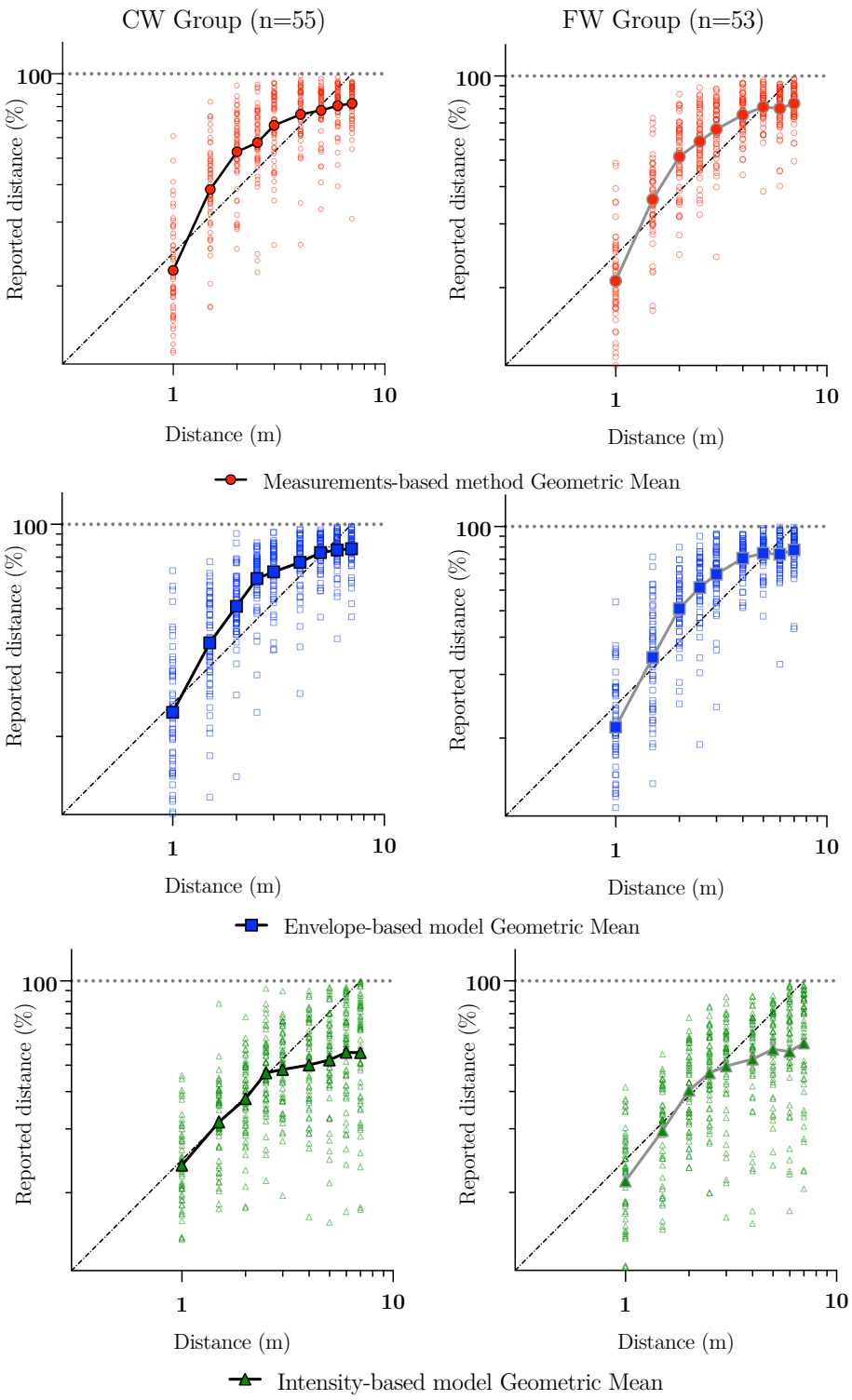


Figure 8.1: Geometric means of reported distance according to the rendering method used to generate the stimuli and the visual spatial boundary condition. Clear dots represent the geometric mean of a single participant for a given distance, opaque dots represent the geometric mean over all participants for a given distance.

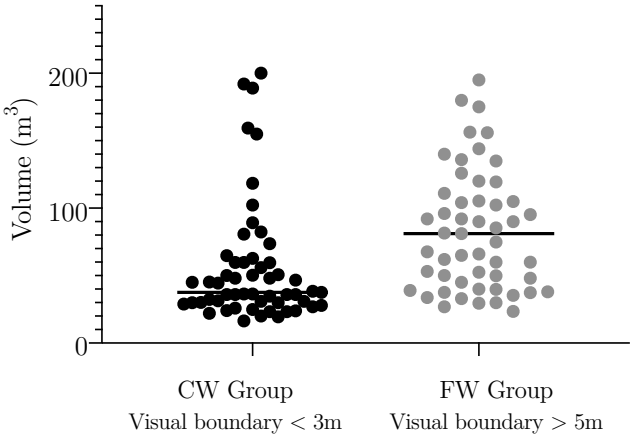


Figure 8.2: Estimated room volume of participants in each group (CW Group: $M = 37.5m^3$; FW Group: $M = 81.7m^3$).

8.3.3 *Compression effect quantification across rendering methods*

The compression effect is quantified with the fitting method described in Section 3.1. The quality of the fitting is quantified with the cross-correlation coefficient R^2. For each rendering method, a positive correlation was found between a fitted power function and the distances reported by each participant for the rendering method based on measurements ($Mean(R^2) = 0.73; SD = 0.12$) and for the envelope-based model ($Mean(R^2) = 0.75$; $SD = 0.11$). A slightly smaller but positive correlation was found for the intensity based model ($Mean(R^2) = 0.68$; $SD = 0.28$). The mean value of R^2 obtained on reported distances of stimuli generated by the intensity-based model is significantly lower than the mean obtained on the envelope-based model. In that respect, stimuli rendered with the intensity-based model tend to produce significantly higher intra-subject variability.

To further investigate the similarity between the rendering method based on measurements and the envelope-based model at an individual level in both groups, each participant's compression coefficients were compared. In Figure 8.3, the individual compression coefficients a and k collected among the two groups are contrasted between those estimated for each model (ordinate) and for the rendering method based on measured BRIRs (abscissa). A linear regression analysis was run for each distribution of dots. In all eight analyses, the null hypothesis "the coefficient of the regression slope is equal to zero" was rejected ($p < 0.001$). The distributions evaluating individual similarity between the envelope-based model and the measurements produce regression slopes close to 1. While the distributions related to the intensity-based model produce regression slopes significantly inferior to 1.

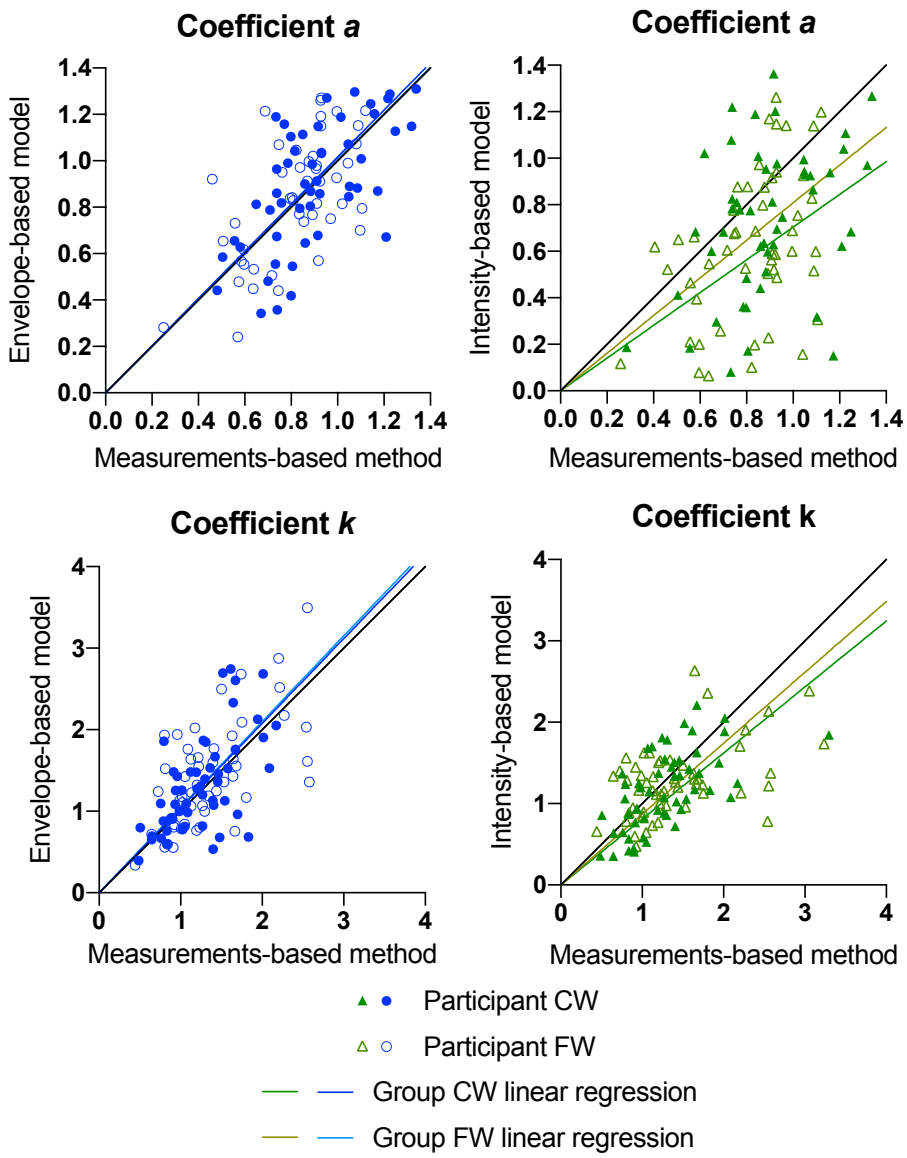


Figure 8.3: Comparison of the individual compression coefficients a (above) and k (below) between the rendering method based on measurements and the models. On the left, the Envelope-based model (blue). On the right, the Intensity-based model (green). Open symbols correspond to coefficients estimated from participants of the FW group, and plain symbols of the CW group. Light lines correspond to the regression curves associated with the FW group, dark lines with the CW group.

The distribution of compression coefficients a and k indicates the presence of a similarity, in terms of auditory compression effect, between the envelope-based model and the rendering method based on measurements. In contrast, the distribution of coefficients for the intensity-based model is more dispersed, showing less correlation at an individual level. The distribution also highlights that most

compression coefficients a estimated for the intensity-based model are lower than those obtained for the rendering method based on the measurements.

8.3.4 *Influence of the visual spatial boudary on compression coefficients*

In order to study the influence of the environment on auditory perceived distance compression, the following analysis method, inspired by the work of Anderson and Zahorik [7] was applied. The mean values of compression coefficients for each distance rendering method are displayed in Figure 8.4 for each group.

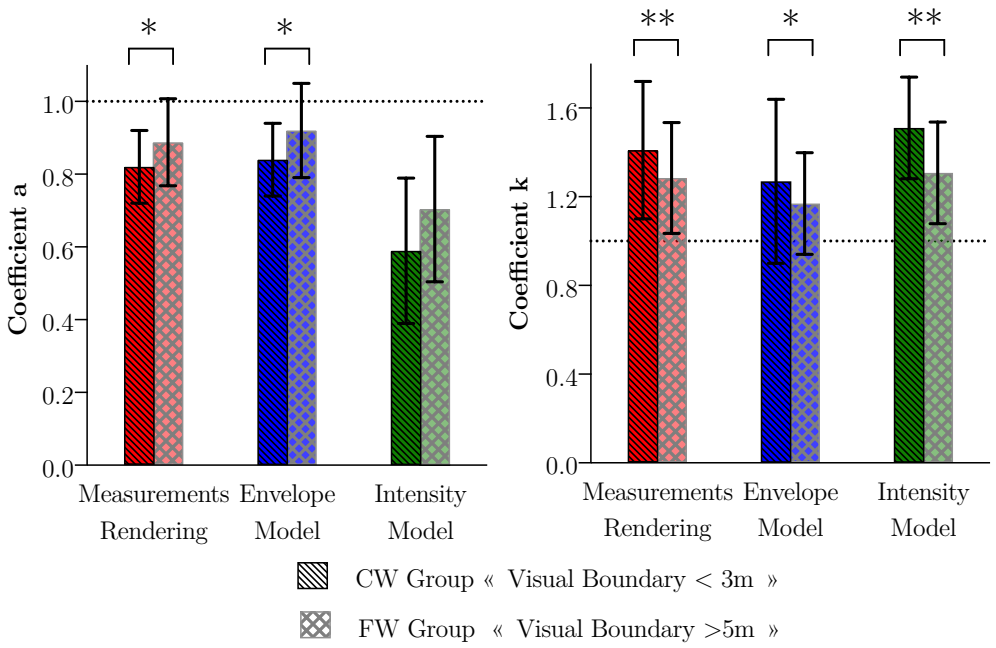


Figure 8.4: Mean and standard deviation of individual fitting coefficient a (left) and k (right) classed by groups (CW group (black, $n = 55$) and FW group (grey, $n = 53$)) and by rendering methods: measurements-based (red), envelope-based model (blue) and the intensity-based model (green). Parameters for FW group are closer to 1 compared to those obtained for CW group demonstrating a lower compression effect of auditory distance perception. ($*$) t-test indicated a p-value < 0.05 ($**$) t-test indicated a p-value < 0.01

The mean values of the coefficients a and k, being both closer to 1, suggest that the compression effect is weaker in the FW group. For each model, independent sample t-tests have been conducted between the values of coefficients a and k estimated from the data of the CW group and those estimated from the data of the FW group. The null hypothesis "the two population means are equal" is rejected for all tests, except for the comparison of the values of the coefficient a for the intensity model. According to these results, the visual spatial boundary had

a moderate but significant effect on the compression effect of auditory distance perception.

8.3.5 *Influence of the room volume on compression coefficients*

The effect of room volume on reported distances was investigated by running a linear regression analysis applied to the individual α compression coefficients estimated in the CW group for the intensity-based model (see Figure 8.5). The logarithm of each reported volume was considered as factor for this regression analysis. According to the results of this analysis, a higher volume is significantly correlated with a lower parameter α ($F = 5.60$; $p = 0.021$). This result confirms that a larger volume is linked to a stronger compression effect of reported distances of stimuli generated by the intensity-based model.

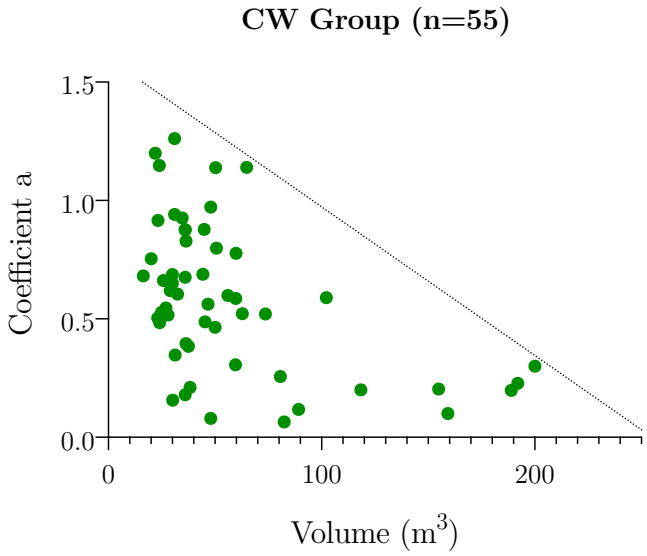


Figure 8.5: Value of the fitting coefficient α (green dots) obtained on the intensity-based model, for participants of Group CW as a function of the self-reported volume of the room.

8.4 DISCUSSION

The current experiment evaluated the influence of environmental context cues on auditory distance perception. The effect of the visual spatial boundary on the auditory distance perception was assessed by comparing the value of the mean compression coefficients α and k between both groups.

The integration of the same set of stimuli used in Experiment I allows us to corroborate the results about acoustic cues weighting strategies (see Section 7.5.3).

The impact of room volume was investigated through linear regression fitting on distance reports and each rendering method associated compression coefficients.

8.4.1 *The influence of the visual spatial boundary*

The influence of visual incongruence was studied by investigating differences between the CW and FW groups. Results in Figure 8.4 demonstrates that FW group's participants that were exposed to a longer spatial boundary, present a smaller compression effect. This result can be compared to past studies concluding that the presence of congruent visual cues could enhance the accuracy of auditory distance perception [7, 33, 86].

In this experiment, a visual incongruence originates from the discrepancy between the reproduced room and the experimental environment specific to each participant. The geometrical characteristics of the reproduced room are, however, closer to the dimensions reported on average by the participants of the FW group than those reported in the CW group. The volume of the reproduced room is 144m^3, while the mean volume is 81.7m^3 in the FW group and 37.5m^3 in the CW group. The presented stimuli intend to reproduce auditory distances up to 7m, while the visual spatial boundary is limited to 3m in the CW group and ranges from 5 to 10 meters for the FW group. Therefore, the presence of more congruent visual cues might have enhanced the auditory distance perception accuracy of FW group.

It has been long argued that the visual context in which auditory judgements occur, contributes to the organisation of the auditory space [33, 171, 172]. This finding about the influence of the visual spatial boundary could be seen as an extension of the hypothesis discussed by Calcagno et al. [33]. In their experiment, participants had to evaluate the distance to a real loudspeaker, presented frontally and emitting white noise bursts. As a result, no room divergence was present. Participants had access to an increasing quantity of visual cues, ranging from being blindfolded to having full vision of the room. Intermediate scenarios in which only 2 or 4 LEDs were lit during the reports were also examined. The authors have demonstrated in this acoustically congruent situation that the presence of congruent visual cues increases the accuracy of auditory distance perception. They hypothesized that visual range information helped calibrate the reported distances of real sound sources, visual distance perception being a sensory modality far more accurate for distance estimations [7]. In the current experiment, the visual spatial boundary condition could also have acted as a calibration of auditory distance perception in an acoustically divergent situation. A long visual spatial boundary condition might have influenced the representation of the auditory space and incited participants to expand the range of the reported distances. This influence on the reporting strategy explains the significantly lower compression effect of the FW group.

8.4.2 *The influence of volume on acoustic cues weighting strategies*

This experiment investigated the influence of volume on auditory distance perception. Results in Section 8.3.5 demonstrated a slight influence on the reported distances of stimuli generated by the intensity-based model. This specific result illustrates the influence of the reported volume on the compression effect of participants on the intensity-based model.

The effect of volume is not observable on the compression coefficients associated with participants of the FW group who benefited from a longer visual spatial boundary.

In the CW group, the mean and maximal value of the non-linear compression coefficient α tend to a smaller value when volume increases. This can be regarded as a reduction in the number of strategies observed for producing auditory distance judgments when the volume increases. In large volumes ($> 100\text{m}^3$), participants seem to only rely on a reverberation-related cue to produce an auditory distance judgement. Under the hypothesis that a larger volume generally leads to the perception of a more reverberant environment [32, 91, 117, 147], this effect can be explained by the expectation induced by these environments. Participants who expect a large variation in a reverberation-related cue may be more likely to rely on it.

It is possible that the statistical effect observed on the α coefficient values is due solely to the 7 participants with a room volume greater than to 100m^3. These participants potentially adopted the same strategies for reporting distances, that imply mainly relying on a reverberant-related cue, without being influenced by their perceived acoustic environment.

8.4.3 *Experiment limitations*

8.4.3.1 *Online aspects*

Practical difficulties linked to online-based experiments may reduce the significance of the observed effects and limit their comparison with similar lab-based studies. The inherent lack of control on the participants' experimental conditions, such as the use of diverse headphones models and different environmental noise levels, may have increased the variability of the reported distances.

Experiment I could be reproduced not only in a congruent environment (*Classroom*) but also divergent situations. A totally new environment could be used, or modification of the congruent environment visual spatial boundary could be also tested. This could be performed by adding a movable surface, such as a curtain, to induce a closer visual spatial boundary. This specific scenario could permit to assess the agnostic nature of the visual spatial boundary criterion.

8.4.3.2 *Binaural reproduction & externalization*

In this experiment, the influence of divergent environmental characteristics on auditory distance perception was examined. Participants were asked to assess the perceived distance of auditory stimuli, but no reporting about externalization was proposed. Externalization and auditory distance perception are two related concepts, however primarily depending on distinct cues [21]. The presence of binaural cues in the sound signal is mandatory for externalization which can be considered a prerequisite for an accurate distance perception for some listeners [141]. To this respect, the perception of externalisation is regarded as dichotomous whereas distance perception is continuous and mainly driven by monaural cues [21, 90]. Therefore, precautions were taken to favor the externalization, such as preserving binaural cues in the auditory stimuli. Generic HRTFs were applied to generate the binaural stimuli, which is known to increase the risk of in-head localization, especially for frontal sources [164]. However, aside from practical constraints linked to the online-based experiment, this decision was motivated by previous findings showing that non-individualized HRTFs had no impact on the auditory distance perception of frontal sound sources [180].

8.5 COMPARISON WITH EXPERIMENT I

In the current experiment, the same auditory stimuli as Experiment I were employed. It permitted us to observe results similar to the one discussed in Experiment I (see Section 7.5). This similarity enables a confirmation of the findings of Experiment I.

8.5.1 *Envelope-based performances*

As expected, similarly to what can be observed in Experiment I, the statistical analysis demonstrated that distance reports associated with the envelope-based model present no significant differences with distance reports associated with the reference. Moreover, the individual compression coefficients α and k (see Figure 8.3) associated with the envelope-based model and the measurements-based method are similar. This reveals the presence of a similarity between both methods at an individual level for participants of this experiment.

8.5.2 *Acoustic cues weighting strategies*

The distributions of dots in Figure 8.3 concerning the coefficients α and k corroborate what was observed in Experiment I on the participants' acoustic cues weighting strategies. A part of the participants of this experiment appeared to rely mainly on intensity to make an auditory distance judgment, resulting in coefficients that are almost identical for both the intensity-based model and the

measurements-based method. Some participants obtain a compression coefficient a nearly equal to 0 for the intensity-based model only. Hence, these participants do not use intensity as a distance cue, but rely primarily on reverberation-related cues to infer an auditory distance judgement. However, as mentioned in Section 8.4.2 the room volume was shown to influence participants strategies. Larger volumes imply a greater reliance on reverberation-related cues. However, the acoustic cues weighting strategies used by participants seem to be mainly linked to idiosyncratic aspects.

8.6 CONCLUSION

In this chapter, we presented and discussed the results of an online experiment investigating how environmental characteristics influence the auditory distance perception of a virtual sound source. Through statistical analysis of mean distance reports and individual quantification of compression effects, the effect of the self-reported visual spatial boundaries and the room volume was investigated.

First, visual spatial boundaries influence auditory distance perception. The closer the visual boundary is, the more compressed the overall distance judgments are. The effect of this boundary on auditory distance perception is thought to be owing to a partial calibration of the auditory space by vision of the room's apparent boundaries.

Second, the volume of the room also has an effect on the acoustic cues weighting strategies. The larger the room is, the more participants rely on reverberation to judge the auditory distance. This could be explained by the greater expectation of reverberation driven by large rooms. However, the limited number of participants that led to this result casts doubt on the significance of the reported effect. Further lab-based studies under controlled conditions are necessary to confirm this finding.

In AAR applications where real and virtual sound sources are present in the same auditory scene, the occurrence of an acoustic divergence between the reproduced room effect and the real listening environment is an inherent challenge. The room divergence effect is then not only characterized by the incongruence between the reproduced room effect and the visual geometry of the listening environment, but also by this acoustic divergence with the actual room effect. Its impact on the auditory distance perception of virtual sound sources will be assessed in Chapter 8.

9

IMPACT OF THE ACOUSTIC DIVERGENCE BETWEEN REPRODUCED ROOM EFFECTS

This chapter describes Experiment IV conducted to determine the impact of an acoustic divergence between a reproduced and and the actual room effect of the listening room. As seen in the Chapter 8 the visual incongruence has a significant influence on the auditory distance perception of virtual sound sources. Besides visual incongruence, another conflict related to the acoustic aspects may also emerge. Acoustic divergence can occur in AAR applications where real and virtual sound sources are part of the overall sound scene heard by the listener. An acoustic divergence might occur between the synthesized room effect and the listening environment. The problem addressed is twofold: 1) what effect does this divergence have on auditory distance perception, and 2) what characteristic of the divergence is primarily responsible for it?

9.1 INTRODUCTION

One of the main challenges of AAR applications is to faithfully reproduce the acoustic environment of the user, in order to seamlessly blend virtual and real auditory events. In AAR scenarios, the acoustic environment may not be accurately characterized and/or faithfully reproduced. This can lead to a divergence between the synthesized room effect and the real acoustic environment.

Two different situations can be distinguished: 1) a real sound source is at a known position in the environment surrounding the listener, 2) the listener is unaware of the exact position of the real sound sources, and the listener cannot directly use this information to infer a judgement of the relative distance between the virtual sound event and the real source. Then the influence of the acoustic divergence can then be envisioned as a calibration effect. The presence of the real sound sources could induce a processing of the acoustic cues conveyed by the virtual sound sources, based on the acoustic properties of the real environment. A similar effect due to the perception of the visual geometry of the environment was observed in Chapter 8.

Kolarik [91] examined the effect of reverberation on auditory distance perception, demonstrating that a more reverberant environment tends to elicit a larger perceived room size and, therefore, longer auditory distance judgments. In this regard, we investigated if an intra-modal calibration effect could be caused by divergent acoustical information carried by co-occurent sound sources. Here, we address the impact of real co-occurring sound sources by studying it through the generation of a virtual sound scene with divergent room effects. The procedure was designed to replicate an AAR scenario in which a single virtual sound source

distance had to be determined in the presence of two co-occurring real sound sources. To introduce an acoustic divergence, two different types of stimuli were generated, each representing either the virtual or the real sound sources.

The virtual sound source is referred to as the "target stimulus". It was generated at different distances with previously used measurements of the *Classroom*, with a binaural rendering. Because the experiment was run with remote participants, the co-occurring sound sources could not be real sound sources, but sound stimuli generated with a divergent room effect, displayed prior to the "target stimulus" (the virtual sound source). Co-occurrent sound sources are referred to as "anchor stimuli". The anchor stimuli were generated with measurements of the *Gallery* to induce an acoustic divergence.

We expected that the perceived distance to the virtual sound source could be biased depending on the characteristics of the acoustic divergence between the reproduced room effects. Therefore, we studied which component(s) of this divergence, between intensity or reverberation-related cues, is(are) primarily responsible for the calibration effect.

9.2 EXPERIMENT IV: AN ACOUSTICALLY DIVERGENT SCENARIO

9.2.1 *Objectives of the experiment*

The experiment's purpose is to determine the significance of an intra-modal calibration effect caused by the acoustic divergence between the room effects used to generate the "target" and "anchor" stimuli. Furthermore, we evaluate whether partially correcting for this divergence in terms of intensity limits its effect.

9.2.2 *Material & Methods*

The experiment was carried out using the previously stated online procedure (see Section 6.2). A total of 120 participants,separated into two groups of 60, evaluated the distance of target stimuli dispatched in several blocks. Within these blocks, the target is presented after two anchor stimuli. Within a given block, a specific acoustic divergence is maintained constant. Both groups share the same control block, where no divergence is present. However, two different divergent blocks were assigned specifically to each group. Indeed, as online-based studies must be brief in order to maintain participants' attention, all conditions could not be gathered into a single session and assigned to a single group.

9.2.2.1 *Auditory stimuli & Conditions*

Each of the conditions reflects a different scenario of acoustic divergence: a *Control* condition, in which there is no acoustic divergence, a *Divergent* condition, where an acoustic divergence between acoustic cues conveyed by the target and the anchor stimuli is present, finally, an *Intensity-equalized* condition, in which the acous-

tic divergence is compensated in terms of perceived intensity. The latter two differ depending on the group of participants. Each condition is constituted of nine different sound sequences, which are organized as follows:

- The nearest of two anchor stimulus is played.
- The farthest anchor stimulus is played.
- One of the 9 target stimuli is played.

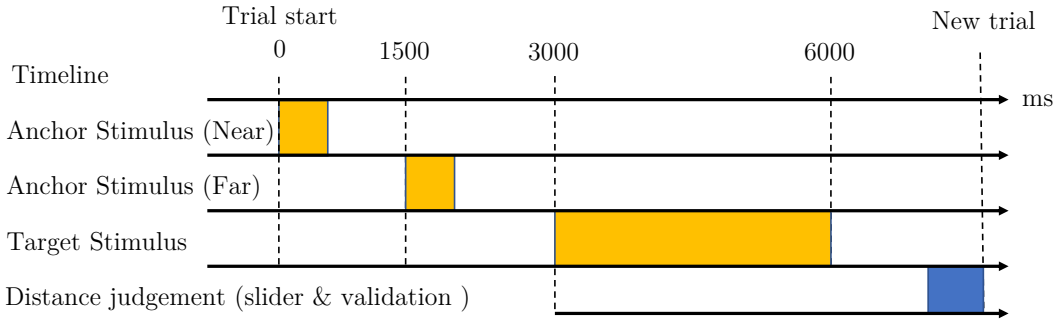


Figure 9.1: Timeline of a sound sequence in each condition of Experiment IV

The acoustic divergence is created by the differences in the rendering methods used for the two types of auditory stimuli:

- **Target stimuli** whose distance is to be evaluated. All nine target stimuli are the same for every conditions and both groups. The 3-second anechoic speech recording used in previous online experiments has been pre-convoluted with each of the BRIRs measured in the *Classroom* (see Figure 7.8). A total of 9 target stimuli were generated at distances ranging from 1 to 7 meters.
- **Anchor stimuli** whose aim is to provide the participant with acoustic information about a divergent listening environment. An anechoic 80ms-long click recording was used to generate a pair of anchor stimuli. It was convoluted with a pair of BRIRs specific to each condition (listed in Table 2).

9.2.2.2 *Groups & Conditions*

	Control condition	*Divergent condition*	*Intensity-equalized condition*
Gr. 1	*Classroom* (1 and 7 meters)	*Gallery* (1 and 7 meters)	*Divergent* condition stimuli (Gr. 1) equalized in loudness with *Control* condition
Gr. 2	*Reference* room (1 and 7 meters)	*Gallery* (1 and 14 meters)	*Divergent* condition stimuli (Gr. 2) equalized in loudness with *Control* condition

Table 2: Experiment IV: Measurements used to generate anchor stimuli for each group and each condition (*Control*, *Divergent*, *Intensity-equalized*). All target stimuli were generated with measurements of the *Classroom* at 9 distances ranging from 1 to 7 meters.

CONTROL CONDITION This condition is the same for both groups. Anchor stimuli were generated by convolution of the anechoic click recording with a BRIR measured at 1m and 7m in the *Classroom*. In that condition, both target and anchor stimuli are generated with measures of the same room.

DIVERGENT CONDITION In Group 1, anchor stimuli were generated by convolution of the anechoic click recording with a BRIR measured at 1m and 7m in the *Gallery*. In Group 2, anchor stimuli were generated by convolution of the anechoic click recording with a BRIR measured at 1m and 14m in the *Gallery*.

INTENSITY-EQUALIZED CONDITION For each group, modified anchor stimuli of the *Divergent* condition are generated. The global intensity of each anchor stimulus was adjusted so that their resulting loudness corresponds to the loudness of anchor stimuli of the *Classroom*, as used in the control condition (e.g. In Group 1 (resp. Group 2) , the loudness of the 7m (resp. 14m) *Gallery* anchor stimulus equals the loudness of the 7m *Classroom* stimulus).

9.2.2.3 *Participants*

Online participants were recruited via the recruitment platform *Prolific*(see Section 6.2). A total of 120 participants were recruited to complete the task. The same inclusion and exclusion criteria as the ones used in the previously described online experiments were used (see Section 6.2). Sixty participants were included in Group 1 (ages ranging from 50 to 18 years old, mean age = 24, SD = 6, 24 females, 36 males), and another 60 participants were included in Group 2 (ages ranging from 54 to 18 years old, mean age = 28, SD = 8, 26 females, 34 males).

9.2.3 *Procedure*

After reading the information note and approving the consent form, participants were directed to the experiment. Participants had to fill out a questionnaire in which they were invited to report their: age, gender, and the type of headphones they used. They were asked to use circumaural headphones if possible. A screening process identical to the one used in previous experiments was used.

Afterwards, a training session began. Participants were instructed that the click sounds were only here to help their sound distance judgement, and to judge their perceived distance to the male speaker using the slider. The training session was composed of three stimuli taken from the control block, with a target stimulus at 1, 2.5 and 7 meters, displayed in a random order.

After the training session, the experiment began. Participants were informed that the procedure is the same as for the training session. Each block contained 27 randomized sound sequences (3 repetitions $\times$ 9 different sound sequences). The order of the 3 blocks within the experiment was also randomized. During a trial, participants triggered the sound sequence playback, but the sequence was only played once. Participants had to report the perceived distance to the target stimuli with a visual analogue slider (See Section 6.1.3). An attention test was triggered between each block. Participants had to recognize which of two recordings corresponded to the spoken content of the target stimuli.

At the end of the experiment, participants were invited to fill out a final questionnaire to collect feedback on the experiment. They had to evaluate the perceived externalization and the general usefulness of the clicks to produce a judgment. The mean duration of the procedure was 20 minutes.

9.3 EXPERIMENT IV: RESULTS

Statistical analysis was performed with *TIBCO Statistica*© except for the power-function fittings, which were done using *Mathworks Inc. MATLAB*©.

OUTLIERS Ten participants in each group were initially excluded from the analysis: 8 of them because they experienced in-head localization, and 3 failed the attention tests, 5 completed the experiment too quickly ($<$ 5 minutes) and 4 others were excluded because a majority of their mean distance reports deviated of more than 2 standard deviations from the general mean calculated on all participants. Ten more participants were recruited in each group so the analysis could be conducted on 60 participants.

DATA SCALING Similarly to the other online experiments reported in this thesis, initial attention was focused on the scaling of the responses in order to create comparable ratings between participants. A min-max feature scaling was performed on the responses of participants using less than 95% of the total slider (see Equation 14).

9.3.1 *Effect of anchor condition*

In the following statistical analysis, the logarithmic value of the geometric mean of each participant's responses computed on the 3 reports associated with a single type of stimulus was considered as the dependent variable (27 different stimuli: 9 distances × 3 anchor conditions). For both groups separately, a focus was put on ensuring the normality of the dependent variables associated with a specific stimulus. A Jarque-Bera test indicated that in all cases for both groups, the null hypothesis "the data is normally distributed" was not rejected.

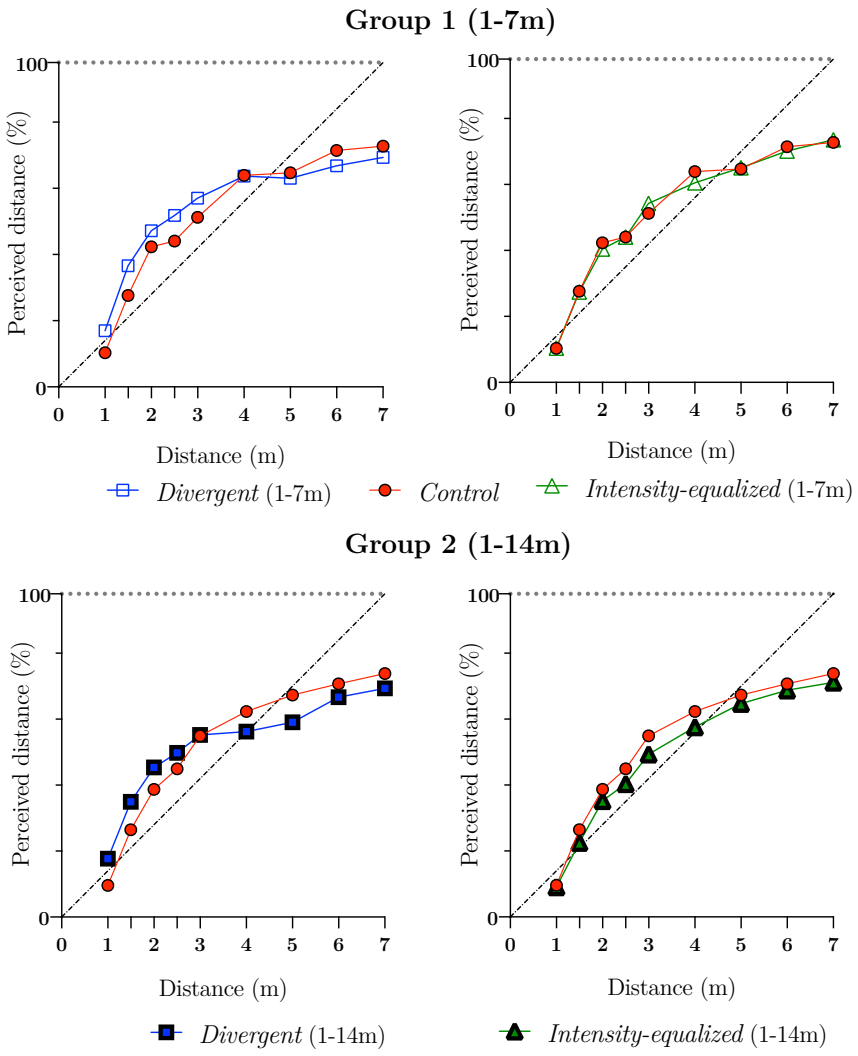


Figure 9.2: Geometric mean perceived distances per group (Group 1: Up, Group 2: down) and per anchor condition (Left: *Control* vs. *Divergent* Right: *Control* vs. *Intensity-equalized*).

In order to investigate possible significant differences between divergent conditions and the control condition, a total of 4 (2 per group) different repeated measure ANOVAs ($\alpha = 0.05$) were run. For each group:

- One ANOVA is conducted on the logarithmic value of the geometric mean of each participant (the dependent variable) collected in the *Control* and the *Divergent* conditions.
- One ANOVA is conducted with the dependent variable collected on the *Control* and the *Intensity-equalized* conditions.

All ANOVAs are conducted with the ANCHOR (2 levels) and DISTANCE (9 levels) as inter-subject factors.

		Anchor		Distance		Distance × Anchor	
		F(1,1)	p-value	F(1,8)	p-value	F(1,15)	p-value
Gr. 1	Divergent vs. Control	1.507	0.224	252.1	<0.001	2.241	0.023
	Equalized vs. Control	0.190	0.664	268.2	<0.001	1.060	**0.385**
Gr. 2	Divergent vs. Control	0.013	0.908	252.9	<0.001	3.049	0.002
	Equalized vs. Control	2.920	0.092	282.8	<0.001	1.274	**0.255**

Table 3: Statistical outputs of the 4 ANOVAs ran for both conditions, in each group. Each line corresponds to the results of a single ANOVA on the dependent variables present in the condition displayed and in the *Control* condition.

In both groups, the ANOVAs revealed:

- a strong significant effect of DISTANCE ($p < 0.001$).
- no significant effect of ANCHOR ($p > 0.05$).
- a significant cross-effect of DISTANCE × ANCHOR was found between the *Divergent* condition and the *Control* condition.
- no significant cross-effect of DISTANCE × ANCHOR was found between the *Intensity-equalized* condition and the *Control* condition.

9.3.2 *Compression effect quantification*

The estimation of the compression effect in distance reports associated with each group and each anchor condition was done with power functions (see Section 3.1). The mean values and standard deviations of the compression coefficients a and k are displayed in Figure 9.3. Independent samples t-tests have been conducted between different groups of parameters to illustrate the influence of the anchor conditions per group.

When compared to the values in the *Control* condition, the coefficients a and k are significantly smaller in the *Divergent* condition. This result implies that the distance reports are more compressed in the *Divergent* condition.

Identical tests were run between the *Control* condition and the *Intensity-equalized* condition. No significant differences were found. Thus, the compression effect in distance reports is comparable in both conditions.

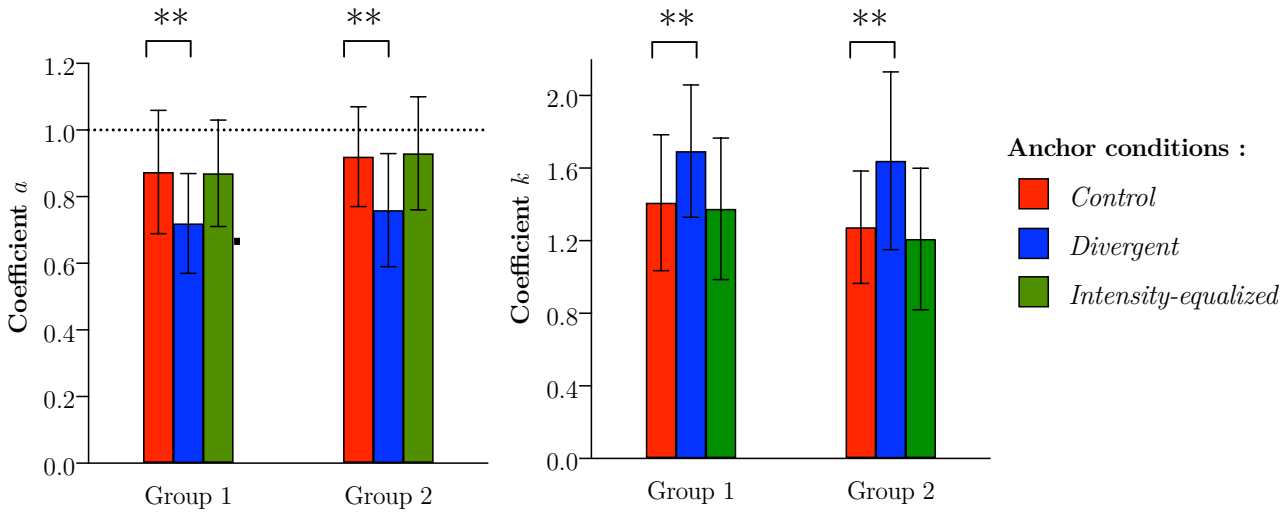


Figure 9.3: Mean and standard deviation of individual fitting coefficients a (left) and k (right) in each condition per group: *Control* (red), *Divergent* (blue), *Intensity-equalized* (green). (*) t-test indicated a p-value < 0.05 (**) t-test indicated a p-value < 0.01.

9.4 DISCUSSION

This experiment evaluated the impact of an acoustic divergence between the acoustic cues conveyed by a virtual sound source (the target stimulus) and co-occurrent sound sources (the anchor stimuli). The procedure was designed to reproduce the effect present in an AAR scenario in which the acoustic environment is not faithfully reproduced.

We intended to see whether, like in relative auditory distance perception scenarios [179], intensity is the main cue driving the calibration effect rather than reverberation-related cues.

9.4.1 *Effect of uncorrected room divergence effect on auditory distance perception*

The presence of divergent anchor stimuli had a significant effect on auditory distance perception. When compared to the *Control* condition in which no mismatch was present between anchor and target stimuli, the *Divergent* condition induced more compressed perceived distances. The *Divergent* condition represents a sce-

nario in which the real environment corresponds to a larger room with a stronger reverberation than the simulated room effect (*Gallery*: $T_{60}(1\text{kHz}) = 0.90\text{s}$, *Classroom*: $T_{60}(1\text{kHz}) = 0.55\text{s}$). The distance report range tends to expand in more reverberant environments [91]. Hence, a sound source at 7meters would be perceived further in the *Gallery* than in the *Classroom*. In group 2, this situation is more extreme, as the furthest stimulus corresponds to a measurement at 14 meters in the *Gallery*. This acoustic divergence induced a calibration of the participants' representation of the auditory space, which was based on the room effect of the *Gallery*. The interpretation of the acoustic cues conveyed by the target stimulus was altered. This is revealed, in each group, by the presence of a stronger compression in distance reports of the *Divergent* condition when compared to the *Control* condition (see Figure 9.3).

It can be argued that the effect of the acoustic divergence on auditory distance perception would have been exacerbated if both categories of stimuli were of the same nature. The anechoic sound used for anchor stimuli was a click, giving no a priori information on the initial power of the sound source to the listener. The use of speech for both categories of stimuli could have led to a stronger effect of anchor conditions as it would have induced a direct comparison of the acoustic cues conveyed by both categories of stimuli.

9.4.2 *Correcting the divergence with loudness matching*

We investigated which characteristics of the acoustic divergence drove the calibration effect. To avoid overlooking the potential impact of reverberation-related cues, and in line with the hypothesis that in weakly reverberant environments, participants are expected to rely weakly on reverberation-related cues, the anchor stimuli were generated in a more reverberant room.

Motivated by studies on relative auditory distance perception [3, 179], the role of intensity was primarily investigated by providing a loudness match with the *Control* anchor stimuli in the *Intensity-equalized* condition. The global intensity of the anchor stimuli was changed so their loudness matched the ones measured in the anchor stimuli of the *Control* condition. No significant differences were observed between the *Intensity-equalized* condition and the *Control* condition. The same results were observed in Group 2, where a stronger acoustic divergence was created by generating an anchor stimulus with a measurement at 14 meters in the *Gallery*.

These findings show that the anchor condition influenced auditory distance judgements primarily through the perception of intensity of the two anchor stimuli positioned at two extreme distances. Reverberation-related cue differences had no significant impact. As a result, only the intensity aspect of the acoustic divergence was responsible for the calibration effect.

9.5 COMPARISON WITH EXPERIMENT III

In order to facilitate comparison, the report method and the instructions given to participants were the same as the ones used in Experiment IV.

9.5.1 *Acoustic and visual divergence*

Gil-Carvajal et al. [64] stated that in binaural AAR scenarios, an emphasis should be put on the matching of the acoustic cues conveyed by real and virtual sound sources instead of visual congruence. In their study, participants were tested in divergent auditory, visual, or auditory-visual situations. However, distance judgements were made relatively to a real loudspeaker, so the incongruence and divergence were mainly driven by source-related cues (see Section 4.2.1).

We focused on the visual and acoustic aspects of the room divergence effect, which was introduced in this experiment and in Experiment III (see Chapter 8). We have attempted to compare the effect of the acoustic divergence produced here in the *Divergent* condition with the impact of the incongruent visual spatial boundary studied in Experiment III.

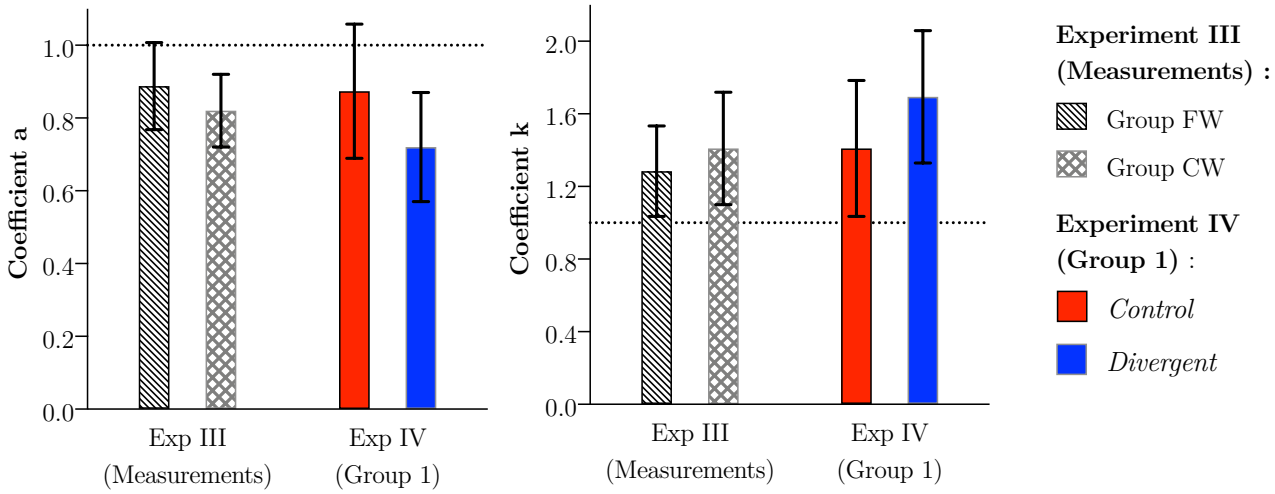


Figure 9.4: Mean and standard deviation of individual fitting coefficients a (left) and k (right) classed by experiments. Experiment III - Measurements-based method (grey) ; Experiment IV - Group 1 *Control* (red) and *Divergent* (blue) conditions.

The effects of the incongruence due to the visual spatial boundary in Experiment III and of the *Divergent* condition have a similar order of magnitude. The acoustic divergence has a slightly more important impact than the one induced by visual spatial boundary differences.

However, they are hardly comparable. The visual spatial boundary was tested in non-controlled environments with criteria only defined as a minimum or a max-

imum for the visual spatial boundary. Moreover, in Experiment III, the evaluation of the visual incongruence effect was done between groups of participants. In this experiment, the acoustic divergence between conditions was controlled. Its effect on auditory distance perception was explored as an intra-subject factor.

As both observed effects are of similar magnitudes, it is then hard to conclude whether the visual incongruence or the acoustic divergence is more important than the other. Lab-based experiments where the experimental conditions and the room divergence are controlled, are needed to achieve this comparison.

9.5.2 *Impact of anchor stimuli in the control condition*

No particular instructions were given in the current experiment concerning the visual conditions, so we hypothesize that participants mainly focused on their computer screens, limiting their visual spatial boundary. The comparison of compression coefficients a and k indicates that participants in both experiments, with limited visual spatial boundary, have comparable compression effects (see Figure 9.4). This result hints that the influence of anchor stimuli in the *Control* condition was hardly significant.

Explicitly informing participants about the position of the anchor stimuli could have resulted in better distance report accuracy. This could have been done by including in the instructions that each anchor position corresponds to the maximum or minimum of the slider. These instructions would have induced a greater weight on relative auditory distance judgments, which are considered more accurate than absolute auditory judgments [3]. However, this experiment focused on calibration aspects and did not aim to create a scenario in which the positions of real sound sources are known to the listener. Consequently, an intra-modal calibration effect is responsible for the observed differences between conditions.

9.6 CONCLUSION

We presented an experiment evaluating the impact of the room divergence effect arising from the presence of co-occurring sound sources. It indicates the presence of an intra-modal calibration driven by the co-occurring sound sources. This calibration was mainly based on the intensity of the anchor stimuli.

In AAR applications, special care should be taken with the reproduction of acoustic environments. The primary cues involved in frontal auditory distance perception are mono-aural reverberation-related cues and intensity. In this specific scenario, with co-occurrent sound sources, we showed that the focus must be put on the reproduction of the intensity decrease with distance. In that case, a faithful reproduction of this cue can prevent an unintended bias due to an intra-modal calibration of the auditory space.

This finding could be subject to further experiments. The role of reverberation-related cues in the intra-modal calibration might have been insignificant because

of the experimental conditions imposed by the online procedure. The potential importance of these cues could be explored in situations where the acoustic divergence, in terms of reverberation-related cues, is more important than in the experiment presented here. It could also be assessed in a scenario where the intensity is an unreliable distance cue. This situation could be achieved by suppressing the sound level variation with distance from the stimuli used as anchors and target stimuli.

10

GENERAL CONCLUSION & PERSPECTIVES

In this thesis, we explored the impact of several effects relevant to the design of auditory distance rendering methods in audio-only augmented reality. This thesis focused on four contributions concerning: 1) the design and application of experimental procedures, 2) the influence of reverberation-related cues on auditory distance perception and how they are relatively weighted with sound level, 3) the impact of environment-related cues linked to geometry, and 4) the influence of co-occurring and acoustically divergent sound sources. Each of these contributions offers perspectives for further research directions in spatial audio rendering for AAR.

10.1 EXPERIMENTAL PROCEDURES

All studies, with the exception of the lab-based Experiment I, were carried out online with remote participants using tools that have been presented in Chapter 6. To the best of our knowledge, no study has used or evaluated the reliability of online methods for auditory distance experiments yet. One of the goals of this thesis was to investigate the impact of environmental cues on auditory distance perception, which was achieved in Experiment II. The significance of two distinct effects related to volume and the visual spatial boundary was successfully underlined through an online-based approach, despite the uncontrolled aspects of the participants' environment.

We encountered the following pitfalls throughout the application of the online procedure:

- Each experiment had a high rate of outliers (about 10%). As reported in Section 6.2, participants' commitment and attentiveness are reduced when they are remote. This situation probably explains the high rate of outliers. The systematic use of attention tests, disseminated along the experimental session, which was only done in Experiment IV, would have helped their identification.

- The inherent lack of control over the participants' experimental conditions, such as the use of diverse headphones models and different environmental noise levels, may have increased the variability of the reported distances. This limitation, and the possible biases it engendered, could be prevented in an online experiment focusing only on within-subject factors. However, such a design would have been difficult to use in Experiment III (study of environment-related cues).

- Each experiment had to be designed with a limited number of stimuli per session due to the time constraint inherent to online experiments. Therefore, the number of within-subject variables was limited and only a few conditions could be examined (e.g. Experiments II and IV, Chapter 9). Online platforms (e.g. *Prolific*, *Gorilla*) now allow researchers to schedule several sessions with the same subjects. This could help overcome the time constraint on the duration of an experiment.

Overall, the design of the procedure could have been more rigorous. An initial assessment of the online procedure's influence may be drawn by comparing the results of lab-based Experiment I to those of online-based Experiment III, which used the same stimuli. The reliability of the online procedure can be examined through the intra and inter-subject variability of reported distances, which are most likely affected by the lack of control over the experimental conditions and the reduced attention inherent to online studies.

Apart from the online aspect, the lab-based procedure differs in the reporting method. In Experiment I, the maximum and minimum of the slider were explicitly associated with a physical position, the position of the listener and the boundary of the room. Moreover, a larger number of presentations per stimulus was used (9 instead of 4). Thus, intra-subject variability must be cautiously compared.

In both experiments, power functions were fitted to participants' data in order to quantify individual compression coefficients. In both experiments, the quality of fit of this function is comparable (for the reference measurements: $R^2 = 0.79$ in Experiment I, and $R^2 = 0.84$ in Experiment III). This suggests that despite a lower number of stimuli, the intra-subject variability is comparable in both experiments. This result hints that the supposed reduced attention of participants has not affected the quality of individual data.

The inter-subject variability in terms of compression effect can be compared through the standard deviation of the collected compression coefficients in each experiment. When comparing these values (reported in Table 4), the inter-subject variability is lower for online participants. This result is mainly driven by the higher number of participants in this experiment. It shows that the lack of control of experimental conditions, which induces increased inter-subject variability across participants, can be overcome by enlarging the pool of participants. One main advantage of online experiments is that they enable the recruitment of large samples of participants easily.

	Measurements		Envelope-based		Intensity-based	
	k	a	k	a	k	a
Experiment I (N=19)	0.5	0.24	0.49	0.33	1.22	0.55
Exp. III - CW Group (N=55)	0.31	0.14	0.25	0.15	0.33	0.24
Exp. III - FW Group (N=53)	0.31	0.12	0.37	0.13	0.33	0.25

Table 4: Standard deviations of compression coefficients a and k collected on participants for each rendering method in Experiment I and Experiment III.

These results corroborate recent research on the reliability of online procedures. The ASA P&P task force on remote testing [1] published their first analysis of studies identifying the impact of remote procedures (online-based approach included) for auditory experiments. A study based on within-subject comparisons of participants performing a similar task in the lab or remotely found that participants are generally truthful about their demographic characteristics, hearing impairments, and listening environment characteristics. Moreover, it also demonstrated that participants performing an auditory detection task in a lab and at home with their own material [137] (headphones, computer) showed similar results. This specific finding suggests that the increased variability, caused by the lack of control over the experimental conditions in online-based studies, has a limited effect on the quality of the collected data.

In our case, a within-subject comparison on the evaluation of visual cues (visual spatial boundary and room volume) with participants in the lab would be necessary to assess the results of Experiment II. The reliability of the characteristics of the room reported by participants can be questioned. Moreover, these reports do not entirely allow us to determine if the observed effect is entirely due to a single visual cue or can be attributed to covarying room characteristics. There, it can be argued that online-based experiments can be subject to stronger variability, when involving thoroughly the user's envrionment. Since experiments in a lab allow to control visual conditions precisely, it can be expected that lab-based results would highlight a more reliable effect of visual cues.

10.2 THE PERCEPTION OF EARLY ENERGY RELATIVELY TO REVERBERATION FOR DISTANCE

The thesis included an investigation into the relevance of the early-to-late energy ratio as a cue for auditory distance perception. The most frequently acknowledged reverberation-related cue is the DRR, whose perceptual relevance has been questioned in previous research.

The first two experiments introduced in Chapter 7, examined if an early-to-late energy ratio could be perceptually more meaningful. In the specific context of the

1 https://www.spatialhearing.org/remotetesting/

lab-based Experiment I, the envelope-based model was as efficient as real-world measurements in rendering sound source distance. Following that, we examined if this perceptual similarity could be explained entirely by the correct reproduction of an early-to-late energy ratio. The *Forward* and *Backward* synthesis methods introduced in Experiment II, were designed in order to faithfully replicate the energy of the early segment of the corresponding measurements, while deliberately introducing strong modifications of their internal reflections distribution, illustrated by significant differences in the D/Ref ratio. The experiment showed that when considering a transition time of 40ms, the distance judgements are close to the measurements and nearly independent from the internal distribution of reflections, as long as the D/Ref ratio does not differ by more than 9dB. In contrast, when considering a larger transition time (i.e. 80ms), the distance judgments are strongly dependent on the internal distribution of reflections and differ significantly from the distances reported with the measurements. As stated in the previous section, online experiments did not allow us to test extensively the effect of the transition time, i.e. testing more transition time values. Further studies should be dedicated to determine the optimal transition time.

The findings of this second experiment indicated that if it is essential to reproduce an early-to-late energy ratio, spectral cues cannot be ignored and their perceptual weight is not negligible for some listeners. Indeed, the *Forward* and *Backward* synthesis methods introduced not only differences in the Direct over First Reflections ratio, but also created significant modifications in the spectral balance and spatial cues. Similarly, spatial characteristics associated with early reflections that have been believed to have a limited effect on auditory distance perception should be examined. It has been argued that the lack of lateralization of early reflections could strengthen their perceptual fusion to the direct sound [26]. As a result, spatial characteristics may influence how early and late energies are perceived, and hence how auditory distance perception is inferred from reverberation. The use of functions specific to SRIRs manipulations, such as the "spatial warping" [96], enables modifications of their spatial characteristics. These functions could be useful in order to study the importance of reverberation-related spatial characteristics (e.g. early reflections direction of arrival) on auditory distance perception.

The impact of reverberation-related spectral and spatial cues on auditory distance perception should be investigated independently in order to quantify their significance relatively to energetic cues (DRR, early-to-late energy ratio). The design of artificial room impulse responses enables the complete control of the energetic and spectral aspects independently, and would therefore be better suited to the evaluation of these cues.

10.3 ACOUSTIC CUES WEIGHTING STRATEGIES AND THE INFLUENCE OF ROOM VOLUME

Auditory distance perception is based primarily on sound intensity and reverberation-related cues. The perceptual system uses weighting strategies on these cues to infer

a distance percept. The weight associated with each cue is known to vary from one individual to another, and depends on the listening situation's characteristics. One of the goals of the thesis was to examine if the participants' strategies to infer a distance judgement were mainly idiosynchratic or could be linked to characteristics of the environment.

Experiments I and III reported respectively in Chapter 7 and Chapter 8, provided an insight into participants' acoustic cues weighting strategies. The use of a model that only included intensity as a relevant distance cue underlined the idiosyncratic nature of these strategies. Despite the fact that all participants were exposed to the same auditory stimuli, a wide diversity of strategies were used by participants. Some participants judged auditory distance primarily on the basis of intensity while others relied entirely on reverberation-related cues. In order to investigate the potential influence of the listening environment on the used strategies, identical stimuli were evaluated with remote participants in Experiment III. Participants were asked to report the dimensions of the room in which they were doing the experiment. The volume of the room was shown to have a small but significant effect on how sound level is weighted in relation to reverberation. A larger room volume is linked to a greater reliance on reverberation.

Experiment II suggested the presence of weighting strategies on reverberation-related cues. Some participants appeared to be impacted by spectral differences in their distance judgements, while others showed comparable performances despite these differences. Thus, some participants based their distance judgements on reverberation-related spectral cues (and/or potentially on spatial cues) while others only used energetic cues such as the early-to-late energy ratio.

Future studies can be envisioned in order to study acoustic cues weighting strategies. Before investigating the impact of individual or environmental characteristics, the relationship between these strategies and externalization should be assessed. In all experiments, binaural stimuli were generated using generic HRTFs, which increases the chances of in-head localisation. Moreover, after each stimulus presentation, the perceived externalization was not assessed by participants. The lack of externalization is known to disable auditory distance perception for some participants [141]. Therefore, changes in the compression effect across rendering methods might be partly related to the externalization induced by them and not to the acoustic cues weighting strategy adopted.

The influence of environmental characteristics on acoustic cues weighting strategies could be evaluated via lab-based experiments in controlled situations. In contrast to an online-based method, it would allow for the examination of the influence of these characteristics as within-subject factors. It has been suggested that the weight associated to each cue may rely on its consistency [183]. Cues that are unreliable (e.g. reverberation-related cues in an anechoic environment) are given less perceptual weight in the combination process. Therefore, the relationship between the expected reliability of acoustic cues due to the perception of the environment, and their perceptual weights, should be examined. Apart from the volume,

the perception of environmental characteristics (such as the nature of the walls material) could drive the expectation of a stronger reverberation.

Finally, the effect of the stimulus type on the strategies could be explored. In all experiments, speech stimuli distance had to be judged. This choice was made because speech is assumed to be equally familiar to all individuals. Because individuals are familiar with the power of speech's sound source, intensity is a more reliable cue in this situation. It would be interesting to determine the extent to which the stimulus's nature has an effect in comparison to the environmental and idiosyncratic characteristics. The impact of the nature of the stimuli on the optimal transition time value could also be investigated.

10.4 VISUAL INCONGRUENCE AND ACOUSTIC DIVERGENCE

This thesis investigated possible effects that could occur when the listening environment is not accurately characterized and/or faithfully reproduced. Experiment III demonstrated how the vision of incongruent spatial boundaries can calibrate auditory distance perception. In auditory-visual incongruent situations, close visual spatial boundaries induce a larger compressive effect on auditory distance perception. This result shows that an incongruence between the perceived visual geometry of the real environment and the reproduced room effect can impair auditory distance perception. This observed effect was independent from the volume reported by participants. Thus, it could be argued that, according to the position of the listener inside the same environment, auditory distance perception varies based on the visual perception of boundaries.

Experiment IV investigated an intra-modal calibration effect that could occur in an AAR scenario in which an acoustic divergence between room effects is present. The findings indicate that it could significantly influence the auditory distance perception of virtual sound sources. In this experiment, we observed a modification of participants' compression effects in acoustically divergent conditions. In these conditions, the effect of the acoustic divergence was mainly driven by the intensity variations with distance. Although these findings require additional perceptual evaluations, they suggest that acoustic divergence becomes an issue for auditory distance perception mainly when the intensity cue is not correctly reproduced.

Apart from assessing these different effects in controlled environments, possible perspectives can be foreseen. Firstly, the impact of these calibration effects could be assessed in the presence of real audio-visual sound sources. It can be hypothesized that the presence of congruent auditory and visual information about a real sound source increases the impact of an acoustic divergence. Gil-Carvajal et al. [64] conducted a study in which they examined the influence of visual-only, acoustic-only, and acoustic-visual divergence in the presence of a real sound source (loudspeaker). The location of the source was disclosed to the participants, and they had to judge the distance to a virtual sound source reproduced with a divergent room

effect. Participants exhibited a larger influence of the room divergence on auditory distance perception when they could hear the real sound source (acoustic-only and acoustic-visual divergence). The impact of the divergence was comparable in both conditions, indicating that participants based their decisions on a direct comparison of acoustic cues. It might be claimed that in an AAR scenario, visual distance perception could serve as a reliable estimate of the distance to a real sound source. As a result, the sound of the real sound source is directly paired with a visual object, establishing an "anchor" for the representation of the auditory space. In this situation, it can be expected that auditory distance perception of virtual sound sources would be made relatively to the position of the "anchor", inducing a direct comparison of the auditory cues conveyed by the actual and reproduced room effect.

Secondly, these calibration effects could be assessed in a scenario where motion tracking is enabled. Most of the studies investigating the role of sensori-motor integration in the calibration of auditory space have mainly focused on angular perception, while its effect on auditory distance perception is relatively unexplored. For individuals with no visual impairments, calibration of the auditory space seems to be primarily achieved by vision [90]. Calibration of the auditory space through sensorimotor integration only, has been proven to be particularly effective for blind individuals [66]. In an AAR scenario for normally-sighted individuals, both self-motion and vision are available. Their integration induces a constant updating of the visual and auditory cues. They provide a greater amount of information on the spatial characteristics of the environment than in static situations. Therefore, self-motion could increase the impact of a potential room divergence effect on spatial localization.

Finally, it could be determined whether listeners could learn or remap their auditory space as a result of a room divergence effect. In the case of azimuth and elevation perception, it is possible to induce an adaptation to altered auditory cues (e.g. through non-individualized HRTF [182] or ear molds [35]). Three main methods have been introduced to induce adaptation to altered auditory cues for angular perception: sound exposure, training with feedback, and explicit training [116].

Adaptation through sound exposure consists of participants learning implicitly the correspondence between altered auditory cues and source position. It is achieved with a continuous multisensory update as a result of motor exploration. Sound exposure must be sustained for an extended length of time in order to be effective with many days of continuous exposure [35].

Contrarily, training with feedback or explicit training consists of short sessions at intervals of one or several days. Adaptation through training is considered more efficient than sound exposure, as it necessitates a shorter amount of time [116]. Various feedback paradigms can be used to design a training task. A training task usually consists of sound localization reporting trials, followed by feedback classifying the response as right or wrong (explicit) or specifying the true location of the stimulus. This positional feedback can be provided with visual, motor, and/or

auditory markers. For example, Zahorik et al. [182] trained subjects on sound localization tasks with audio-visual feedback displayed on a motion-tracked head-mounted display. They notably demonstrated that the benefits of the training were still effective 4 months later with the same participants. Shinn-Cunningham et al. [152] used light flashes to indicate the correct direction to a virtual sound source after each spatialized sound stimulus presentation. Stitt et al. [158] designed a game task in which participants had to search for virtual sound sources by pointing with their hand, on which a motion-tracked device was placed. An auditory feedback designed as a "Geiger counter" metaphor was used. It consisted of pink and white noise bursts, whose alternation rate got smaller when participants were getting closer to the right direction.

It could be interesting to investigate if adaptation to altered auditory distance cues can also be achieved by providing a similar training. Explicit training tasks could be designed using visual markers in the real world or in VR. Outside the field of view, a training task adapted from the one used by Stitt et al. could be applied. If participants are effectively able to adapt to distance rendering laws that deviate from the rules of the physical world, it could bring into question the standard to be reached, in terms of accuracy in the reproduction of acoustic cues, by perceptually-motivated auditory distance rendering methods in AAR applications. Mendoncca et al. [116] have proposed that most likely, humans are able to represent several auditory cues combination rules at once to infer a localization judgement. This would allow AAR users to adapt to non-physical laws without experiencing real world object localization disruption. However, it presumes that users can dissociate the virtual or real nature of the perceived sound objects. Consequently, the question of the aftereffects linked to the adaptation to divergent acoustic cues must be raised. Auditory distance perception is critical for survival and altering the perception of the physical environment could have dangerous consequences. Therefore, using non-realistic auditory cues to simulate distance in AAR requires careful consideration of the ethical implications associated with the integration of virtuality into our daily lives reality.

Part IV

APPENDIX

a

APPENDIX, PRELIMINARY EXPERIMENT

This appendix presents a preliminary experiment investigating the perceptual similarity between BRIRs converted from SRIRs (see Chapter 5) and actual measurements of BRIRs. The procedure is the same as the one used in Experiment I (see Chapter 7). The procedure only differs in terms of participants and auditory stimuli.

A.1 METHODS

The procedure used for this experiment is the same as the one presented in Experiment I. Only the differences with the procedure presented in Chapter 7 will be presented.

A.1.1 *Auditory stimuli*

Auditory stimuli are based on the convolution of BRIRs with a speech recording. Three categories of measurements were evaluated. All measurements were made in the *Classroom* at IRCAM, used in all experiments (see Figure 7.2).

Different BRIRs were measured in a classroom at IRCAM with a dummy head *Neumann KU100*. Nine BRIRs were measured for distances ranging from 1 to 7m (1, 1.5, 2, 2.5, 3, 4, 5, 6, 7m) by changing the speaker position and with a single microphone position.

For the same source-receiver positions SRIRs were measured with a spherical microphone array *Eigenmike©EM32*, and converted to BRIRs following the procedure described in Chapter 5.

Finally, BRIRs were generated with the envelope-based model used in Experiment I for the same source-receiver distances.

A.1.2 *Participants*

A total of 20 (8 women) participants, ages ranging from 22 to 52 (mean age: 29), took part in the study. Inclusion and exclusion criteria reported in Chapter 6 were applied. Participants were recruited among Sorbonne-Université students and people working at IRCAM.

A.1.3 *Procedure, listening environment & report method*

The procedure, listening environment, and report method were identical to the ones presented in Section 7.3 for Experiment I.

A.2 RESULTS

A repeated measures ANOVA applied to the geometric mean distances of each participant was carried out, with the within-subject factors DISTANCE (9 levels from 1 to 7m) and MODEL (3 levels: 2 models and the reference). The values of the geometric means over all participants are displayed in Figure a.1.

- The main effect DISTANCE was significant ($F(1,8) = 509,4$ $p < 0.01$, $Partial - \eta^2 = 0,9802$)
- The main effect MODEL was not significant($F(1,2) = 0,73$, $p = 0.25$, $Partial - \eta^2 = 0.0488$)
- and the interaction DISTANCE $\times$ MODEL was not significant ($F(1,15) = 0,981$, $p = 0.11$, $Partial - \eta^2 = 0.0756$).

The similarity in distance reports across the methods was further investigated using a post-hoc analysis (Fisher LSD). For each distance, no significant differences ($p > 0.05$) between the three different methods were found.

A.3 CONCLUSION

The results of this preliminary experiment illustrate the similarity between actual BRIRs and those converted from SRIRs. If a perceptual difference exists, our report procedure indicates that it is not significant.

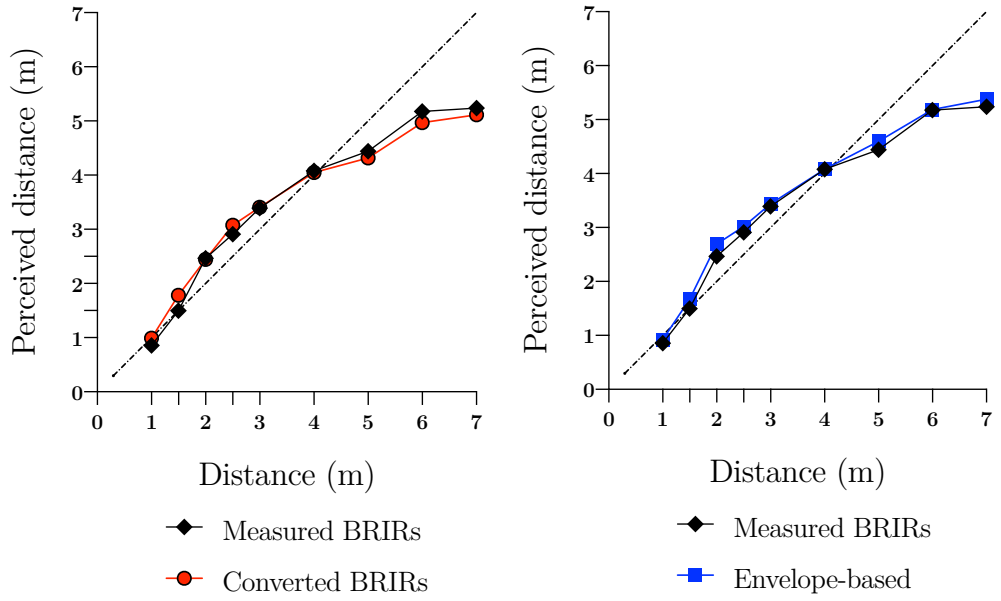


Figure a.1: Geometric mean (over 20 participants) perceived distances according to the method used to generate the sound source: Actual BRIRs(black), BRIRs converted from measured SRIRs(red), and BRIRs generated by the envelope-based model (blue).

b

APPENDIX, CHAPTER 7 (EXPERIMENT II)

This appendix provides an illustration of the energetic differences between the *Forward* and *Backward* synthesized BRIRs with actual measurements. The differences in terms of early energy of the synthetic responses with corresponding real measurements can be found in Figure b.1 and Figure b.2. The figures depict the early energy differences with measurements as a function of the offset time t of the BRIRs early part:

$$E_{dif}(t, d, T_{trans}) = E_{early,syn}(t, d, T_{trans}) - E_{early,meas}(t, d) \qquad \textit{(in dB)} \quad (16)$$

With $E_{early,syn}(t, d, T_{trans})$ the energy contained in the early part $[0, t]$ of a synthesized impulse response for a distance d and an offset time t.

And $E_{early,meas}(t, d)$ the energy contained in the early part $[0, t]$ of the measured impulse response at a distance d and a considered offset time t.

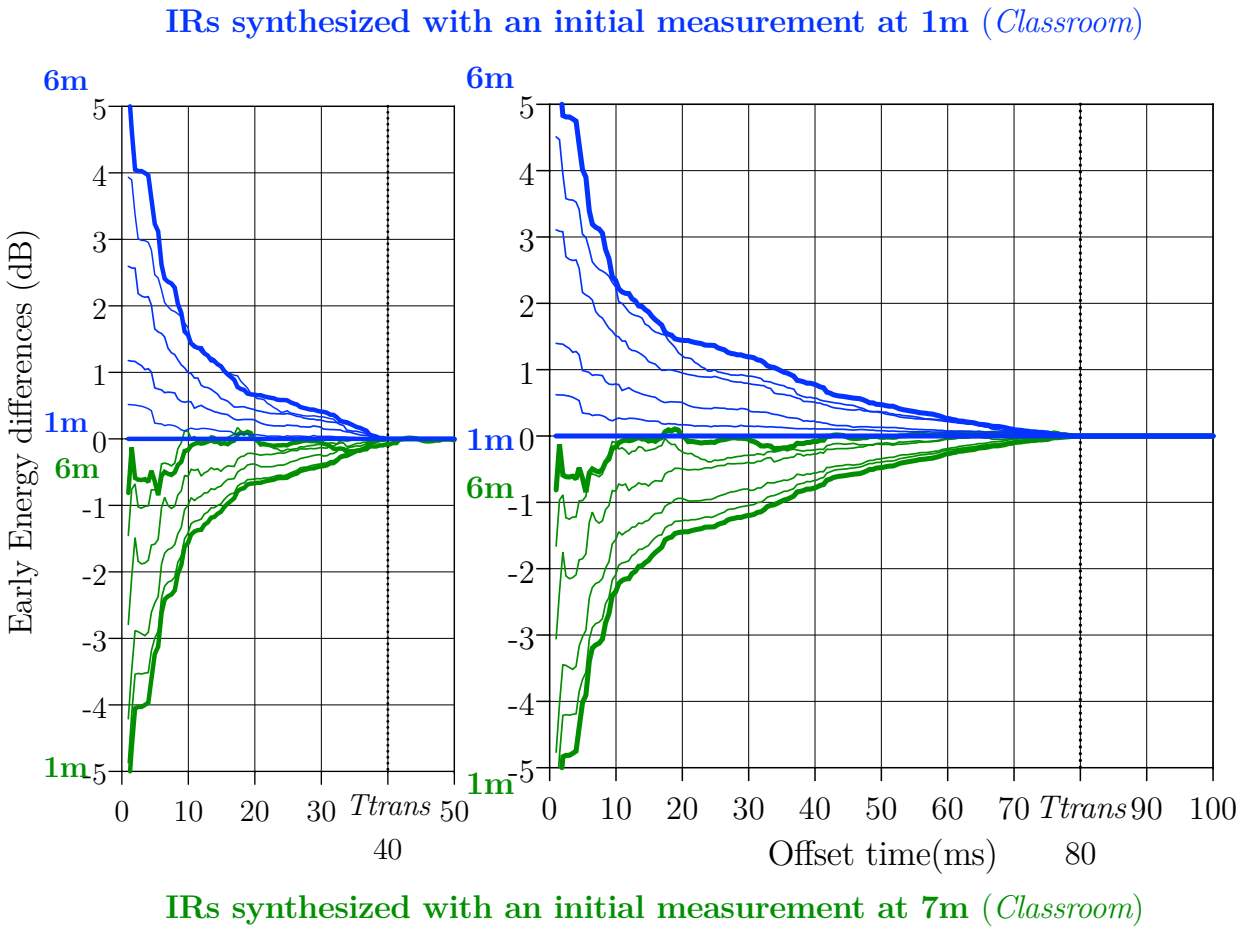


Figure b.1: Early energy differences as a function of its considered offset time, between synthesized and measured BRIRs of the *Classroom*, for T_{trans} equal to 40ms(left) and 80ms (right). An initial impulse response at 1m (blue, up) and 7m(green, down) were used as the initial measurement of the syntheses to create respectively an excess or a lack of energy in the early part ($t < T_{trans}$ when compared to real measurements.

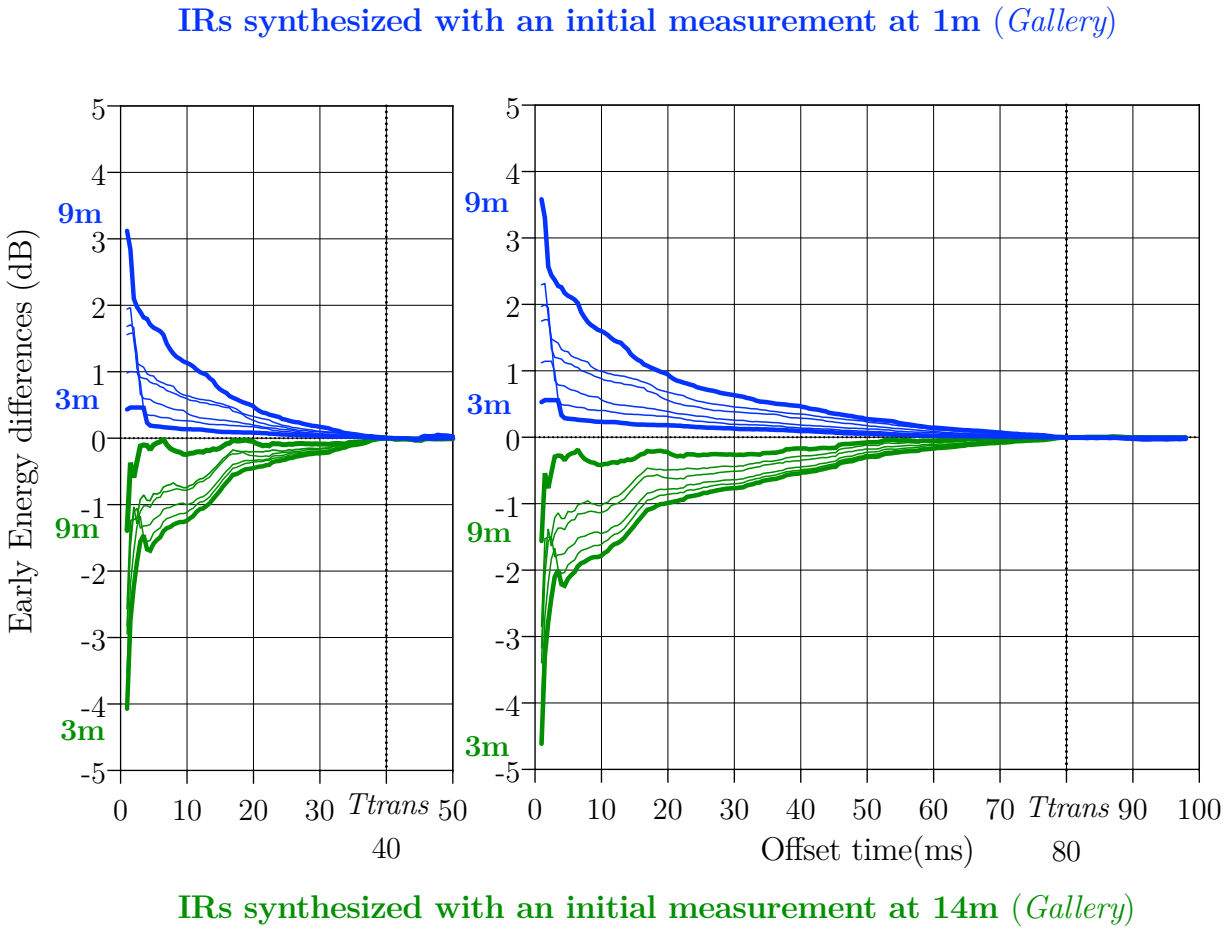


Figure b.2: Early energy differences as a function of its considered offset time, between synthesized and measured BRIRs of the *Gallery*, for T_{trans} equal to 40ms(left) and 80ms (right). An initial impulse response at 1m (blue, up) and 7m (green, down) were used as the initial measurement of the syntheses to create respectively an excess or a lack of energy in the early part ($t < T_{trans}$) when compared to real measurements.

C

PUBLICATIONS

[1] Vincent Martin, Isabelle Viaud-Delmon, and Olivier Warusfel. "Virtual sound source distance evaluation in acoustically and visually incongruent contexts." In: *The Journal of the Acoustical Society of America* 150.4 (2021), A140–A141.

[2] Vincent Martin, Isabelle Viaud-Delmon, and Olivier Warusfel. "Effect of Environment-Related Cues on Auditory Distance Perception in the Context of Audio-Only Augmented Reality." In: *Applied Sciences* 12.1 (2022). ISSN: 2076-3417. DOI: 10.3390/app12010348. URL: https://www.mdpi.com/2076-3417/12/1/348.

[3] Vincent Martin, Olivier Warusfel, and Isabelle Viaud-Delmon. "Source distance modelling in the context of Audio Augmented Reality." In: *Forum Acusticum*. 2020, pp. 1369–1376.

BIBLIOGRAPHY

[1] Jonathan S Abel and Patty Huang. "A simple, robust measure of reverberation echo density." In: *Audio Engineering Society Convention 121*. Audio Engineering Society. 2006.

[2] Thibaut Ajdler, Luciano Sbaiz, and Martin Vetterli. "The plenacoustic function and its sampling." In: *IEEE transactions on Signal Processing* 54.10 (2006), pp. 3790–3804.

[3] Michael A Akeroyd, Stuart Gatehouse, and Julia Blaschke. "The detection of differences in the cues to distance by elderly hearing-impaired listeners." In: *The Journal of the Acoustical Society of America* 121.2 (2007), pp. 1077–1089.

[4] David Alais and David Burr. "The ventriloquist effect results from near-optimal bimodal integration." In: *Current biology* 14.3 (2004), pp. 257–262.

[5] Ralph Algazi, Richard O Duda, Ramani Duraiswami, Nail A Gumerov, and Zhihui Tang. "Approximating the head-related transfer function using simple geometric models of the head and torso." In: *The Journal of the Acoustical Society of America* 112.5 (2002), pp. 2053–2064.

[6] Jont B Allen and David A Berkley. "Image method for efficiently simulating small-room acoustics." In: *The Journal of the Acoustical Society of America* 65.4 (1979), pp. 943–950.

[7] Paul W Anderson and Pavel Zahorik. "Auditory/visual distance estimation: accuracy and variability." In: *Frontiers in psychology* 5 (2014), p. 1097.

[8] Jeffrey Andre and Sheena Rogers. "Using verbal and blind-walking distance estimates to investigate the two visual systems hypothesis." In: *Perception & Psychophysics* 68.3 (2006), pp. 353–361.

[9] Alexander L Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K Evershed. "Gorilla in our midst: An online behavioral experiment builder." In: *Behavior research methods* 52.1 (2020), pp. 388–407.

[10] Nobuharu Aoshima. "Computer-generated pulse signal applied for sound measurement." In: *The Journal of the Acoustical Society of America* 69.5 (1981), pp. 1484–1488.

[11] Daniel H Ashmead, DeFord L Davis, and Anna Northington. "Contribution of listeners' approaching motion to auditory distance perception." In: *Journal of experimental psychology: Human perception and performance* 21.2 (1995), p. 239.

[12] Daniel H Ashmead, Deford Leroy, and Richard D Odom. "Perception of the relative distances of nearby sound sources." In: *Perception & psychophysics* 47.4 (1990), pp. 326–331.

[13] Ronald T Azuma. "A survey of augmented reality." In: *Presence: teleoperators & virtual environments* 6.4 (1997), pp. 355–385.

[14] Amit Barde, William S Helton, Gun Lee, and Mark Billinghurst. "Binaural spatialization over a bone conduction headset: Minimum discernable angular difference." In: *Audio Engineering Society Convention 140*. Audio Engineering Society. 2016.

[15] Amit Barde, Robert W Lindeman, Gun Lee, and Mark Billinghurst. "Binaural Spatialization over a Bone Conduction Headset: The Perception of Elevation." In: *Audio Engineering Society Conference: 2019 AES International Conference on Headphones Technology*. Audio Engineering Society. 2019.

[16] Andrew Barron, Jorma Rissanen, and Bin Yu. "The minimum description length principle in coding and modeling." In: *IEEE Transactions on information theory* 44.6 (1998), pp. 2743–2760.

[17] Michael Barron and Arthur Harold Marshall. "Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure." In: *Journal of sound and Vibration* 77.2 (1981), pp. 211–232.

[18] Dwight W Batteau. "The role of the pinna in human localization." In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 168.1011 (1967), pp. 158–180.

[19] Benjamin B Bederson. "Audio augmented reality: a prototype automated tour guide." In: *Conference companion on Human factors in computing systems*. 1995, pp. 210–211.

[20] Durand R Begault. "Perceptual effects of synthetic reverberation on three-dimensional audio systems." In: *Journal of the Audio Engineering Society* 40.11 (1992), pp. 895–904.

[21] Virginia Best, Robert Baumgartner, Mathieu Lavandier, Piotr Majdak, and Norbert Kopčo. "Sound externalization: A review of recent research." In: *Trends in Hearing* 24 (2020).

[22] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[23] Costas Boletsis and Dimitra Chasanidou. "Smart tourism in cities: Exploring urban destinations with audio augmented reality." In: *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. 2018, pp. 515–521.

[24] John S Bradley, R Reich, and SG Norcross. "A just noticeable difference in C50 for speech." In: *Applied Acoustics* 58.2 (1999), pp. 99–108.

[25] John S Bradley and Gilbert A Soulodre. "Objective measures of listener envelopment." In: *The Journal of the Acoustical Society of America* 98.5 (1995), pp. 2590–2597.

[26] Adelbert W Bronkhorst. "Modeling auditory distance perception in rooms." In: *Forum Acusticum, Sevilla, Spain*. 2002.

[27] Adelbert W Bronkhorst and Tammo Houtgast. "Auditory distance perception in rooms." In: *Nature* 397.6719 (1999), pp. 517–520.

[28] Douglas S Brungart and William M Rabinowitz. "Auditory localization of nearby sources. Head-related transfer functions." In: *The Journal of the Acoustical Society of America* 106.3 (1999), pp. 1465–1479.

[29] Douglas S Brungart, William M Rabinowitz, and Nathaniel I Durlach. "Evaluation of response methods for the localization of nearby objects." In: *Perception & psychophysics* 62.1 (2000), pp. 48–65.

[30] Douglas S Brungart and Kimberly R Scott. "The effects of production and presentation level on the auditory distance perception of speech." In: *The Journal of the Acoustical Society of America* 110.1 (2001), pp. 425–440.

[31] Robert A Butler, Elena T Levy, and William D Neff. "Apparent distance of sounds recorded in echoic and anechoic chambers." In: *Journal of Experimental Psychology: Human Perception and Performance* 6.4 (1980), p. 745.

[32] Densil Cabrera, Caeup Jeong, Hyun Jeong Kwak, and Ji-Young Kim. "Auditory room size perception for modeled and measured rooms." In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 2005. 5. Citeseer. 2005, pp. 2995–3004.

[33] Esteban R Calcagno, Ezequiel L Abregu, Manuel C Eguía, and Ramiro Vergara. "The role of vision in auditory distance perception." In: *Perception* 41.2 (2012), pp. 175–192.

[34] Céline Cappe, Antonia Thelen, Vincenzo Romei, Gregor Thut, and Micah M Murray. "Looming signals reveal synergistic principles of multisensory integration." In: *Journal of Neuroscience* 32.4 (2012), pp. 1171–1182.

[35] Simon Carlile, Kapilesh Balachandar, and Heather Kelly. "Accommodating to new ears: the effects of sensory and sensory-motor feedback." In: *The Journal of the Acoustical Society of America* 135.4 (2014), pp. 2002–2011.

[36] Simon Carlile, Philip Leong, and Stephanie Hyams. "The nature and distribution of errors in sound localization by human listeners." In: *Hearing research* 114.1-2 (1997), pp. 179–196.

[37] Thibaut Carpentier, Markus Noisternig, and Olivier Warusfel. "Twenty years of Ircam Spat: looking back, looking forward." In: *41st International Computer Music Conference (ICMC)*. 2015, pp. 270–277.

[38] Thomas Chatzidimitris, Damianos Gavalas, and Despina Michael. "SoundPacman: Audio augmented reality in location-based games." In: *2016 18th Mediterranean Electrotechnical Conference (MELECON)*. IEEE. 2016, pp. 1–6.

[39] Laurence Cliffe, James Mansell, Joanne Cormac, Chris Greenhalgh, and Adrian Hazzard. "The audible artefact: Promoting cultural exploration and engagement with audio augmented reality." In: *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*. 2019, pp. 176–182.

[40] Paul D Coleman. "Failure to localize the source distance of an unfamiliar sound." In: *The Journal of the Acoustical Society of America* 34.3 (1962), pp. 345–346.

[41] Paul D Coleman. "An analysis of cues to auditory depth perception in free space." In: *Psychological Bulletin* 60.3 (1963), p. 302.

[42] Paul D Coleman. "Dual role of frequency spectrum in determination of auditory distance." In: *The Journal of the Acoustical Society of America* 44.2 (1968), pp. 631–632.

[43] Michael Cook. "The judgment of distance on a plane surface." In: *Perception & Psychophysics* 23.1 (1978), pp. 85–90.

[44] Trevor J Cox, William J Davies, and Yiu W Lam. "The sensitivity of listeners to early sound field changes in auditoria." In: *Acta Acustica united with Acustica* 79.1 (1993), pp. 27–41.

[45] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research." In: *PloS one* 8.3 (2013), e57410.

[46] James E Cutting and Peter M Vishton. "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth." In: *Perception of space and motion*. Elsevier, 1995, pp. 69–117.

[47] José Aparecido Da Silva. "Scales for perceived egocentric distance in a large open field: Comparison of three psychophysical methods." In: *The American Journal of Psychology* (1985), pp. 119–144.

[48] Jérôme Daniel and Sébastien Moreau. "Further study of sound field coding with higher order ambisonics." In: *Audio Engineering Society Convention 116*. Audio Engineering Society. 2004.

[49] Giovanni Del Galdo, Maja Taseska, Oliver Thiergart, Jukka Ahonen, and Ville Pulkki. "The diffuse sound field in energetic analysis." In: *The Journal of the Acoustical Society of America* 131.3 (2012), pp. 2141–2151.

[50] Richard O Duda and William L Martens. "Range dependence of the response of a spherical head model." In: *The Journal of the Acoustical Society of America* 104.5 (1998), pp. 3048–3058.

[51] Chris Dunn and Malcolm J Hawksford. "Distortion immunity of MLS-derived impulse response measurements." In: *Journal of the Audio Engineering Society* 41.5 (1993), pp. 314–335.

[52] Nathaniel I Durlach, A Rigopulos, XD Pang, WS Woods, A Kulkarni, H Steven Colburn, and Elizabeth M Wenzel. "On the externalization of auditory images." In: *Presence: Teleoperators & Virtual Environments* 1.2 (1992), pp. 251–257.

[53] Inger Ekman. "Sound-based gaming for sighted audiences–experiences from a mobile multiplayer location aware game." In: *Proceedings of the 2nd audio mostly conference*. 2007, pp. 148–153.

[54] Isaac Engel and Lorenzo Picinali. "Long-term user adaptation to an audio augmented reality system." In: *24th international congress on Sound and vibration, London*. 2017.

[55] Nicolas Epain and Craig T Jin. "Spherical harmonic signal covariance and sound field diffuseness." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10 (2016), pp. 1796–1807.

[56] Pablo E Etchemendy, Ezequiel Abregú, Esteban R Calcagno, Manuel C Eguia, Nilda Vechiatti, Federico Iasi, and Ramiro O Vergara. "Auditory environmental context affects visual distance perception." In: *Scientific reports* 7.1 (2017), pp. 1–10.

[57] Pablo E Etchemendy, Ignacio Spiousas, Esteban R Calcagno, Ezequiel Abregú, Manuel C Eguia, and Ramiro O Vergara. "Direct-location versus verbal report methods for measuring auditory distance perception in the far field." In: *Behavior research methods* 50.3 (2018), pp. 1234–1247.

[58] Angelo Farina. "Simultaneous measurement of impulse response and distortion with a swept-sine technique." In: *Audio Engineering Society Convention 108*. Audio Engineering Society. 2000.

[59] Eleanor A Gamble. "Minor studies from the psychological laboratory of Wellesley College: Intensity as a criterion in estimating the distance of sounds." In: *Psychological Review* 16.6 (1909), p. 416.

[60] Hannes Gamper et al. "Enabling technologies for audio augmented reality systems." In: (2014).

[61] Mark B Gardner. "Historical background of the Haas and/or precedence effect." In: *The Journal of the Acoustical Society of America* 43.6 (1968), pp. 1243–1248.

[62] Mark B Gardner. "Distance estimation of 0 or apparent 0-oriented speech signals in anechoic space." In: *The Journal of the Acoustical Society of America* 45.1 (1969), pp. 47–53.

[63] Michael A Gerzon. "Periphony: With-height sound reproduction." In: *Journal of the audio engineering society* 21.1 (1973), pp. 2–10.

[64] Juan C Gil-Carvajal, Jens Cubick, Sébastien Santurette, and Torsten Dau. "Spatial hearing with incongruent visual or auditory room cues." In: *Scientific reports* 6.1 (2016), pp. 1–10.

[65] Robert L Goldstone. "Influences of categorization on perceptual discrimination." In: *Journal of Experimental Psychology: General* 123.2 (1994), p. 178.

[66] Monica Gori, Giulio Sandini, Cristina Martinoli, and David C Burr. "Impairment of auditory spatial localization in congenitally blind human subjects." In: *Brain* 137.1 (2014), pp. 288–293.

[67] Marcin Gorzel, David Corrigan, Gavin Kearney, John Squires, and Frank Boland. "Distance perception in virtual audio-visual environments." In: *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World (2012)*. 2012, pp. 1–8.

[68] Rishabh Gupta, Risabh Ranjan, Jianjun He, and Woon Seng Gan. "Study on differences between individualized and non-individualized hear-through equalization for natural augmented listening." In: *Audio Engineering Society Conference: 2019 AES International Conference on Headphones Technology*. Audio Engineering Society. 2019.

[69] Helmut Haas. "The influence of a single echo on the audibility of speech." In: *Journal of the Audio Engineering Society* 20.2 (1972), pp. 146–159.

[70] Deborah A Hall and David R Moore. "Auditory neuroscience: The salience of looming sounds." In: *Current Biology* 13.3 (2003), R91–R93.

[71] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho. "Augmented reality audio for mobile and wearable appliances." In: *Journal of the Audio Engineering Society* 52.6 (2004), pp. 618–639.

[72] William M Hartmann and Andrew Wittenberg. "On the externalization of sound images." In: *The Journal of the Acoustical Society of America* 99.6 (1996), pp. 3678–3688.

[73] William Morris Hartmann. "Listening in a room and the precedence effect." In: *Binaural and spatial hearing in real and virtual environments* (1997), pp. 191–210.

[74] BG Haustein. "Hypotheses about the one-eared distance perception of human obedience." In: *high frequency tech. and electroacoustics* 79 (1969), pp. 46–57.

[75] Florian Heller and Jan Borchers. "Audioscope: Smartphones as directional microphones in mobile audio augmented reality systems." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, pp. 949–952.

[76] Felix Henninger, Yury Shevchenko, Ulf K Mertens, Pascal J Kieslich, and Benjamin E Hilbig. "lab. js: A free, open, online study builder." In: *Behavior Research Methods* (2021), pp. 1–18.

[77] Joseph Henrich, Steven J Heine, and Ara Norenzayan. "Most people are not WEIRD." In: *Nature* 466.7302 (2010), pp. 29–29.

[78] Takayuki Hidaka, Leo L Beranek, and Toshiyuki Okano. "Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls." In: *The Journal of the Acoustical Society of America* 98.2 (1995), pp. 988–1007.

[79] Robert E Holt and Willard R Thurlow. "Subject orientation and judgment of distance of a sound source." In: *The Journal of the Acoustical Society of America* 46.6B (1969), pp. 1584–1585.

[80] Daniel P Jarrett, Emanuël AP Habets, and Patrick A Naylor. *Theory and applications of spherical microphone array processing*. Vol. 9. Springer, 2017.

[81] Walt Jesteadt, Craig C Wier, and David M Green. "Intensity discrimination as a function of frequency and sensation level." In: *The Journal of the acoustical society of America* 61.1 (1977), pp. 169–177.

[82] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel. "Analysis and synthesis of room reverberation based on a statistical time-frequency model." In: *Audio Engineering Society Convention 103*. Audio Engineering Society. 1997.

[83] Jean-Marc Jot and Antoine Chaigne. "Digital delay networks for designing artificial reverberators." In: *Audio Engineering Society Convention 90*. Audio Engineering Society. 1991.

[84] Jean-Marc Jot, Veronique Larcher, and Olivier Warusfel. "Digital signal processing issues in the context of binaural and transaural stereophony." In: *Audio Engineering Society Convention 98*. Audio Engineering Society. 1995.

[85] Eunice Jun, Gary Hsieh, and Katharina Reinecke. "Types of motivation affect study selection, attention, and dropouts in online experiments." In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–15.

[86] Gavin Kearney, Marcin Gorzel, Henry Rice, and Frank Boland. "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields." In: *Acta Acustica united with Acustica* 98.1 (2012), pp. 61–71.

[87] Hae-Young Kim, Yôiti Suzuki, Shouichi Takane, and Toshio Sone. "Control of auditory distance perception based on the auditory parallax model." In: *Applied Acoustics* 62.3 (2001), pp. 245–270.

[88] Sang-Myeong Kim and Wonjae Choi. "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach." In: *The Journal of the Acoustical Society of America* 117.6 (2005), pp. 3657–3665.

[89] Stephen Martin Kirkup. *The boundary element method in acoustics*. Integrated sound software, 2007.

[90] Andrew J Kolarik, Brian CJ Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan. "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss." In: *Attention, Perception, & Psychophysics* 78.2 (2016), pp. 373–395.

[91] Andrew J Kolarik, Shahina Pardhan, Silvia Cirstea, and Brian CJ Moore. "Using acoustic information to perceive room size: effects of blindness, room reverberation time, and stimulus." In: *Perception* 42.9 (2013), pp. 985–990.

[92] Norbert Kopčo, Matt Schoolmaster, and Barbara Shinn-Cunningham. "Learning to judge distance of nearby sounds in reverberant and anechoic environments." In: *Proc. Joint congress CFA/DAGA*. 2004.

[93] Norbert Kopčo and Barbara G Shinn-Cunningham. "Effect of stimulus spectrum on distance perception for nearby sources." In: *The Journal of the Acoustical Society of America* 130.3 (2011), pp. 1530–1541.

[94] Konrad Kowalczyk and Maarten Van Walstijn. "Room acoustics simulation using 3-D compact explicit FDTD schemes." In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.1 (2010), pp. 34–46.

[95] Asbjørn Krokstad, Staffan Strom, and Svein Sørsdal. "Calculating the acoustical room response by the use of a ray tracing technique." In: *Journal of Sound and Vibration* 8.1 (1968), pp. 118–125.

[96] Matthias Kronlachner and Franz Zotter. "Spatial transformations for the enhancement of Ambisonic recordings." In: *Proceedings of the 2nd International Conference on Spatial Audio, Erlangen*. 2014.

[97] Erik Larsen, Nandini Iyer, Charissa R Lansing, and Albert S Feng. "On the minimum audible difference in direct-to-reverberant energy ratio." In: *The Journal of the Acoustical Society of America* 124.1 (2008), pp. 450–461.

[98] Hyunkook Lee. "Apparent source width and listener envelopment in relation to source-listener distance." In: *Audio engineering society conference: 52nd international conference: Sound field control-engineering and perception*. Audio Engineering Society. 2013.

[99] Paul C Leopardi. "Distributing points on the sphere: partitions, separation, quadrature and energy." PhD thesis. University of New South Wales, Sydney, Australia, 2007.

[100] Robert W Lindeman, Haruo Noma, and Paulo Goncalves de Barros. "An empirical study of hear-through augmented reality: Using bone conduction to deliver spatialized audio." In: *2008 IEEE Virtual Reality Conference*. IEEE. 2008, pp. 35–42.

[101] Robert W Lindeman, Haruo Noma, and Paulo Gonçalves De Barros. "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio." In: *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE. 2007, pp. 173–176.

[102] JPA Lochner and JF Burger. "The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech." In: *Acta Acustica united with Acustica* 8.1 (1958), pp. 1–10.

[103] Tapio Lokki, Heli Nironen, Sampo Vesa, Lauri Savioja, Aki Härmä, and Matti Karjalainen. "Application scenarios of wearable and mobile augmented reality audio." In: *The 116th Convention of the Audio Engineering Society, Berlin, May 8-11 2004*. Audio Engineering Society. 2004, preprint–6026.

[104] Jack M Loomis, Roberta L Klatzky, John W Philbeck, and Reginald G Golledge. "Assessing auditory distance perception using perceptually directed action." In: *Perception & Psychophysics* 60.6 (1998), pp. 966–980.

[105] Justin A MacDonald, Paula P Henry, and Tomasz R Letowski. "Spatial audio through a bone conduction interface: Audición espacial a través de una interfase de conducción ósea." In: *International journal of audiology* 45.10 (2006), pp. 595–599.

[106] James C Makous and John C Middlebrooks. "Two-dimensional sound localization by human listeners." In: *The journal of the Acoustical Society of America* 87.5 (1990), pp. 2188–2200.

[107] Vincent Martin, Isabelle Viaud-Delmon, and Olivier Warusfel. "Virtual sound source distance evaluation in acoustically and visually incongruent contexts." In: *The Journal of the Acoustical Society of America* 150.4 (2021), A140–A141.

[108] Vincent Martin, Isabelle Viaud-Delmon, and Olivier Warusfel. "Effect of Environment-Related Cues on Auditory Distance Perception in the Context of Audio-Only Augmented Reality." In: *Applied Sciences* 12.1 (2022). ISSN: 2076-3417. DOI: 10.3390/app12010348. URL: https://www.mdpi.com/2076-3417/12/1/348.

[109] Vincent Martin, Olivier Warusfel, and Isabelle Viaud-Delmon. "Source distance modelling in the context of Audio Augmented Reality." In: *Forum Acusticum*. 2020, pp. 1369–1376.

[110] Pierre Massé, Thibaut Carpentier, Olivier Warusfel, and Markus Noisternig. "A robust denoising process for spatial room impulse responses with diffuse reverberation tails." In: *The Journal of the Acoustical Society of America* 147.4 (2020), pp. 2250–2260.

[111] Pierre Massé, Thibaut Carpentier, Olivier Warusfel, and Markus Noisternig. "Denoising directional room impulse responses with spatially anisotropic late reverberation tails." In: *Applied Sciences* 10.3 (2020), p. 1033.

[112] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. "OpenSesame: An open-source, graphical experiment builder for the social sciences." In: *Behavior research methods* 44.2 (2012), pp. 314–324.

[113] David McGookin and Stephen Brewster. "PULSE: An auditory display to provide a social vibe." In: *Proceedings of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*. 2011, pp. 12–15.

[114] Harry McGurk and John MacDonald. "Hearing lips and seeing voices." In: *Nature* 264.5588 (1976), pp. 746–748.

[115] Kittiphong Meesawat and Dorte Hammershoi. "The time when the reverberation tail in a binaural room impulse response begins." In: *Audio Engineering Society Convention 115*. Audio Engineering Society. 2003.

[116] Catarina Mendonça. "A review on auditory space adaptations to altered head-related cues." In: *Frontiers in neuroscience* 8 (2014), p. 219.

[117] Donald H Mershon, William L Ballenger, Alex D Little, Patrick L McMurtry, and Judith L Buchanan. "Effects of room reflectance and background noise on perceived auditory distance." In: *Perception* 18.3 (1989), pp. 403–416.

[118] Donald H Mershon, Douglas H Desaulniers, Thomas L Amerson, and Stephan A Kiefer. "Visual capture in auditory distance perception: proximity image effect reconsidered." In: *Journal of Auditory Research* (1980).

[119] Donald H Mershon and L Edward King. "Intensity and reverberation as factors in the auditory perception of egocentric distance." In: *Perception & Psychophysics* 18.6 (1975), pp. 409–415.

[120] Jean-Christophe Messonnier and Alban Moraud. "Auditory distance perception: criteria and listening room." In: *Audio Engineering Society Convention 130*. Audio Engineering Society. 2011.

[121] John C Middlebrooks, James C Makous, and David M Green. "Directional sensitivity of sound-pressure levels in the human ear canal." In: *The Journal of the Acoustical Society of America* 86.1 (1989), pp. 89–108.

[122] Paul Milgram and Fumio Kishino. "A taxonomy of mixed reality visual displays." In: *IEICE Transactions on Information and Systems* 77.12 (1994), pp. 1321–1329.

[123] George A Miller. "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness." In: *The Journal of the Acoustical Society of America* 19.4 (1947), pp. 609–619.

[124] Henrik Møller. "Fundamentals of binaural technology." In: *Applied acoustics* 36.3-4 (1992), pp. 171–218.

[125] Han-gil Moon, Jung-uk Noh, Koeng-mo Sung, and Dae-Young Jang. "Reverberation cue as a control parameter of distance in virtual audio environment." In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 87.7 (2004), pp. 1822–1826.

[126] Masayuki Morimoto and Kazuhiro Iida. "A practical evaluation method of auditory source width in concert halls." In: *Journal of the Acoustical Society of Japan (E)* 16.2 (1995), pp. 59–69.

[127] Masayuki Morimoto and Z Maekawa. "Effects of low frequency components on auditory spaciousness." In: *ACTA Acustica united with Acustica* 66.4 (1988), pp. 190–196.

[128] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. "Ambix-a suggested ambisonics format." In: *Ambisonics Symposium, Lexington*. 2011, p. 11.

[129] Anna N Nagele, Valentin M Bauer, Patrick GT Healey, Joshua D Reiss, Henry Cooke, Tim Cowlishaw, Chris Baume, and Chris Pike. "Interactive Audio Augmented Reality in Participatory Performance." In: *Frontiers in Virtual Reality* 1 (2020), p. 46.

[130] Rozenn Nicol. "Binaural Technology. AES Monograph." In: *Audio Engineering Society Inc* (2010).

[131] Søren H Nielsen. "Auditory distance perception in different rooms." In: *Audio Engineering Society Convention 92*. Audio Engineering Society. 1992.

[132] Toshiyuki Okano. "Judgments of noticeable differences in sound fields of concert halls caused by intensity variations in early reflections." In: *The Journal of the Acoustical Society of America* 111.1 (2002), pp. 217–229.

[133] Toshiyuki Okano, Leo L Beranek, and Takayuki Hidaka. "Relations among interaural cross-correlation coefficient (IACC E), lateral fraction (LF E), and apparent source width (ASW) in concert halls." In: *The Journal of the Acoustical Society of America* 104.1 (1998), pp. 255–265.

[134] Stefan Palan and Christian Schitter. "Prolific. ac—A subject pool for online experiments." In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27.

[135] Eyal Peer, David M Rothschild, Zak Evernden, Andrew Gordon, and Ekaterina Damer. "MTurk, Prolific or panels? Choosing the right audience for online research." In: *Choosing the right audience for online research (January 10, 2021)* (2021).

[136] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. "PsychoPy2: Experiments in behavior made easy." In: *Behavior research methods* 51.1 (2019), pp. 195–203.

[137] Z Ellen Peng, Emily Buss, Yi Shen, Hari Bharadwaj, G Christopher Stecker, Jordan A Beim, Adam K Bosen, Meredith Braza, Anna C Diedesch, Claire M Dorey, et al. "Remote testing for psychological and physiological acoustics: Initial report of the p&p task force on remote testing." In: *Proceedings of Meetings on Acoustics 179ASA*. Vol. 42. 1. Acoustical Society of America. 2020, p. 050009.

[138] John W Philbeck, Jack M Loomis, and Andrew C Beall. "Visually perceived location is an invariant in the control of action." In: *Perception & Psychophysics* 59.4 (1997), pp. 601–612.

[139] Jean-Dominique Polack. "La transmission de l'énergie sonore dans les salles." PhD thesis. Le Mans, 1988.

[140] Jean-Dominique Polack. "Modifying chambers to play billiards: the foundations of reverberation theory." In: *Acta Acustica united with Acustica* 76.6 (1992), pp. 256–272.

[141] Luna Prud'Homme and Mathieu Lavandier. "Do we need two ears to perceive the distance of a virtual frontal sound source?" In: *The Journal of the Acoustical Society of America* 148.3 (2020), pp. 1614–1623.

[142] Boaz Rafaely. "Analysis and design of spherical microphone arrays." In: *IEEE Transactions on speech and audio processing* 13.1 (2004), pp. 135–143.

[143] Boaz Rafaely. *Fundamentals of spherical array processing*. Vol. 16. Springer, 2018.

[144] Jussi Rämö and Vesa Välimäki. "Digital augmented reality audio headset." In: *Journal of Electrical and Computer Engineering* 2012 (2012).

[145] Rishabh Ranjan and Woon-Seng Gan. "Natural listening over headphones in augmented reality using adaptive filtering techniques." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.11 (2015), pp. 1988–2002.

[146] Francis Rumsey. "Headphone Technology: Hear-Through, Bone Conduction, and Noise Canceling." In: *Journal of the Audio Engineering Society* 67.11 (2019), pp. 914–919.

[147] Jesper Sandvad. "Auditory perception of reverberant surroundings." In: *The Journal of the Acoustical Society of America* 105.2 (1999), pp. 1193–1193.

[148] Marian Sauter, Dejan Draschkow, and Wolfgang Mack. "Building, hosting and recruiting: A brief introduction to running behavioral experiments online." In: *Brain sciences* 10.4 (2020), p. 251.

[149] Manfred R Schroeder. "Integrated-impulse method measuring sound decay without using impulses." In: *The Journal of the Acoustical Society of America* 66.2 (1979), pp. 497–500.

[150] Manfred R Schroeder and KH Kuttruff. "On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima." In: *The Journal of the Acoustical Society of America* 34.1 (1962), pp. 76–80.

[151] Russell D Shilling and Barbara Shinn-Cunningham. "Virtual auditory displays." In: *Handbook of Virtual Environments*. CRC Press, 2002, pp. 105–132.

[152] Barbara G Shinn-Cunningham, Nathaniel I Durlach, and Richard M Held. "Adapting to supernormal auditory localization cues. I. Bias and resolution." In: *The Journal of the Acoustical Society of America* 103.6 (1998), pp. 3656–3666.

[153] Barbara G Shinn-Cunningham, Norbert Kopco, and Tara J Martin. "Localizing nearby sound sources in a classroom: Binaural room impulse responses." In: *The Journal of the Acoustical Society of America* 117.5 (2005), pp. 3100–3115.

[154] T Shuku and K Ishihara. "The analysis of the acoustic field in irregularly shaped rooms by the finite element method." In: *Journal of Sound and Vibration* 29.1 (1973), 67–IN1.

[155] Jon M Speigle and Jack M Loomis. "Auditory distance perception by translating observers." In: *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*. IEEE. 1993, pp. 92–99.

[156] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. "Comparison of different impulse response measurement techniques." In: *Journal of the Audio engineering society* 50.4 (2002), pp. 249–262.

[157] Rebecca Stewart and Mark Sandler. "Statistical measures of early reflections of room impulse responses." In: *Proc. of the 10th int. conference on digital audio effects (DAFx-07), Bordeaux, France*. 2007, pp. 59–62.

[158] Peter Stitt, Lorenzo Picinali, and Brian FG Katz. "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning." In: *Scientific reports* 9.1 (2019), pp. 1–14.

[159] Gijsbert Stoet. "PsyToolkit: A software package for programming psychological experiments using Linux." In: *Behavior research methods* 42.4 (2010), pp. 1096–1104.

[160] Ivan E Sutherland. "A head-mounted three dimensional display." In: *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 1968, pp. 757–764.

[161] Peter W Tappan. "Proximaural loudspeakers (-nearphones-)." In: *Audio Engineering Society Convention 16*. Audio Engineering Society. 1964.

[162] Miikka Tikander, Matti Karjalainen, and Ville Riikonen. "An augmented reality audio headset." In: *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08), Espoo, Finland*. 2008.

[163] Alan M Turing and J Haugeland. *Computing machinery and intelligence*. MIT Press Cambridge, MA, 1950.

[164] Jesper Udesen, Tobias Piechowiak, and Fredrik Gran. "The effect of vision on psychoacoustic testing with headphone-based virtual sound." In: *Journal of the Audio Engineering Society* 63.7/8 (2015), pp. 552–561.

[165] Daniel L Valente and Jonas Braasch. "Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues." In: *The Journal of the Acoustical Society of America* 128.4 (2010), pp. 1952–1964.

[166] Chiara Valzolgher, Mariam Alzhaler, Elena Gessa, Michela Todeschini, Pauline Nieto, Gregoire Verdelet, Romeo Salemme, Valerie Gaveau, Mathieu Marx, Eric Truy, et al. "The impact of a visual spatial frame on real sound-source localization in virtual reality." In: *Current Research in Behavioral Sciences* 1 (2020), p. 100003.

[167] Virginie Van Wassenhove, Ken W Grant, and David Poeppel. "Temporal window of integration in auditory-visual speech perception." In: *Neuropsychologia* 45.3 (2007), pp. 598–607.

[168] Michael Vorländer and Jason E Summers. "Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and acoustic virtual reality." In: *Acoustical Society of America Journal* 123.6 (2008), p. 4028.

[169] Bruce N Walker and Jeffrey Lindsay. "Navigation performance in a virtual environment with bonephones." In: Georgia Institute of Technology. 2005.

[170] H Kenneth Walker, W Dallas Hall, and J Willis Hurst. "Clinical methods: the history, physical, and laboratory examinations." In: (1990).

[171] David H Warren. "Intermodality interactions in spatial localization." In: *Cognitive Psychology* 1.2 (1970), pp. 114–133.

[172] David H Warren, Robert B Welch, and Timothy J McCarthy. "The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses." In: *Perception & Psychophysics* 30.6 (1981), pp. 557–564.

[173] Richard M Warren. *Auditory perception: A new synthesis*. Vol. 109. Elsevier, 2013.

[174] Robert B Welch. "Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions." In: *Advances in psychology*. Vol. 129. Elsevier, 1999, pp. 371–387.

[175] Robert B Welch and David H Warren. "Immediate perceptual response to intersensory discrepancy." In: *Psychological bulletin* 88.3 (1980), p. 638.

[176] Stephan Werner, Florian Klein, Thomas Mayenfels, and Karlheinz Brandenburg. "A summary on acoustic room divergence and its effect on externalization of auditory events." In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2016, pp. 1–6.

[177] Stephan Werner, Florian Klein, Annika Neidhardt, Ulrike Sloma, Christian Schneiderwind, and Karlheinz Brandenburg. "Creation of Auditory Augmented Reality Using a Position-Dynamic Binaural Synthesis System—Technical Components, Psychoacoustic Needs, and Perceptual Evaluation." In: *Applied Sciences* 11.3 (2021), p. 1150.

[178] Pavel Zahorik. "Distance localization using nonindividualized head-related transfer functions." In: *The Journal of the Acoustical Society of America* 108.5 (2000), pp. 2597–2597.

[179] Pavel Zahorik. "Assessing auditory distance perception using virtual acoustics." In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1832–1846.

[180] Pavel Zahorik. "Auditory display of sound source distance." In: *Proc. Int. Conf. on Auditory Display*. 2002, pp. 326–332.

[181] Pavel Zahorik. "Direct-to-reverberant energy ratio sensitivity." In: *The Journal of the Acoustical Society of America* 112.5 (2002), pp. 2110–2117.

[182] Pavel Zahorik, Philbert Bangayan, V Sundareswaran, Kenneth Wang, and Clement Tam. "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals." In: *The Journal of the Acoustical Society of America* 120.1 (2006), pp. 343–359.

[183] Pavel Zahorik, Douglas S Brungart, and Adelbert W Bronkhorst. "Auditory distance perception in humans: A summary of past and present research." In: *ACTA Acustica united with Acustica* 91.3 (2005), pp. 409–420.

[184] Massimiliano Zampini, Steve Guest, David I Shore, and Charles Spence. "Audio-visual simultaneity judgments." In: *Perception & psychophysics* 67.3 (2005), pp. 531–544.

[185] Andreas Zimmermann and Andreas Lorenz. "LISTEN: a user-adaptive audio-augmented museum guide." In: *User Modeling and User-Adapted Interaction* 18.5 (2008), pp. 389–416.

[186] Franz Zotter and Matthias Frank. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[187] Franz Zotter, Matthias Frank, and Hannes Pomberger. "Comparison of energy-preserving and all-round ambisonic decoders." In: *Fortschritte der Akustik, AIA-DAGA,(Meran)* (2013).

[188] Franz Zotter, Hannes Pomberger, and Markus Noisternig. "Energy-preserving ambisonic decoding." In: *Acta Acustica united with Acustica* 98.1 (2012), pp. 37–47.