

Auditory externalization of a remote microphone signal

Présentée le 25 mars 2022

Faculté des sciences et techniques de l'ingénieur
Laboratoire de traitement des signaux 2
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Vincent Pierre Olivier GRIMALDI

Acceptée sur proposition du jury

Prof. J.-Ph. Thiran, président du jury
Prof. P. Vandergheynst, Dr H. Lissek, directeurs de thèse
Prof. P. Zahorik, rapporteur
Dr C. Faller, rapporteur
Prof. J.-M. Odobez, rapporteur

The final question will be:
Is the soundscape of the world an indeterminate composition over which we have no control,
or are we its composers and performers, responsible for giving it form and beauty?
— R. Murray Schafer, *The Soundscape*

To my family and friends . . .

Acknowledgements

En préambule de ce manuscrit, je souhaite adresser mes plus sincères remerciements aux personnes qui ont contribué à la réussite de cette thèse de doctorat.

J'aimerais tout d'abord remercier mon superviseur et co-directeur de thèse, le Dr. Hervé Lissek. Merci de m'avoir permis de rejoindre le groupe acoustique et donné la chance de travailler sur ce sujet stimulant et passionnant, répondant totalement à mes aspirations. Je te remercie également pour ta confiance et pour m'avoir permis de travailler librement et en autonomie, tout en me faisant profiter de ton expérience et de tes conseils aux moments opportuns. J'aimerais également remercier mon directeur de thèse, le Prof. Pierre Vandergheynst pour sa confiance, et grâce à qui le Groupe Acoustique possède un toit au LTS2 pour mener de passionnantes recherches.

Ensuite, je tiens à remercier le Dr. Gilles Courtois qui a initié le beau projet qui m'a permis de rejoindre le Groupe Acoustique. C'était une véritable chance de collaborer avec toi dès le début de ma thèse, et de pouvoir profiter de ton expertise et de tes connaissances pour rentrer directement dans le bain. Merci également pour la supervision depuis Stäfa par la suite, et de manière générale pour toutes les discussions intéressantes qui ont contribué aux succès des projets menés ensemble. J'exprime ensuite ma gratitude au Dr. Laurent Simon, avec qui j'ai eu également un grand plaisir à collaborer. Merci pour ton enthousiasme et pour tous les échanges passionnants que nous avons eus, en particulier sur les sujets de perception auditive. Ces discussions, ainsi que tes conseils avisés, ont été précieux dans le succès de cette thèse. Merci également pour ta relecture détaillée de ce manuscrit.

J'adresse un grand merci à Miquel Sans pour son temps, son implication et le partage de son expertise qui ont été fructueux lors d'une étape importante de cette thèse. Merci également à la Dr. Eleftheria Georganti qui a participé à la supervision depuis Stäfa avec bienveillance au début de mon doctorat. Je remercie de manière plus globale l'ensemble des membres de Sonova qui ont pu ponctuellement procurer d'intéressantes idées ou intuitions. Je souhaite aussi remercier la Dr. Ina Kodrasi pour son aide précieuse en début de thèse.

J'ai une pensée émue à la mémoire de Philippe Estoppey, audioprothésiste de renom, avec qui j'ai eu le bonheur de travailler. Merci d'avoir rendu possible une étude si enrichissante avec tes patient-e-x-s. Merci pour ton professionnalisme, ta bienveillance et ton enthousiasme.

Je tiens à adresser mes remerciements au Prof. Pavel Zahorik, au Dr. Christof Faller ainsi qu'au Prof. Jean-Marc Odobez de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse. J'ai fortement apprécié les conversations passionnantes et enrichissantes que nous avons eues lors de la soutenance. Merci également au Prof. Jean-Philippe Thiran d'avoir présidé et orchestré la défense.

Je remercie ensuite mes collègues du Groupe Acoustique : Maxime Volery, Stanislav Sergeev, Mathieu Padlewski, Xinxin Guo, Etienne Rivet et Romain Boulandet pour la bonne ambiance, les cafés, les bonnes bières, et les super discussions. Un salut particulier à Thach Pham Vu et Thomas Laurence pour tous les bons moments partagés en Crète et à Montréal, les dégustations et les sessions "music room". Merci à Eva Mompert et Mercedes Quintas pour l'efficacité et la bonne humeur constante.

Je souhaite exprimer ma gratitude à tou-te-x-s les participant-e-x-s qui ont pris part aux études de perception que j'ai pu mener lors de ces quatre années. Merci également aux collègues du couloir ELB0 (Ismael et Adrian en particulier), ainsi qu'à Simon et Patrick qui ont bien voulu jouer les cobayes lors des préparations de ces études.

Merci aux étudiant-e-s dont j'ai eu grand plaisir à superviser les projets de semestre et stages, et qui ont participé avec enthousiasme à des projets liés à mes recherches: Thien-Anh Nguyen, Gloria Dal Santo, Melina Chrysanthou, Alon Tchelet, Samuel Beuret, David Sanchez, Lucien Barret, Alexandre Damia et Simon Canales.

Enfin, je souhaite aussi remercier ma famille pour son soutien sans faille. Je salue également mes ami-e-s, avec un clin d'œil particulier à Louis et Sara. Et bien sûr, merci à Fanny pour ton amour inconditionnel, ton soutien, et le bonheur quotidien que tu m'apportes.

Lausanne, le 21 Février 2022

Vincent

Abstract

A remote microphone (RM) system can be used in combination with wearable binaural communication devices, such as hearing aids (HAs), to improve speech intelligibility. Typically, a speaker is equipped with a body-worn microphone which enables to pick up their voice with a high signal-to-noise ratio (SNR). However, if this signal is played diotically through the receivers (i.e. the same signal in both ears), the necessary cues enabling the auditory system to locate the sound source are bypassed. This can affect the ability of the listener to feel immersed in the environment or to follow a conversation, especially for hearing-impaired (HI) listeners. Auditory sound source localization in humans is performed by the auditory system owing to the interpretation of various cues related to the physical sound propagation from this source to the eardrums of the listener. This is mainly enabled by the binaural structure of the auditory system. Previous works have successfully developed a method to provide a simplified spatialization of the RM signal which enabled normal-hearing (NH) and HI listeners to locate sound sources in azimuth, while preserving speech intelligibility. However, the proposed method yielded common spatial hearing perceptual artefacts, such as in-head localization and front/back confusion.

This thesis is devoted to the investigation and the perceptual evaluation of wearable devices-compatible audio playback solutions aiming at enhancing the realism of this spatialization. In particular, the goal is to improve the externalization of a sound source, i.e. to ensure it is perceived outside the head, and possibly at a certain distance corresponding to its physical location in the environment. For this purpose, early reflections (ERs) in a room play a key role. Hence, several signal processing approaches to provide ERs in the binaural synthesis were investigated. Subsequently, a subjective listening study was conducted, in which NH and HI aided listeners had to evaluate the perceived auditory distance with various binaural rendering strategies. The results show that the superimposition of ERs with the considered methods significantly improves the perception of auditory distance for NH and HI listeners. The study also provides insights about auditory distance perception in aided HI listeners with severe-to-profound hearing loss. A follow-up study was conducted with NH listeners and showed that a complete implementation of these strategies might improve auditory distance perception while preserving spatial awareness. In these studies, subjects had their head fixed. Previous studies have shown that head movements coupled with head-tracking can contribute

to the auditory externalization of a virtual sound source. The next part of this thesis reports the development of a head-tracking algorithm compatible with wearable devices, relying solely on two 3-axis accelerometers. While showing promising results, the limitations of the developed algorithm still yield mismatches in the estimation. Consequently, a subjective listening test was conducted with NH listeners to study the effect of head-tracking artefacts on the perception of externalization and the performance in azimuth localization of a sound source. The results suggest that auditory externalization is not affected by a large latency or the amplitude mismatch. Latency did not decrease the performance in localization either, contrarily to the amplitude mismatch.

Key words: auditory externalization, auditory distance, hearing aids, hearables, remote microphone, early reflections, head-tracking, head movements

Résumé

Un système de microphone distant (RM) peut être utilisé en combinaison avec des dispositifs de communication binaurale portables tels que des aides auditives, pour améliorer l'intelligibilité de la parole. En général, un-e locuteur-rice est équipé d'un microphone qui permet de capter sa voix avec un rapport signal/bruit (SNR) élevé. Toutefois, si ce signal est diffusé de manière diotique dans les récepteurs de l'appareil (i.e. le même signal est utilisé pour les deux oreilles), les indices nécessaires à la localisation de la source sonore par le système auditif ne sont pas restitués. Cela peut affecter la capacité de l'auditeur-rice à se sentir inclus-e dans l'environnement ou à suivre une conversation, en particulier chez les personnes malentendantes (ME). La localisation d'une source sonore chez l'humain peut être effectuée par le système auditif grâce à l'interprétation de divers indices liés à la propagation physique du son depuis cette source jusqu'aux tympans de l'auditeur-rice. Cette capacité est notamment rendue possible par la structure binaurale du système auditif. Des travaux antérieurs ont permis de développer une méthode de spatialisation simplifiée du signal RM qui permet aux auditeur-rice-s normo-entendant-e-s (NE) et ME de localiser les sources sonores en azimut, tout en préservant l'intelligibilité de la parole. Cependant, la méthode proposée a donné lieu à des artefacts de perception auditive spatiale courants, tels que la localisation à l'intérieur de la tête ou la confusion avant/arrière.

Cette thèse est consacrée à l'étude et à l'évaluation perceptive de solutions techniques compatibles avec les dispositifs portables et visant à améliorer le réalisme de cette spatialisation. En particulier, l'objectif est d'améliorer l'externalisation d'une source sonore, c'est-à-dire de permettre qu'elle soit perçue en dehors de la tête, et éventuellement à une certaine distance correspondant à son emplacement physique dans l'environnement. À cette fin, les réflexions précoces dans une pièce jouent un rôle clé. Par conséquent, plusieurs approches de traitement du signal visant à intégrer des réflexions précoces dans la synthèse binaurale ont été étudiées. Par la suite, une étude d'écoute subjective a été menée : des participant-e-s NE et ME devaient évaluer la distance auditive perçue avec différentes stratégies de synthèse binaurale. Les résultats montrent que la superposition des réflexions précoces grâce aux méthodes considérées améliore significativement la perception de la distance auditive chez les auditeur-rice-s NE et ME. L'étude fournit également de nouveaux résultats sur la perception de la distance auditive chez les sujets ME portant des aides auditives et présentant une perte auditive sévère.

à profonde. Une étude similaire a été menée avec des auditeur·rice·s NE et a montré qu'une implémentation complète de ces stratégies pourrait améliorer la perception de la distance auditive tout en préservant la conscience spatiale des autres sons de l'environnement. Dans ces études, les sujets avaient la tête fixe. Des études précédentes ont démontré que les mouvements de la tête couplés au "head-tracking" contribuent à améliorer l'externalisation d'une source sonore virtuelle. La partie suivante de cette thèse rapporte le développement d'un algorithme de head-tracking compatible avec les dispositifs d'écoute portables, reposant uniquement sur deux accéléromètres 3-axes. Bien que les résultats soient prometteurs, les limitations de l'algorithme développé entraînent encore des erreurs d'estimation. Par conséquent, un test d'écoute subjectif a été réalisé avec des participant·e·s NE pour étudier l'effet des artefacts de suivi de la tête sur la perception de l'externalisation et la performance en localisation azimutale d'une source sonore. Les résultats suggèrent que l'externalisation auditive n'est pas affectée par une latence importante ou un décalage d'amplitude de l'estimation. La latence ne diminue pas non plus les performances de localisation, contrairement au décalage d'amplitude.

Mots clefs : perception, auditive, distance, externalisation, aides auditives, mouvements de tête, head-tracking, hearables.

Contents

Acknowledgements	i
Abstract (English/Français)	iii
List of Figures	xi
List of Tables	xv
List of Acronyms	xvii
1 Introduction	1
1.1 Context and motivation	1
1.1.1 Hearing aids	1
1.1.2 Hearables	2
1.2 Outline and contributions	3
2 Spatialization of the remote microphone and auditory externalization	5
2.1 The remote microphone system	5
2.2 Fundamentals of spatial hearing	6
2.2.1 Localization in azimuth	6
2.2.2 Localization in elevation	7
2.2.3 Head-related transfer functions	8
2.2.4 Localization in a room	9
2.2.5 Dynamic binaural synthesis	9
2.3 Spatialization of the remote microphone	10
2.4 Perceptual limitations of the baseline binaural synthesis	12
2.5 Auditory externalization	13
2.5.1 Definition	13
2.5.2 Relation to auditory distance: a continuum or two different dimensions ?	14
2.5.3 Methods for measuring externalization and auditory distance	14
2.6 Externalization and distance perception for normal-hearing subjects	15
2.6.1 Role of reflections	16
2.6.2 Role of binaural cues	17
2.6.3 Level cues	18
2.6.4 Spectral cues	18

2.6.5	Laterality	19
2.6.6	Visual cues	19
2.6.7	Head movements and head-tracking	20
2.6.8	Room divergence effect	21
2.7	Auditory externalization and distance perception in HI listeners and HAs users	21
3	Adding early reflections to the remote microphone signal	25
3.1	Multi-Channel Wiener Filter	25
3.1.1	Model	25
3.1.2	Principle	26
3.1.3	Insertion of an additional remote microphone	27
3.1.4	Implementation and tuning	28
3.1.5	Objective assessment of the MWF method	29
3.2	Multi-band envelope processing	31
3.2.1	Principle	31
3.2.2	Avantages and outlook	33
3.3	Coherence-based method	34
3.4	Partitioned convolution	38
3.4.1	Uniform partitioned convolution	38
3.4.2	Non-uniform partitioned convolution	40
3.5	Discussion	41
3.5.1	Performance	41
3.5.2	Perceptual considerations	42
4	Auditory distance perception in NH and HI listeners with various binaural rendering strategies	45
4.1	Context and motivation	45
4.2	Methods	47
4.2.1	Participants	47
4.2.2	Experimental setup and measurement phase	48
4.2.3	Stimuli	49
4.2.4	Task	50
4.3	Results	50
4.3.1	Auditory distance perception in NH listeners	51
4.3.2	Auditory distance perception in HI listeners	52
4.3.3	Auditory distance perception and PTA	54
4.4	Discussion	54
4.4.1	Perceived auditory distance in NH and HI listeners	54
4.4.2	Considerations on the RM system for HA application	56
4.4.3	Limitations of the study	57
4.5	Conclusion and perspectives	57

CONTENTS

5	Auditory distance perception of a RM signal using low computational cost algorithms	59
5.1	Context and motivation	59
5.2	Description of the algorithms	61
5.2.1	Remote Microphone + ambient Microphone (RMM)	61
5.2.2	Early Reflections extraction and Cleaning (ERC)	61
5.2.3	Partitioned Convolution (PConv)	63
5.3	Experiment 1: Auditory distance perception	63
5.3.1	Setup	63
5.3.2	Stimuli	64
5.3.3	Procedure	65
5.3.4	Results	65
5.3.5	Discussion	67
5.4	Experiment 2: Spatial awareness	68
5.4.1	Setup	68
5.4.2	Stimuli	68
5.4.3	Procedure	69
5.4.4	Results	69
5.4.5	Discussion	71
5.5	Conclusion and perspectives	72
6	Design of a head-tracking algorithm using two 3-axis accelerometers	73
6.1	Motivation to include head-tracking in binaural communication devices	73
6.2	Equipment	74
6.2.1	Reference tracking	74
6.2.2	Prototype	74
6.2.3	Turntable	75
6.3	Context of the implementation	76
6.4	Description of the algorithm	78
6.4.1	Description of the model	78
6.4.2	Base algorithm	79
6.4.3	Upgrades of the algorithm for the head tracking application	81
6.4.4	Real-time implementation	91
6.5	Evaluation of the algorithm	92
6.5.1	Database	92
6.5.2	Evaluation Method	94
6.5.3	Results	94
6.6	Limitations of the algorithm	97
6.6.1	Measure dependent performance	97
6.6.2	Other limitations and related perspectives for improvement	99
6.7	Conclusion	100

7	Effects of head-tracking artefacts on externalization and localization	101
7.1	Introduction	101
7.1.1	Context and motivation	101
7.1.2	Contribution of head-tracking on spatial perception in literature	102
7.1.3	Goal	104
7.2	Spatialization algorithm and simulated artefacts	105
7.2.1	Binaural synthesis algorithm	105
7.2.2	Head-tracking conditions	106
7.3	Participants	107
7.4	Externalization experiment protocol	108
7.4.1	Setup	108
7.4.2	Preparation / Training	109
7.4.3	Stimuli and conditions	110
7.4.4	Task	111
7.5	Externalization experiment results	112
7.5.1	Auditory externalization	112
7.5.2	Head movements	114
7.6	Externalization experiment discussion	116
7.6.1	Effect of the head-tracking condition	116
7.6.2	Effect of the early reflections and binaural cues	117
7.6.3	Considerations for applications to binaural communication devices	117
7.7	Localization experiment: protocol	119
7.7.1	Setup	119
7.7.2	Stimuli	120
7.7.3	Procedure	120
7.8	Localization experiment: results	121
7.8.1	Localization error	121
7.8.2	Localization time	125
7.9	Localization experiment: discussion	127
7.9.1	Localization error	127
7.9.2	Localization time	128
7.9.3	Considerations for applications to binaural communication devices	128
7.10	Summary and conclusion	129
8	Conclusion	131
A	Additional figures and data	137
	Bibliography	159
	Curriculum Vitae	161

List of Figures

2.1	Wireless microphone system for hearing aids [44].	6
2.2	Example of hearing aid equipped with two omni-directional microphones: Phonak Audéo P.	6
2.3	Representation of the dimensions used to locate a sound source in spherical coordinates: azimuth, elevation and distance (taken from [140]).	7
2.4	Illustration of the geometric difference in distance used to compute the ITD with Woodworth's formula.	8
2.5	Components of a room impulse response.	9
2.6	Illustration of the cone of confusion, with a sound source and its mirror image.	10
2.7	Initial position (left panel) and position after head movement (right panel), Without head-tracking it is not possible to provide the synthesis of a source that appears at the correct position regardless of the head orientation (B), with head-tracking, the filtering can be adapted to the head orientation (A).	10
2.8	Schematic representation of the principle of the localization and spatialization algorithm for remote microphone systems (applicable to HAs or hearables). . .	11
2.9	In-head localization (left panel) vs. externalization (right panel) of a sound source.	13
3.1	Multi-channel Wiener Filter of a 2x2 microphones binaural hearing device. . .	26
3.2	Multi-channel Wiener Filter of a 2x2 microphones binaural hearing device with additional remote signal.	28
3.3	Male speech in babble noise, effect of the MWF processing for two SNR settings (SNR = 0 or 10 dB).	29
3.4	Energy comparison between the satellite and the RM signals in different Bark bands.	32
3.5	Normalized cross-correlation between the envelopes of the signals picked up by the remote and satellite microphones for various rooms and speaker (male/fe- male).	32
3.6	Male speech in babble noise, effect of the Multi-band envelope processing for two SNR settings (SNR = 0 or 10 dB).	34
3.7	Simplified structure of the coherence-based method for ERs extraction, for the left HA.	35
3.8	Definition of a mapping function used to compute a gain, based on the coher- ence of the two signals and their cumulative distributions (cum. dist.) [37]. . . .	36

3.9	Objective metrics as a function of the input SNR, for the performance comparison between the MFW, the MWF with RM (MWF - RM) and the coherence-based method (COH). The measures were made with diffuse babble noise in a reverberant room (classroom at EPFL) (taken from [42]).	37
3.10	Impulse response partition in the case of a uniform partition.	39
3.11	Graphic representation of the real-time implementation of the partitioned convolution algorithm (taken from [5]).	39
3.12	Impulse response partition in the case of a non-uniform partition.	40
4.1	Schematic representation of the experimental setup mounted in a classroom. .	48
4.2	Auditory distance as evaluated by NH listeners on a continuous scale with the following markers: <i>Center of the head</i> (0), <i>Boundary of the head</i> (20), <i>At Loudspeaker 1</i> (40), <i>At Loudspeaker 2</i> (60), <i>At Loudspeaker 3</i> (80) and <i>Further than Loudspeaker 3</i> (80 to 100).	51
4.3	Auditory distance as evaluated by HI listeners on a continuous scale with the following markers: <i>Center of the head</i> (0), <i>Boundary of the head</i> (20), <i>At Loudspeaker 1</i> (40), <i>At Loudspeaker 2</i> (60), <i>At Loudspeaker 3</i> (80) and <i>Further than Loudspeaker 3</i> (80 to 100). WDRC-A is the case where the WDRC is performed after the spatial processing, and WDRC-B is the case where the WDRC is performed before. . .	52
4.4	Perceived auditory distance as a function of the PTA at best ear for the three stimuli with significant correlation.	55
5.1	Block diagram of the RMM algorithm.	61
5.2	Block diagram of the ERC algorithm.	62
5.3	Example of some complementary filter gains at a random frame (frequency bins), A_{erc} and A_{res} , with $G_{res} = 0.25$	62
5.4	Block diagram of the PConv algorithm.	63
5.5	Schematic representation of the setup for Experiment 1, mounted in a listening room.	64
5.6	Auditory distance estimations evaluated on a continuous scale with the following markers: <i>Inside of the head</i> (0), <i>Border of the head</i> (20), <i>At Loudspeaker (LS) 1</i> (40), <i>At LS 2</i> (60), <i>At LS 3</i> (80) and <i>Further than LS 3</i> (80 to 100).	66
5.7	Auditory scene for Experiment 2, simulated over headphones using binaural synthesis.	68
5.8	Example of stimuli structure in time, for the main and competing speakers (Experiment 2).	69
5.9	Scores for the number of detections of the "reverse" speech in the competing speaker.	70
5.10	Reaction time in the speech reverse detection task.	71
6.1	Head-tracker software with real-time Euler angles tracking and 3D visualization of the tracker.	74
6.2	Prototype HAs and audio interface.	75

LIST OF FIGURES

6.3	Setup for the control of the turntable.	75
6.4	Two IMU/AHRS devices mounted on a turntable.	76
6.5	Artificial head equipped with the prototype, and mounted on the motor of the turntable.	76
6.6	Human head angular motions, with the yaw, pitch and roll (image from [112]). .	77
6.7	Accelerometers setup. The accelerometers A1 and A2 are spaced by a distance D and supposed to be aligned together with the center of the system on the y-axis. .	79
6.8	Example of the computation of Ω_T in relation to the evolution of Ω_r	82
6.9	Power spectral density as a function of frequency in slow (a) and fast (b) mode. .	85
6.10	Mapping function used to define $\lambda(n)$	87
6.11	Simplified block diagram of the real-time implementation, showing the main steps and data. $(a_{x'1}, a_{y'1}, a_{z'1})$ and $(a_{x''2}, a_{y''2}, a_{z''2})$ denote the accelerations of the sensors 1 and 2 after applying their respective rotation matrices.	91
6.12	Schematic representation of the setup used for recording the database, the loudspeakers were placed respectively at -90° , -45° , -0° , $+45^\circ$ and $+90^\circ$	93
6.13	Setup for the acquisition of the database.	94
6.14	Fitting of the prototype with embedded accelerometers.	94
6.15	RMS Error as a function of the measurement. The letter corresponds to the anonymized participant, and the number is the repetition when the subject's recordings led to several usable measures. $n = 576$ combinations of the parameters were performed for each measure.	95
6.16	Examples of various degrees of performance of the algorithm. The plots show the estimations and the related references for several measures with different participants and their best setting. (a) Illustrates an example of a good performance of the algorithm. (b) Shows an example of an overall acceptable estimation with small errors yielding local mismatches. (c) Illustrates a scenario where most of the movements are not detected. (d) Shows the worst type of performance, where large errors yield a general drift of the estimation.	96
6.17	Mean b_0 and main effects for the seven parameters and twelve selected measurements obtained from the factorial analysis.	98
7.1	Block diagram of the non-individualized spatialization algorithm used in this experiment.	105
7.2	Example of the simulation of the latency artefact, in this case with a delay of 400 ms.	107
7.3	Example of the simulation of the amplitude mismatch artefact.	108
7.4	Participant equipped with the headphones and head-tracking device in the listening room.	109
7.5	Example of the simulation of the amplitude mismatch artefact adapted to the externalization experiment.	111
7.6	Externalization ratings per head-tracking condition and DRR setting (z-scores transformed).	113

7.7	Schematic representation of the setup used of the localization experiment, with 8 loudspeakers hidden behind black acoustically transparent curtains, and the listener equipped with headphones and the IMU/AHRS head-tracking device. .	119
7.8	Absolute error of localization for each head-tracking condition.	121
7.9	Absolute error in localization, effect of the target azimuth for each head-tracking condition.	124
7.10	Scores for localization time per head-tracking condition.	125
7.11	Localization time as a function of the repetition number.	126
A.1	Audiograms, measured at the best ear, of the NH listeners (blue) and HI listeners (green, yellow, orange). The upper and lower thicker black lines corresponds to the average of the NH listeners and HI listeners respectively. (Chapter 4).	137
A.2	Intelligibility scores in the spatial awareness experiment (Chapter 5).	138
A.3	Histogram of the z-score externalization rating per head-tracking condition (Chapter 7).	140
A.4	QQ-Plots of the residuals per head-tracking condition for the externalization ratings (z-scores) (Chapter 7).	141
A.5	Histogram of the absolute error of localization (cubic root transformed) per head-tracking condition (Chapter 7).	142
A.6	QQ-Plots of the residuals per head-tracking condition for the absolute error of localization (cubic root transformed) (Chapter 7).	143
A.7	Histogram of the absolute error of localization (cubic root transformed) per head-tracking condition (Chapter 7).	144
A.8	QQ-Plots of the residuals per head-tracking condition for the absolute error of localization (cubic root transformed) (Chapter 7).	145
A.9	Complete <i>Post hoc</i> Tukey HSD for the effect of the target azimuth on the localization absolute error in the Real Sources (a) and Amp. Artefact (b) conditions. The values in the "contrast" column correspond to the target azimuths in degrees (Chapter 7).	146

List of Tables

3.1	Scores obtained with the objective metrics, with the original recording, the Muli-channel Wiener filter, and the Muli-channel Wiener filter with RM.	30
3.2	Summary of the main advantages and drawbacks of the investigated methods	43
4.1	Perceived auditory distance (NH listeners), Wilcoxon signed-rank tests results (significant p -values are in blue).	51
4.2	Perceived auditory distance (HI listeners, WDRC-A), Wilcoxon signed-rank tests results (significant p -values are in blue).	53
4.3	Perceived auditory distance (HI listeners, WDRC-B), Wilcoxon signed-rank tests results (significant p -values are in blue).	53
4.4	Spearman's correlations coefficient (ρ) and associated p -values for the perceived auditory distance and the PTA at the best ear of HI listeners in the WDRC-A and WDRC-B conditions. Significant correlations are in blue.	54
5.1	Perceived auditory distance, Wilcoxon signed-rank tests results (significant p -values are in blue).	66
6.1	List of the tested parameters and their values.. . . .	95
6.2	Coefficients weights of the seven parameters used in the factorial analysis	99
7.1	List of conditions used in the externalization experiment.	110
7.2	Medians, interquartile ranges and Spearman's ρ in relation to the z-score of externalization for several dependent variables of the participant's movements.	115
7.3	List of conditions used for the localization experiment.	120
7.4	Medians and interquartile ranges of the absolute error in azimuth localization for each head-tracking condition.	121
7.5	Medians and interquartile ranges of the absolute error in azimuth localization for each head-tracking condition and target azimuth, formatted as median(interquartile range).	123
7.6	Medians and interquartile ranges of the time of localization for each head-tracking condition.	125
A.1	WHO's grades of hearing impairment [132].	137

LIST OF TABLES

A.2	ANOVA Table with Greenhouse-Geisser correction : Z-score of externalization (Chapter 7).	138
A.3	ANOVA Table : Error of localization (Chapter 7).	139
A.4	ANOVA Table : Time of localization (Chapter 7).	139

List of Acronyms

ADP	Auditory Distance Perception
ANOVA	Analysis Of Variance
BRIR	Binaural Room Impulse Response
CG	Coherence Gain (block)
CPU	Central Processing Unit
DM	Digital Modulation
DRR	Direct-to-Reverberant Ratio
DSP	Digital Signal Processing
EG	Envelope Gain (block)
ERs	Early Reflections
ERC	Early Reflections extraction and Cleaning
FFT	Fast Fourier Transform
FM	Frequency Modulation
GP	Coherence Gain (block)
GUI	Graphical User Interface
HA	Hearing Aid
HASQI	Hearing Aid Speech Quality Index
HATF	Head-Absent Transfer Function
HI	Hearing Impaired
HL	Hearing Loss
HPIR	Headphone-to-ear Impulse Response
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function

HSD	Honestly Significant Difference
IFFT	Inverse Fast Fourier Transform
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
JND	Just-Noticeable Difference
MMSE	Minimum Mean-Square Error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
MTF	Modulation Transfer Function
MWF	Multichannel Wiener Filter
NH	Normal-Hearing
PConv	Partitioned Convolution
PESQ	Perceptual Evaluation of Speech Quality
PNE	Partial Noise Estimation
PTA	Pure-Tone Average
RMM	Remote Microphone and ambient Microphone
RMSE	Root-Mean-Square Error
RSSID	Received Signal Strength Indication Difference
SegSNR	Segmental Signal-to-Noise Ratio
SDW	Speech Distortion Weighting
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
STD	Standard Deviation
STFT	Short Time Fourier Transform
VAD	Voice Activity Detector
WDRC	Wide Dynamic Range Compression

1 Introduction

Foreword

The term "binaural communication device" is used in this thesis in order to denote two types of devices for which most of the investigations and conclusion drawn along this dissertation can be applied : hearing aids (HAs) and hearables, which are briefly introduced in this section. A common use case of these two types of devices is the wireless remote microphone (RM) application, which is introduced in this section as well as in Chapter 2. This thesis mainly focuses on investigating how to improve the binaural spatialization of the sound rendered in the RM use case. In particular, several algorithms that aim at enhancing the realism of the binaural synthesis in this application were developed.

1.1 Context and motivation

1.1.1 Hearing aids

In 2021, over 5% of the world's population suffers from disabling hearing loss [131]. Among those 466 million people, 34 million are children. By 2050, it is estimated that this number should reach 700 million, which represents one in every ten people.

Hearing aids (HAs) are the most frequently used medical device to mitigate hearing loss. In some situations, such as noisy or reverberant environments, hearing-impaired (HI) subjects still suffer from speech intelligibility degradation, despite the use of HAs. This is particularly critical for tasks requiring efficient and effortless communication such as attending a classroom or a meeting at the workplace. Assisting listening devices can be used in combination with HAs to address this issue, in particular with RM systems. The principle usually consists in retrieving a "clean" version of the speech captured close to the speaker which is characterized by a high signal-to-noise (SNR) ratio. This signal provides a higher speech understanding but does not usually include most of the spatial information associated with the localization of the source. For this purpose, the human auditory system exploits differences between the signals

received in the two ears, called binaural cues. Binaural technology aims at simulating those cues to provide an immersive experience to the listener where sources are correctly located in space, and spatial acoustic properties of the environment are rendered. Additionally, a better spatialization of the speaker and acoustic scene can, in theory, contribute to speech intelligibility enhancement.

Recent works proposed a novel method to account for sound spatialization in assisting listening systems while taking into consideration the technical constraints of HAs [43, 44]. In [43], the performance in speech intelligibility was improved for the subjects presenting a moderate-to-severe hearing impairment with the spatialization feature. In the same study, in opposition to normal-hearing (NH) listeners, a majority of the HI listeners had similar performance with natural and artificial spatial hearing. These encouraging results suggest that the localization abilities of certain HI listeners are maintained in the frontal horizontal plane when resorting to a simplified spatialization.

Nevertheless, the current method induces several perceptive flaws commonly encountered with artificial spatial rendering such as front-back confusion and in-head localization, which is the perception of sound coming from within the head rather than externalized in the environment. Thus, it is of particular interest to investigate if a better perception of externalization can be rendered for HI listeners and what technical solutions can be practically implemented in regards of the current technical constraints of hearing devices. This thesis aims at developing new algorithms based on perceptual properties of the auditory system in order to provide a better externalization perception to HAs users, as well as a feeling of being immersed in a more realistic acoustic scene corresponding to the visual and physical reality. After development and implementation, the new features are tested through subjective listening tests in order to assess their perceptual effect.

1.1.2 Hearables

The word "hearables" was built as a neologism from the terms "wearable" and "headphones", and can be referenced as a sub category of ear-worn wearable devices [75]. These devices can be seen as miniaturized ear-worn computers. They were primarily engineered for the purposes of mobile communication and real-time information services in various contexts, but other uses have been developed notably regarding various monitoring applications related to the user's health and their body performance [66]. A wide variety of other uses can be found in the literature for this type of device [136].

Nevertheless, the present work only focuses on the transmission of audio information to the user. The challenge is to provide auditory spatialization for those sounds, with similar constraints as in HAs regarding the computational power, saving battery or using little memory storage.

1.2 Outline and contributions

This dissertation follows roughly the chronological order of the research that has been conducted. It is organized as follows.

Chapter 2 aims to provide the necessary information about the context of the RM for binaural communication devices, which is the context of the works proposed in this thesis. Some fundamentals of spatial hearing are also summarized in this chapter. The previous implementation of the binaural synthesis that served as a starting point for the proposed developments is briefly described, as well as the limitations which motivated the conduct of further works to improve the algorithm for perceptual purposes. Then, a description of auditory externalization is given, with some considerations about its precise definition. Finally, a review of the main cues involved in the perception of externalization and auditory distance is provided. Those cues motivated the works conducted in the next chapters.

Chapter 3 reports the work that was conducted with the implementation of DSP algorithms which aim at introducing early reflections in the binaural processing. The algorithm which served as starting point corresponds to an anechoic binaural spatialization, and the superimposition of early reflections to it is mainly motivated by the potential improvements in terms of auditory externalization which could be provided. The implemented methods are composed of a state-of-the-art method for sound dereverberation that was applied to the context of the RM in this work. Then an original approach as well as a method relying on coherence are also introduced for the same purpose. Finally, a different strategy is described, which consists in superimposing synthetic early reflections.

Chapter 4 concerns a first subjective listening test that was conducted with 10 NH and 20 HI listeners. The study aimed at testing several types of binaural rendering strategies and studying the perceptual effects of such algorithms on the perception of auditory distance in NH and HI listeners. Some of those strategies are based on methods described in Chapter 3, and mainly provide a certain amount of reverberation to the listeners. The goal was to verify that adding ERs to the binaural rendering of the RM system improves auditory distance perception in HI listeners. The study also provided insight about the perception of auditory distance in HI aided listeners.

Chapter 5 is devoted to a second listening experiment with 25 NH listeners. The framework of this chapter was binaural communication devices in the general sense, such as hearables. The goal of the study was to discuss the perception of auditory distance with low-cost algorithms that aim at providing a more realistic experience to the listener with a certain amount of externalization thanks to the introduction of ERs. The algorithms also take into account the problem of providing spatial awareness of other sound events of the environment, which was assessed during a second part of the experiment.

Chapter 6 reports the implementation of a head-tracking algorithm aiming at estimating the orientation of the head in the horizontal plane, i.e. the azimuth, relying solely on the

use of two 3-axis accelerometers. The main goal was to investigate the possibility to provide dynamic binaural synthesis with the technical constraints of HAs. Several perceptual considerations which motivates this work are described first. The baseline algorithm aiming at estimating rotational measurements with two accelerometers which served as a starting point is described. Then several improvements and upgrades to this algorithm are presented, which aimed to make the algorithm relevant to the tracking of head movements with the sensors potentially available on wearable devices. Finally, this algorithm was evaluated with several measurements on human subjects.

Chapter 7 concerns a third subjective listening test which aimed to evaluate the perceptual effects of head-tracking artefacts, such as those that can be encountered with constrained algorithms as described in Chapter 6. In particular, head movement coupled with head-tracking are known to potentially provide improvements in perceived externalization and localization accuracy. The listening test included the simulation of amplitude mismatch tracking and latency artefacts. The effects of such artefacts on both auditory externalization and performance in localization in the horizontal plane were evaluated.

2 Spatialization of the remote microphone and auditory externalization

This chapter first describes the context of the remote microphone (RM) system. Some fundamentals of spatial hearing are briefly described along with the implementation which served as a starting point for the work achieved in this thesis. The chapter then discusses the definition of auditory externalization and how to measure it. Finally, a review of the main cues involved in the perception of auditory externalization and auditory distance perception is made. Some of these cues justify in particular the investigation and development of the algorithms described in Chapter 3 and Chapter 6. Many of the cues listed here were used, or at least taken into account for the experimental design, in the perceptual experiments described in Chapter 4, Chapter 5 and Chapter 7.

2.1 The remote microphone system

This section introduces the context of the assistive listening device. Examples and applications are mainly described for the case of HAs. Nevertheless most of those considerations can be extended to binaural communication devices with a RM in general, including hearables. In applications such as a classroom or workplace meeting, a remote wireless microphone system for HAs (example illustrated in Fig. 2.2) can be used to provide a clean and intelligible speech signal to HI listeners [17]. The voice of the speaker is picked up close to the source by a body-worn microphone, as depicted in Fig. 2.1. A beamformer can be used on the body worn device, so as to pick up mostly the sound coming from the voice of the speaker with a high signal-to-noise ratio (SNR).

Currently, this signal is usually transmitted wirelessly through Digital Modulation (DM, formerly Frequency Modulation (FM)) to the HAs and rendered diotically, i.e. the same signal is streamed to both ears. This method allows for a large improvement in terms of speech

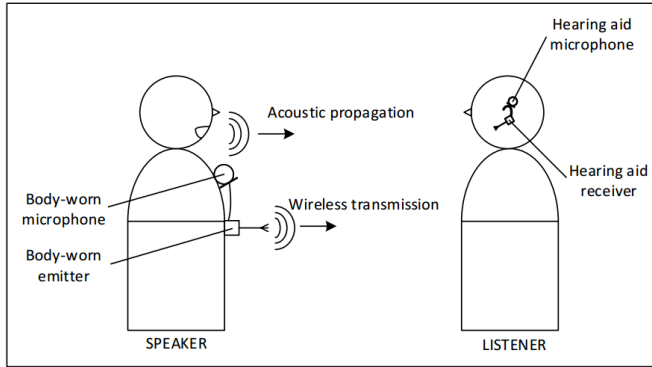


Figure 2.1: Wireless microphone system for hearing aids [44].



Figure 2.2: Example of hearing aid equipped with two omni-directional microphones: Phonak Audéo P.

intelligibility and reduced listening effort in noisy environments [172].

However, the resulting sound lacks auditory localization cues, especially if the RM signal is not mixed with the acoustic signal of the local HA microphone. Indeed, to localize a sound source, the auditory system relies on several cues, such as interaural differences in time and level, and the spectral filtering caused by the shape of the head, pinna and torso. Those are mainly included in the acoustic transfer functions from the position of the sound source to the eardrum of the listener. The next section aims at summarizing briefly the fundamentals of spatial hearing necessary for the global comprehension of this thesis.

2.2 Fundamentals of spatial hearing

Locating a sound source in space, here expressed in a spherical coordinate system, consists in determining its position regarding three dimensions: azimuth, elevation and distance, as depicted in Fig. 2.3.

For those three dimensions, the auditory system relies on various cues which are described in this section for the azimuth and elevation, and later in this chapter for distance.

2.2.1 Localization in azimuth

The localization in azimuth, is mainly affected by binaural cues, in particular interaural time differences (ITDs) and interaural level differences (ILDs). This was first theorized and developed in 1907 by Lord Rayleigh [110]. In particular, the "duplex theory" of sound localization suggested that ITDs and ILDs had a frequency-dependent influence. For a given listener position and sound source position, the ITD is the difference in sound propagation time between the positions of the two ears. A common approximation, assuming the sound to be a planar wave (source at infinite distance), consists in computing the ITD with Woodworth's

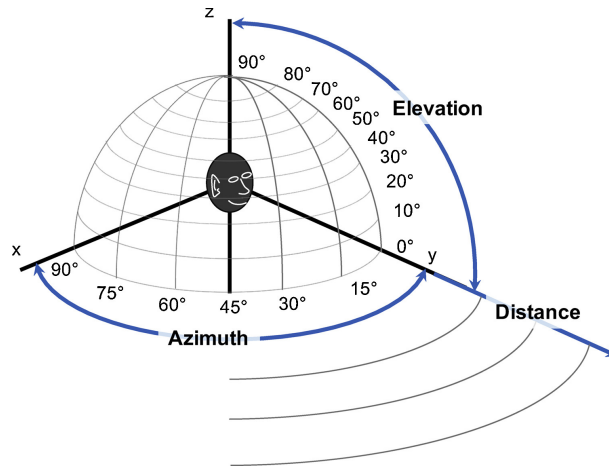


Figure 2.3: Representation of the dimensions used to locate a sound source in spherical coordinates: azimuth, elevation and distance (taken from [140]).

formula [173] which can be derived from the geometric representation depicted in Fig. 2.4:

$$ITD = \frac{r(\theta + \sin\theta)}{c}, \quad (2.1)$$

where c is the speed of sound. The ITD is primarily useful to the auditory system for the localization of sound sources with substantial energy below 1.5 kHz [110, 123]. Mills found in [123] that the minimum threshold of detection of ITD could be as low as $10 \mu s$.

The ILD is mainly caused by the head shadow effect [158]. Indeed, by reflection and absorption, the head partially reduces the acoustic energy for wavelengths that are shorter than the approximated head diameter. Thus, it is often considered that the ILD has no effect below 1.5 kHz [110, 123]. The minimum threshold of detection in ILD was shown to be about 0.5 dB in [123].

2.2.2 Localization in elevation

Localization in elevation is mainly affected by monaural spectral information [15]. This spectral information is generated by reflections, absorption and diffraction of the incident sound waves by the shapes of the pinna, head, shoulders and torso. The pinna in particular was shown to play an important role in the human localization, with the "pinna effect" [10]. For example, in [29], listeners were able to determine the vertical position of high frequency sounds (above 3 kHz) with only monaural information. This was not the case for lower frequency sounds, which might be explained by the small dimensions of the pinna, which makes it capable to affect only sounds with short wavelength.

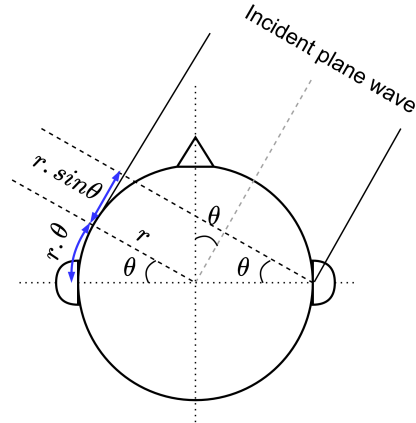


Figure 2.4: Illustration of the geometric difference in distance used to compute the ITD with Woodworth's formula.

The perception of distance will be discussed further in this section. This thesis focused on the localization in the horizontal plane, i.e. azimuth and distance, which are the main dimensions of interest in the context of RMs for wearable binaural devices, as in most use cases, the speakers should be located at the same height than the listener.

2.2.3 Head-related transfer functions

Head-related transfer functions (HRTFs) [169] are the filters which encode all the useful information for sound localization cited above, including ITDs, ILDs and the spectral information caused by the shape of the pinna, head, shoulders and torso. Their time-domain representations are referred to as head-related impulse responses (HRIRs).

All these cues are dependent on the various related anthropometric measures of the listener, thus each individual is used to locate sound with their own HRTFs. Measuring individual HRTFs for a listener is a long procedure which requires some specific equipment such as anechoic rooms and arrays of loudspeakers. Because of this, binaural spatial synthesis is often performed using generic HRTFs measured with manikins equipped with artificial ears, that are designed to represent average anatomical dimensions of humans. In [167], a similar performance in localization accuracy could be obtained with non-individual HRTFs compared to free-field acoustic sources, at the exclusion of front-back confusions. However, in [124], Minaar et al. compared real-life listening, real head recordings and artificial head recordings. They found that localization accuracy in real life was better than with recordings, and that real head recordings were better than with artificial heads. The adaptation to altered HRTFs takes a long time (several days) as suggested in [116]. Mendonça et al. [118] found that with appropriate active training, a learning effect with generic HRTFs could be obtained to yield a large improvement in localization performance.

2.2.4 Localization in a room

The sound reflections occurring in rooms are also an important element for spatial auditory perception. Including reflections in the simulation of virtual spatial rendering allows to recreate the spatial auditory perception of a sound in a room and can help to perceive the distance of the virtual source [181] (this will be discussed in more detail in Section 2.6). The corresponding impulse response at the two ears, called the binaural room impulse response (BRIR) can be obtained directly from measurement in the room or by simulation using geometrical acoustics. The auralized sound can be obtained by convolving an anechoic signal with the BRIRs of both ears. A complete room impulse response is composed of three elements: the direct sound, early reflections (ERs) and late reverberation, as depicted in Fig. 2.5.

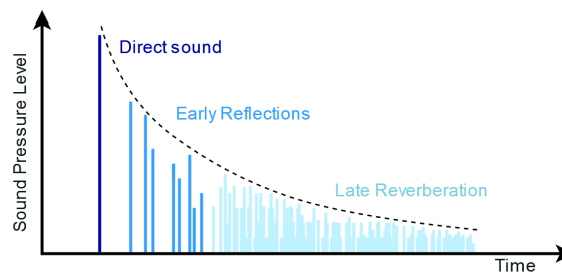


Figure 2.5: Components of a room impulse response.

2.2.5 Dynamic binaural synthesis

The cues described so far and included in the BRIRs allow one to synthesize a convincing stable virtual sound source for a listener standing still. However, when the listener rotates their head, the simulated location of the sound source becomes erroneous if the filtering is not updated. The use of a head-tracker allows one to render a dynamic binaural synthesis of the sound sources. This provides several advantages regarding perceptual considerations.

Front-back confusion

Front-back confusion, first mentioned by Wallach [162], is a common issue that can happen in binaural sound synthesis. Indeed, a sound source and its mirror image as pictured in Fig. 2.6, are associated with similar binaural cues: ILDs and ITDs. Wallach [162] made the hypothesis that head-movement should enable the elimination of most front-back confusions, also called reversals, which has been confirmed in several studies [155, 128, 171]. Indeed, head motion can resolve the so called cone of confusion [173], owing to the differential integration of interaural cues over time [15].

In [115], the necessary range of head movement that helps improving azimuth localization accuracy was investigated. They found that head rotation with an azimuth window of only 4°

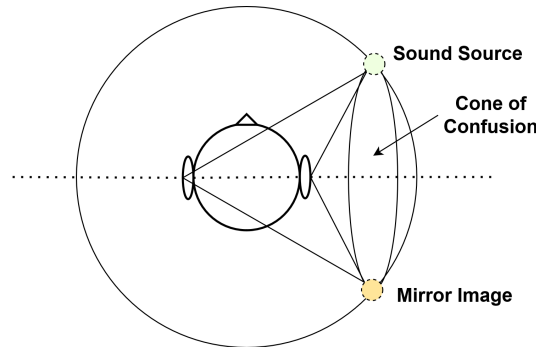


Figure 2.6: Illustration of the cone of confusion, with a sound source and its mirror image.

could already significantly reduce the rate of reversals.

Plausibility

With the use of a head-tracker, the angular position of the head can be retrieved, and the binaural rendering can be adapted accordingly by selecting the correct pair of HRTFs. Hence, the position of the sound sources remains stable in the perception of the rendering regardless of the listener's head position. Without head-tracking, the sound source would appear to be moving by the same azimuth as the listener's head, which would make the plausibility of the rendering collapse, as schematized in Fig. 2.7.

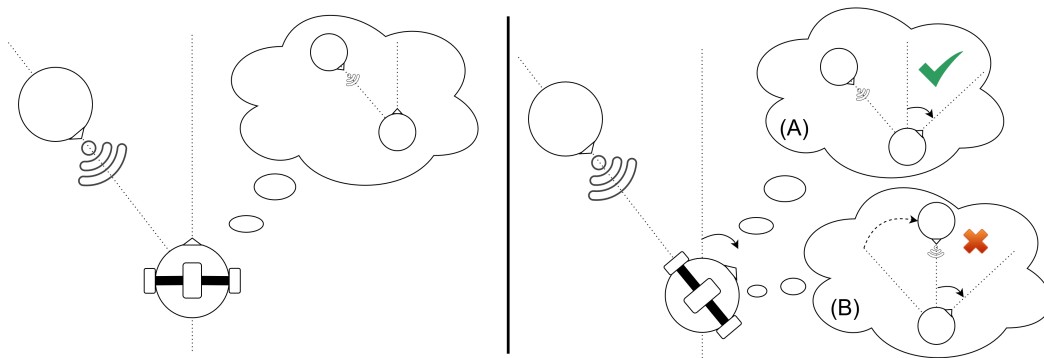


Figure 2.7: Initial position (left panel) and position after head movement (right panel), Without head-tracking it is not possible to provide the synthesis of a source that appears at the correct position regardless of the head orientation (B), with head-tracking, the filtering can be adapted to the head orientation (A).

2.3 Spatialization of the remote microphone

There are several motivations to include spatial hearing in RM systems. Binaural hearing contributes to improved speech intelligibility in challenging situations such as noisy conditions

or complex scenes [32]. An intelligibility increase can be observed when a spatial separation is introduced between the target speech and the sound maskers [48, 59], which is referred to as the spatial release from masking. Moreover, being able to localize the sound source when the speaker is not initially in the visual field should help the listener find the speaker more quickly to access lip reading, which is crucial for HI listeners. A real-time binaural localization and tracking algorithm for HAs was introduced in [38, 44]. A typical use-case is depicted in Fig. 2.8. In the depicted situation, two speakers are equipped with remote microphones, and one listener is equipped with a wearable binaural communication device (HAs or hearables) with receivers.

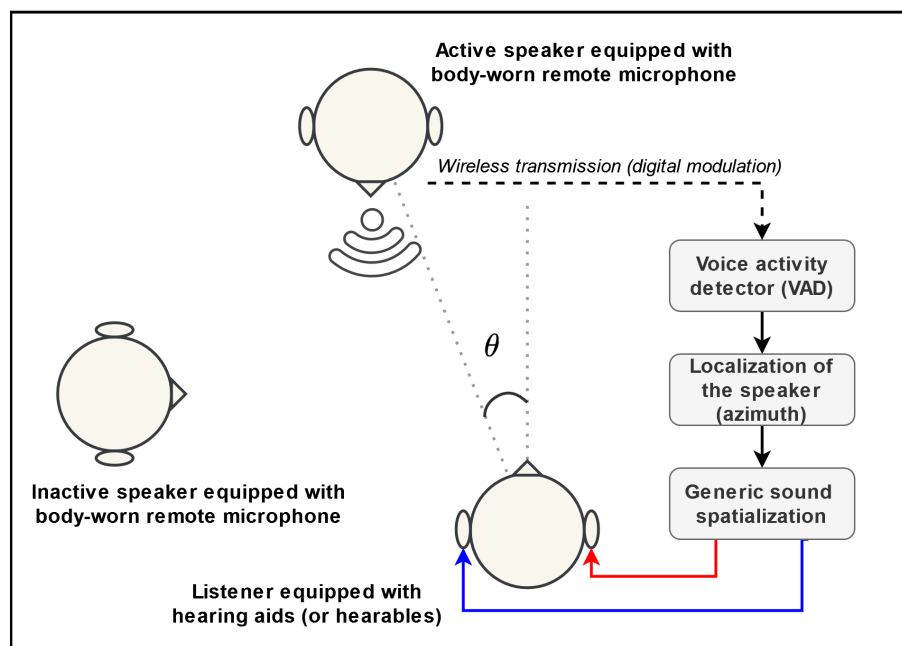


Figure 2.8: Schematic representation of the principle of the localization and spatialization algorithm for remote microphone systems (applicable to HAs or hearables).

The first step consists in using a voice activity detector (VAD), which enables to detect which speaker is currently talking. Then, the position relative to the listener is estimated with an algorithm relying on the weighted combination of three cues. Two of those cues are related to the acoustic signal received by the HAs' microphones: the interaural phase difference (IPD) and the ILD. The third cue is related to the wireless transmission with the Received Signal Strength Indication Difference (RSSID). In [44], the spatial resolution was limited to five zones in the $\pm 90^\circ$ area. This is justified as the listener might have access to the visual cue most of the time in the typical use-case (attending a class, meeting at work), and because HI listeners have lower performance for the localization in azimuth [53]. However, it is imaginable to use finer discretizations in other contexts.

Finally, the clean speech signal can be spatialized according to the estimated direction, using

non-individualized HRTFs. The frame size used in the HA considered is of 128 samples (5.8 ms at 22.05 kHz) for the real-time processing, which encourages to use 5.8-ms impulse responses. This corresponds to a truncated version of a binaural impulse response that resembles an anechoic impulse response, containing most of the time only the direct sound with its binaural cues, and is mostly similar to the HRTF. The algorithm takes advantage of the fact that HRTFs can be modeled as minimum-phase functions, meaning that the excess phase can be considered almost linear and thus can be approximated with a pure delay, hence,

$$HRTF \approx |HRTF| e^{j\phi_{min}} e^{j\omega\tau} \quad (2.2)$$

with ϕ_{min} the minimum phase and τ a pure delay. This allows splitting the binaural synthesis processing in two steps. First, the ILD and spectral cues are introduced at the stage of filtering of the spectrum of the incoming signal by the magnitude spectrum of the HRTF. Second, a reproduction of the ITD is achieved by delaying one of the left or right channel by a pure delay.

The mentioned feature is intended to increase comfort, the feeling of immersion inside the scene and, presumably, the intelligibility for the aided users. In [43], speaker localization in azimuth and speech intelligibility experiments were conducted with 40 participants divided in four groups of 10 subjects each, with either normal-hearing, moderate, severe, and profound hearing loss (HL). Binaural and anechoic speech stimuli were used for both tasks. The results showed that most HAs users had similar performances with natural and artificial spatial hearing, suggesting that HI subjects keep their localization abilities in the frontal plane with simple generic HRTFs. Additionally, for HI subjects with moderate and severe HL, a better intelligibility performance was observed for stimuli spatialized with the above described algorithm compared to the diotic presentation. For the listeners with profound HL, the results suggest that the spatialization feature described in this section should not degrade intelligibility. The algorithm was implemented with a sampling frequency of 22.05 kHz, which is a typical rate in HA.

2.4 Perceptual limitations of the baseline binaural synthesis

In the baseline spatialization algorithm described above, several issues commonly appearing with non individualized HRTFs were reported, in particular front and back confusions and the perception of sound coming from within the head (internalization, also sometimes referred as lateralization) [156]. In a survey related to the use of remote systems in classrooms, users reported a preference for the remote signal over the HA microphone signal with respect to intelligibility. However, the users did not always choose the remote system due to its poor performance in externalization, spatial awareness, and sound quality, despite an easier listening effort [129]. This calls for the need to improve the audio signal processing algorithm to retrieve a better perception of the acoustic scene when the speaker's voice seems to come from outside the head, and the general rendering is perceived as more natural and matching the visual cues.

Moreover the above mentioned binaural synthesis did not include head-tracking to achieve dynamic binaural synthesis. This means that plausibility might collapse in certain situations, such as when the listener turns their head in the horizontal plane with a certain azimuth, the sound source will be perceived as turning by the same azimuth if the binaural synthesis is not adapted depending on the listener's head movements. This motivates to investigate the possibility to include head-tracking to achieve dynamic binaural rendering. Head movements are also known to resolve most of the front-back confusions [162], which is another motivating reason to include head-tracking in spatialization algorithms for RM applications.

In Section 2.3, the baseline RM spatialization method used as a starting point for this work has been introduced. As mentioned, a limitation from this method is the perception of sounds inside the head (i.e. internalized) rather than outside the head (i.e. externalized), meaning that the sense of distance is lost. Section 2.5.2 aims at reviewing the main cues involved in the perception of externalization and auditory distance.

2.5 Auditory externalization

2.5.1 Definition

The notion of auditory externalization relates to the perception of sound events as coming from locations that are outside of the head. This is the case for most sounds from the physical world, when they are directly perceived by the auditory system for NH listeners. The presentation of diotic sounds over headphones, was shown to lead to the perception of sound images that were inside rather than outside the head [77]. This is a common issue when listening to binaural audio over headphones, and is illustrated in Fig. 2.9. Including the spatial cues of the filtering caused by the head, pinna and torso (as contained in the HRTFs) as well as reflections of the environment, allows to render a convincing externalized acoustic scene over binaural headphone reproduction [71, 13]. The first review of the cues affecting auditory externalization was presented in [52] and a more recent review can be found in [14].

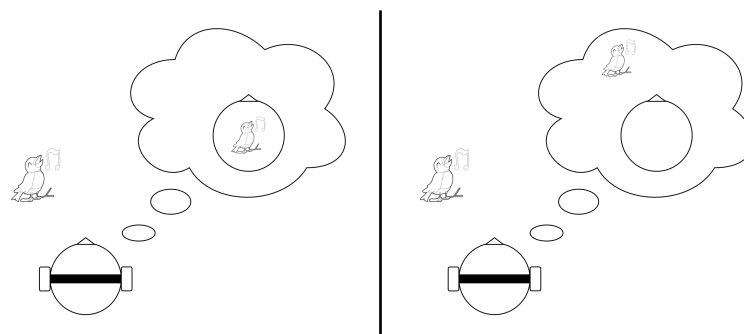


Figure 2.9: In-head localization (left panel) vs. externalization (right panel) of a sound source.

2.5.2 Relation to auditory distance: a continuum or two different dimensions ?

In the attempts to precisely define and study auditory externalization, an interesting discussion concerns its relation to auditory distance perception [14]. The most common approach consists in seeing the link between auditory distance and externalization as a continuum. Indeed, it is possible to imagine that the center of the head could be the minimal distance that is possible to perceive, and that by following a continuous line, the perception can be shifted continuously to the border of the head and finally to sound sources that are outside the head. In one of the experiments described in [71], a source could be moved continuously from inside the head to outside the head, which the authors interpreted as evidence for a continuum between internalized and outside-of-the-head sound sources. In [52], it is suggested that externalization can be interpreted continuously from outside to inside the head and through the border of the head, as a matter of degree of externalization.

However, a second approach consists in considering that auditory distance and externalization actually belong to two distinct dimensions. The main argument for this paradigm is that they are known to be mainly influenced by different types of cues [14]. Distance perception is mainly affected by monaural cues, such as sound level or direct-to-reverberant ratio (DRR) [91, 178]. On the other end, externalization perception might be more influenced by binaural cues [18]. Indeed, when the amount of binaural cue information is reduced in the presentation of sounds over headphones, the perception of externalization is decreased [71, 33]. For example, it is possible to imagine a diotic sound presented over headphones with a large amount of reverberation and a low DRR. It is likely that such sound would give the impression to be somehow distant because of the low DRR, while still sounding internalized as no binaural cues necessary for externalization would be available.

So far, no clear statement can be made to settle if auditory externalization and auditory distance do share the same continuum or if they should be approached only as two different dimensions. Many overlapping cues are involved in their respective perception. The next section is dedicated to describing the cues affecting auditory distance and externalization perception. Certain of these cues motivated the signal processing developments described in Chapter 3 and 6, and most of these were considered in the design of the experiments described in Chapter 4, 5 and 7.

2.5.3 Methods for measuring externalization and auditory distance

As a result of the above mentioned considerations, evaluating the perception of externalization can be achieved with various methods, described exhaustively in [14]. Even though free-field acoustic signals can be perceived as internalized [20], it is most common to focus on the evaluation of externalization with sounds presented over headphones. Those sounds are usually processed with binaural rendering methods, whether anechoic with HRTFs, with room information included in BRIRs, or hybrid strategies.

To assess the perceived externalization, it is common to resort to discrete categorical scales with several levels to which are associated some reference elements. For example in [71], the following terms are used on a 4-levels categorical scale: "(0) The source is in my head; (1) The source is not well externalized. It is at my ear, or on my skull, or very diffuse; (2) The source is externalized but it is diffuse or else at the wrong place; (3) The source is externalized, compact, and located in the right direction and at the right distance". In this case, a loudspeaker was available as a visual reference. Other studies have used comparable types of scales with various terminologies, mostly related to distance [33, 74]. This type of scale can also be represented graphically on a sheet given to the participant, as done in [63, 87]. It is also possible to resort to continuous scales. In this case, listeners are usually asked to rate the distance on a scale which continuously represents the possible locations in space along a line starting at the center of the head, through the border of the head, to references that are outside the head. This can be done with a graphic representation available on a touchscreen on which the listener can give the rating directly [117]. Scales in inches have also been used in [13]. With such scales and labels and the availability of visual references to match in the room, it can be argued that the simulation not only has to produce a sound perceived outside the head, but also at the correct distance of the location of the visual reference. Procedures with eyes closed such as in [97] aim at mitigating this effect. The continuous scale can be provided with only two minimum and maximum labels such as "completely internalized" associated with the 0% rating and "completely externalized" associated with the 100% rating, as done in [97, 18] and in the experiment described in Chapter 7. It is also possible to assess the perceived externalization by asking a binary question if the sound source is perceived inside or outside the head, as done in [20] for example. Such a method gives indication about the likelihood of a stimulus to be perceived as externalized, but differences between e.g. stimuli that are always barely externalized vs. stimuli that are always neatly externalized might be missed.

In the case of experiments focusing on auditory distance, continuous scales can be provided with several labels [41]. Those labels are usually associated with visual reference cues in the environment such as a loudspeaker in the room, or parts of the head for internalized sound sources, such as e.g. "border of the head" or "center of the head". A last approach consists in providing a sheet with a sketch representing the head of the listener and their environment, and let them draw the perceived location of the sound source [73, 46]. These methods might also be affected by vision and effects such as "visual capture" (this is discussed further in Section 2.6.6).

2.6 Externalization and distance perception for normal-hearing subjects

As seen in Section 2.5.2, auditory distance perception and externalization are closely related. Thus, the cues involved in their perception overlap, even though some cues might have more influence on one attribute compared to the other. Consequently, this section focuses on the most relevant cues in the scope of this thesis dissertation, and the effects on both attributes

are presented in a mixed manner. To go further, an exhaustive review of cues affecting auditory distance perception can be consulted in [178, 91], while a recent review of cues, focusing on auditory externalization can be consulted in [14].

2.6.1 Role of reflections

Reflections have a crucial role in the perception of externalization [71] as well as perceived distance of the source [122, 23, 178]. The direct-to-reverberant energy ratio (DRR) corresponds to the ratio of energy reaching the listener's eardrum directly to the energy reaching the listener after reflections on the walls and various surfaces in the room or environment. It is defined as:

$$DRR = 10 \log_{10} \left(\frac{\int_0^{t_D} h(t) dt}{\int_{t_D}^{t_R} h(t) dt} \right), \quad (2.3)$$

where t_D is the time corresponding to the end of the direct sound and start of the ERs, and t_R corresponds to the full length of the BRIR. While the contribution of the direct sound will be dependent on the source distance, the reverberant contribution in the DRR is mainly determined by the dimensions of the room and the acoustic properties of the surfaces. The DRR is an important cue that explains how the auditory system estimates sound source distance [23, 92]. The DRR decreases as the listener moves farther from the source. Indeed, the direct sound decreases (by 6 dB for a doubling of the distance) while the energy of the reverberant sound caused by the reflections on walls and objects remains almost constant with position [174].

The length of the impulse response (i.e. the number of reflections) affects the perceived externalization [45], especially for higher order reflections (longer than 25 ms) [159] and with a larger influence up to approximately 30-40 ms [149]. Generally, reflections occurring after around 80 ms are not considered useful for externalization. The amount of reverberation due to the environment is also likely to affect distance perception, as it was reported that distance estimations were larger in rooms with longer reverberation times compared to room with shorter reverberation time [120]. DRR is mainly useful in enclosed environments and determined by the shape of the room as well as the absorption coefficient of the various walls, surfaces and objects in the environment. DRR has been shown to be an absolute cue in the evaluation of auditory distance [122].

A recent study investigated the possibility of generating externalized virtual sources using different BRIR modification techniques while preserving sound quality, i.e. in this study, the similarity to the original anechoic recording [64]. For this purpose, the authors modified several parameters such as the impulse response length (truncation), the reverberation time (using a decay envelope), or the DRR, and investigated the resulting effect on externalization and attributes of sound quality. They found that while externalization could be increased as

expected with an increased amount of reverberation for all methods, a trade-off has to be found to optimize sound quality and each method has their advantage. They suggest that truncation, although being the best option to yield short BRIRs has no physical equivalent and may thus sound unnatural. DRR modifications can lead to timbral artefact which might affect naturalness extensively. Finally they suggest that a compromise can be found by playing on the reverberation time with a decaying gain, which enables to reduce the BRIRs length as well in a more naturally sounding way.

However, reverberation cues alone might not be sufficient to provide a convincing perception of externalization [18].

2.6.2 Role of binaural cues

The resort to non-individualized HRTFs can reduce the degree of perceived externalization [52, 87]. Leclère et al. [97] found that individualized BRIRs did not yield a better perception of externalization compared to non-individualized BRIRs when presenting noise stimuli processed with BRIRs from three different rooms. The study did not include a visual source to match, which could explain the different conclusion compared to the studies cited above (the influence of vision on such an experiment is discussed further in Section 2.6.6). The ILD is an important cue related to distance perception and has substantial variations in the near-field [25], thus it provides accurate distance information up to 1 m [174]. The natural ILDs fluctuations were altered in [33], resulting in a significant reduction of perceived externalization. On the other end, the ITD can be considered to be fairly constant with distance and does not provide a useful cue for distance estimation.

Auditory distance perception is more accurate for sources located laterally relative to the listener in near-field [92] (this is discussed further in Section 2.6.5). Moreover, for nearby sources, lower frequency ILD cues (<3 kHz) for distance estimation are considered robust relatively to room reverberation, i.e. when DRR cues are available in addition to ILD cues [146]. For a source located laterally, it was found that monaural reverberation cues are sufficient to provide externalization, whereas for a frontal source binaural cues of the reflections are needed [34]. In the same work, it was reported that the interactions of interaural cues of reverberation and the direct sound also affects externalization perception. The auditory system have been shown to suppress the localization information of reflections arriving shortly after the direct sound [105]. This phenomenon is referred to as the precedence effect. Thus, it is suggested in [34] that the precedence effect might be involved in the auditory processing of the dynamic binaural cues that are used for externalization perception. Li et al. found in [100] that externalization of a lateral sound source is more affected by the reverberation at the contralateral ear has compared to that at the ipsilateral ear.

Interaural coherence is a measure of the similarity of the signals at the two ears. It was showed in [22] that artificial changes in interaural coherence could affect the perception of auditory distance. A recent study [138], suggests that naive listeners might not use interaural coherence

as a cue used for auditory distance estimation of a virtual frontal source. This was observed both with and without visual cues. However, the auditory externalization over headphones was degraded when binaural information of the sound was completely removed. The sound level was the main cue used by naive listeners to estimate auditory distance. They could, in some cases, rely on reverberation-related changes in spectral content. The distance estimations were not affected by the use of non-individualized spatialization instead of individualized spatialization.

2.6.3 Level cues

Sound pressure level is another important cue related to distance estimation [7, 174]. An increasing source distance is associated to a decreasing sound level at the eardrum of the listener following the inverse-square law (for a spherical sound wave) in an anechoic environment:

$$dL = Lp_1 - Lp_2 = 10 \log\left(\frac{D_1}{D_2}\right)^2 = 20 \log\left(\frac{D_1}{D_2}\right), \quad (2.4)$$

where dL is the difference in sound pressure level, Lp_k the sound pressure level at location k and D_k the distance from the source to location k . Hence, the level decreases by 6 dB when the distance doubles (i.e. if $D_2 = 2D_1$). This reduction rate is reduced in a reverberant environment with reflections [174]. Contrarily to DRR, sound level is a relative cue and requires an *a priori* knowledge of the expected loudness of the source, which is the case for most daily life sounds such as the human voice. This effect is referred to as the "stimulus familiarity". Level combined with DRR are an efficient combination of cues for distance estimation [181, 92].

2.6.4 Spectral cues

The importance of pinna cues in the perception of externalization was suggested by Plenge in [137], which he referred as "outside-head localization". The role of the spectral detail included in the BRIR for externalization was studied in [72]. The results suggest that a reduction of spectral detail in the direct sound affects externalization while reduction of spectral detail of the reverberant part had little influence on externalization. In [95], HRTFs were significantly smoothed in frequency, and the results suggested that the fine spectral details might not be critical for perceiving the auditory externalization with broadband noises. In [102], it was shown that for more lateral sound sources, the magnitude spectral detail in the direct sound at the ipsilateral ear became increasingly important for auditory externalization in comparison to the contralateral ear. In [18], externalization was evaluated with both in-the-ear and behind-the-ear positioning of the HA microphone. Externalization was improved with in-the-ear positioning, suggesting that pinna-induced spectral cues are important for externalization.

Spectral cues are also a useful cue for distance perception of nearby sources due to the

diffraction caused by the head shape which varies with distance and frequencies [27], and is different for each individual. For nearby sources the effect of spectrum was found to be stronger for frontal sources and more accurate for sounds containing more energy in low frequencies [92]. For distances larger than 15m, due to the air absorption, sounds coming from farther away will be characterized by high frequencies being more attenuated than lower frequencies. Therefore, sounds presenting more attenuation of high frequencies will tend to be perceived farther away [30, 106].

2.6.5 Laterality

Several studies demonstrated that there is a strong dependency of the laterality of the source on the perception of externalization. More lateral sources were shown to be perceived as more externalized than frontal sources in several studies. This was assessed with binary judgement in [20] with anechoic conditions, the availability of head-tracking and small movement. This was also shown in [81] for both anechoic and reverberant conditions, and using a continuous percentage scale, while listeners had access to visual cues. In [97], the same effect was also observed using a continuous scale, for both anechoic and reverberant conditions when the listeners had their eyes closed. It can be expected that a lateral source (e.g. at 90°) may be perceived, at worst, near the position of the ipsilateral ear [14]. For example, a monaural anechoic sound presented over headphones and fully panned to one side may be perceived as lateralized, but already near the position of the corresponding ear. Conversely, a frontal source (close to 0°) can potentially be perceived at the center of the head. Consequently, a scale which starts at the center of the head could yield better externalization ratings for more lateral sources. This could create a bias in externalization experiments which would not differentiate between azimuths.

The azimuth of the sound source also has an influence on the perceived auditory distance. Brungart et al. [25] found that distance evaluations for lateral sounds were more accurate than for frontal sounds. The authors suggested that the variations in ILD with azimuth and distance provides a useful cue for the distance estimation of nearby sources (< 1 m). Indeed, for lateral nearby sources, the ILD increases as the source arises. Conversely, for further sources, the ILD is only affected by direction. Consequently, the range of possible ILD values is larger when the source is more lateral, while it tends to zero (at 0°) for sources in the median plane.

2.6.6 Visual cues

Vision is known to be the predominant sense used by human beings in many situations, and visual distance estimation is also known to be more accurate than estimation relying on auditory cues [147, 109]. Visual information was shown to play an important role for distance perception [31, 177, 47]. In particular, visual cues and the combination of auditory and visual cues have been shown to yield more accurate distance estimations compared to auditory cues alone [4].

However, vision could be predominant compared to audition in certain situations. This is referred to as visual capture, and was first suggested to influence the perception of auditory distance by Gardner [61]. In particular, he found a tendency in listeners to perceive sound at the nearest visible potential sound source instead of the actual sound source position emitting the sound. He referred to this phenomenon as the "proximity-image effect". This hypothesis was challenged by Mershon et al. in [121], where the distance of sound sources were not only underestimated, but a hidden closer sound source could be perceived at the position of a further dummy loudspeaker. Another famous example of visual capture is the well-known ventriloquism effect [163]. It consists in the source being perceived as coming from a potential visual location of the source, rather than the actual localization where it is emitted. For example, this could happen when watching a speaker on a screen while listening via headphones. This effect was shown to have an impact on distance perception, and it is suggested in [31] that this effect could be persistent as listener keep in memory a visual representation of the environment that help them estimate auditory distance afterwards with sound only.

The effect on localization in azimuth can be explained by the so-called "localization blur", which is greater for audition than for vision. Moreover, the mismatch between the visual impression of the room and the reflections (e.g. absence of reflections or reflections corresponding to another environment) prevents from a correct estimation of the distance and externalization [63], which is the case in the situation described in the previous section with the use of generic HRTFs which do not include the room reflections.

In [97], listeners had to close their eyes to evaluate the degree of perceived externalization under different binaural conditions. In particular, they found that non-individualized BRIRs and individualized BRIRs yielded comparable perceived degree of externalization when the listener did not have access to a visual reference to match. The rationale behind this strategy was to remove the constraint for the listener to have to visually match the perceived location of the sound source. It can be suggested that this enabled to disentangle externalization from auditory distance perception.

2.6.7 Head movements and head-tracking

Head movements can also provide useful information likely to improve externalization. In [20], it was found that, in the case of individualized HRTFs, head movement without head-tracking substantially reduced the perceived externalization [20]. Indeed, the use of a head-tracker allows monitoring head angular position and select the suitable HRTF so as to simulate a plausible situation where the sound source remains at the same position as the listener rotates their head. In particular, large head-movements coupled with head-tracking were shown to improve externalization in non-individualized binaural synthesis, for frontal and rear sources in particular [74]. It was also shown to provide at least a better feeling of "spaciousness" [54]. Simulating a corresponding movement of the source in the azimuthal plane while the

head remained fixed (called pseudo-tracking) provided the same benefits in those studies for large movements, as well as in a recent work exploring a variety of source trajectories [98]. A common finding is that small movements of the head or source provide small or even no improvement regarding the perception of externalization.

A more detailed review of the literature concerning the contribution of head-tracking and head movement can be found in Section 7.1.2.

2.6.8 Room divergence effect

Room congruence, i.e. the coincidence of the acoustic properties of the room used for playback and the BRIRs used for the virtual sound synthesis, has been shown to have a significant effect on externalization. In [170], Werner et al. found that a divergence between the binaurally rendered room and the listening room lead to a decrease of perceived externalization whereas congruence between the rooms yielded an increase. A recent study investigated if a similar observation could be made in the case of virtual reality (VR) applications in which listeners would get a visual impression of the virtual room, but might not be aware of the auditory information of this room [103]. They found that non-individual BRIRs need to match listeners' expectations based on the visual impression of the room. Nevertheless, when listening in an unusual virtual room with atypical acoustic properties for a naive listener, in this case an anechoic room, the correct room information degraded the perceived externalization. More rooms would have to be tested to verify that this is not just the particular case of an anechoic room.

In [151], subjects were asked to adjust the DRR to adapt the binaural playback of a synthesized scene to their expected DRR of the listening room, which they managed to do reliably without explicit external reference. Nevertheless, only a slight (non-significant tendency) externalization improvement was found when the divergent room condition was DRR-adapted to the listening room in comparison to the original divergent room condition.

It can be noted that in general, humans tend to be less accurate in auditory distance estimation compared to their accuracy abilities in angular localization [178]. Moreover, the perceived auditory distance is often overestimated for close sources and underestimated for faraway sources [174].

2.7 Auditory externalization and distance perception in HI listeners and HAs users

While many studies cover the effect of HL on the ability to localize sounds in the azimuthal plane, the literature only contains a few studies related to how HL affects distance perception and externalization. In [2], the ability of elderly HI subjects to discriminate differences in the cues to the distance of speech stimuli was assessed. The elderly HI listeners performed

similarly as NH listeners when both DRR and level information was available. Nevertheless, their performance was degraded compared to NH listeners when the level cue was fixed and only the DRR cue was available, suggesting difficulties to resort to DRR to evaluate distance for this population.

Amplitude compression is an important feature in HAs necessarily applied to face the recruitment of loudness phenomenon [58]. The goal is to apply high gain for low sound levels so that they become audible for the HI listener, and lower gains to high sound levels so as to not reach excessive and potentially harmful sound pressure level (SPL). It is frequency-dependent and non-linear. Thus, it could be expected that compression might alter the level cue and information available from the DRR for a typical experimental localization task. Nevertheless, in real life situations, and for continuous speech stimuli, reverberation is rarely separated from direct sound in time (only reverb tails for end of sentence or pauses), thus compression might not have much of a detrimental effect. In [3], experienced HA users were performing a distance discrimination with continuous speech stimuli and both level and DRR cues available. Compression did not have an effect on the performance which can be explained by the accustomization of subjects who are used to their own HAs. It is also worth mentioning that only small compression was applied in this study compared to conventional compression that are used for severe or profound HI subjects. The task only consisted in a relative distance judgment and an absolute judgment would be worth investigating. In [148], the authors underlined the importance of the transparency of the method of amplification regarding the distortion of localization auditory cues, notably ILD cues used for distance estimation. Fast compression (short attack time) was revealed to increase the just-noticeable differences (JND) in ILD compared to linear amplification, in a localization task including both NH and HI subjects [126].

In [18], NH and HI subjects rated the degree of externalization of speech stimuli, made of a mix of internalized and externalized speech sequences presented over headphones. The main conclusion was that HI listeners experienced a contracted perception of externalization as stimuli were never as externalized or internalized as for NH listeners. The observation was correlated with high-frequency HL.

A method relying on a so-called "structural binaural model" [24], consists in combining separate filters representing all the components of the HRIR (head, pinna, ear canal) to generate an overall HRIR. This approach allows including anthropometric measures of the listener to customize the various elements of the HRIR accordingly. Those individual adjustments were shown to provide an improved perception of externalization of speech stimuli for NH subjects, in particular when associated with reverberation [88]. Nevertheless the method was not evaluated with HI listeners in the latter. Recently, a novel algorithm aiming at improving externalization in HAs without the resort to individual HRIRs was introduced in [79, 81] and relies on the model introduced in [24]. The study evaluated the effect of various procedures for speech externalization presented through earphones to ten NH subjects and ten moderate HI subjects. As hypothesized, the externalization degree with individual anechoic HRIRs

was enhanced over simple stereo panning. The addition of room reverberation produced a significant improvement in externalization compared to the anechoic HRIRs. Interestingly, the structural binaural model with simulated reverberation resulted in similar performances as the use of the listener's own HRIRs. Compared to the method using the average adult values for tuning the various filters, a best fit of the 125 ratings provided by the listeners gave a slight but significant increase in externalization. Conversely, a selection of the model based on anthropometric measurements did not produce any significant benefit over the use of the average adult values. There was no significant differences between NH and HI subjects. However, the HI subjects were only moderately impaired, with the criteria to have 22 dB HL averages or greater for four-frequency (500, 1000, 2000, 4000 Hz) pure-tones.

A recent study demonstrated that the superimposition of ERs to the RM signal can significantly improve externalization with artificially synthesized ERs [80]. The listener ratings show that a trade-off exists between spatial attributes as auditory externalization and source width and attributes of intelligibility and sound quality. They suggest that the interaural cross correlation might be the explanation of the dependent link between those two types of attributes.

As mentioned in this chapter, ERs contribute substantially to the perception of externalization. The next chapter is dedicated to the investigation of digital signal processing (DSP) strategies aiming at adding ERs to the baseline anechoic binaural algorithm described in Section 2.3.

3 Adding early reflections to the remote microphone signal

Chapter 2, presented the key role of reverberation, and in particular early reflections (ERs) in the perception of externalization and auditory distance. The present chapter aims at describing several spatial binaural processing methods which were implemented or investigated during this thesis to introduce ERs in the RM signal. There are two main approaches to introducing ERs in the RM signal. The first approach consist in extracting the ERs from the signals of the microphones placed on the HI or hearables. This can be done by using either multi-channel Wiener filtering, a multi-band envelop processing, or a coherence-based method. The second approach consists in generating artificial ERs from the RM signal using partitioned convolution. All the methods were implemented with a sampling frequency of 22.05 kHz, which is a typical rate in HAs.

3.1 Multi-Channel Wiener Filter

Multi-channel Wiener filtering (MWF) [49, 36] is commonly used for noise reduction in speech enhancement applications and is an adequate candidate for dereverberating a speech signal [94]. The method consists in performing a minimum mean-square error (MMSE) estimation of a reference signal. Second order statistics of the signal during speech-and-noise and noise-only periods are computed in order to achieve this estimation and consequently filter the signal.

3.1.1 Model

Let's consider a system with two hearing devices, each being equipped with M microphones. The signal in the m^{th} microphone in the left HA can be written as:

$$y_{L,m}(\omega) = x_{L,m}(\omega) + n_{L,m}(\omega) \quad (3.1)$$

where $x_{L,m}(\omega)$ is the speech component and $n_{L,m}(\omega)$ is the noise component. ω will be omitted in the following for the sake of simplicity. The computation is made in each of the FFT bins in the implementation. Then we define the stacked M-dimensional vector containing the microphone signals for the left device:

$$y_L = \begin{bmatrix} y_{L,1} \\ \vdots \\ y_{L,M} \end{bmatrix} \quad (3.2)$$

and the stacked 2M-dimensional vector containing all the microphone signals (x and n are defined similarly):

$$y = \begin{bmatrix} y_L \\ y_R \end{bmatrix} \quad (3.3)$$

Then we define the speech and noise correlation matrices as:

$$R_{xx} = \varepsilon \{xx^H\} \quad \text{and} \quad R_{nn} = \varepsilon \{nn^H\}, \quad (3.4)$$

where $\varepsilon\{\cdot\}$ is the expectation operator and \cdot^H is the Hermitian transpose.

3.1.2 Principle

Each device estimates the speech component of a chosen reference microphone signal by applying the MWF to all the available microphone signals. The diagram in Fig. 3.1 depicts the specific case of binaural HAs where each device is equipped with two microphones (which is often the case). The reference microphone is preferably chosen as the one offering the highest SNR. Here the front microphone is used as a reference on each side.

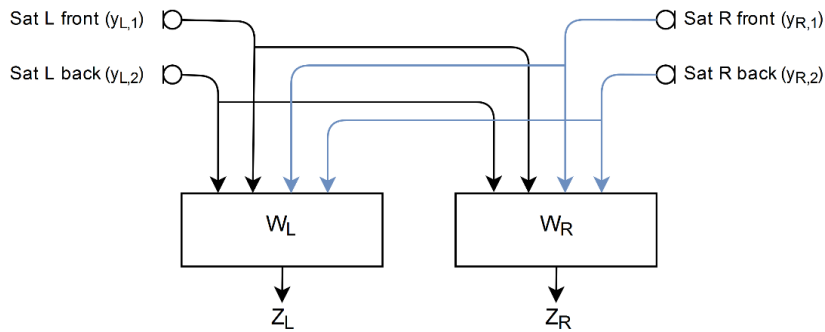


Figure 3.1: Multi-channel Wiener Filter of a 2x2 microphones binaural hearing device.

The first step requires the use of an efficient voice activity detector (VAD) to compute the correlation matrix $R_{yy} = \varepsilon \{yy^H\}$ during speech-and-noise periods and R_{nn} during noise only

periods. One must be careful that the periods used for computation are long enough to compute second order statistics. Empirically and from the literature, 1-2 seconds seems to be a minimum [127]. The speech correlation matrix R_{xx} can be estimated in a simple manner as:

$$R_{xx} = R_{yy} - R_{nn} \quad (3.5)$$

Finally the solution of the MMSE criterion leads to the MWF with speech distortion weighting (SDW):

$$W_L = \frac{R_{nn}^{-1} R_{xx} e_L}{\mu + \text{Tr}\{R_{nn}^{-1} R_{xx}\}} \quad \text{and} \quad W_R = \frac{R_{nn}^{-1} R_{xx} e_R}{\mu + \text{Tr}\{R_{nn}^{-1} R_{xx}\}} \quad (3.6)$$

where $\mu \geq 0$ is a trade-off parameter allowing to emphasize noise reduction at the cost of a larger amount of speech distortion, e_L and e_R are unit vectors with one at the index of the reference (front) microphone, and $\text{Tr}\{\cdot\}$ is the trace of the matrix. Finally, the output signals for the left and right HAs are obtained by filtering and summing all microphones signals from both devices, i.e.

$$Z_L = W_L^H y \quad \text{and} \quad Z_R = W_R^H y \quad (3.7)$$

The MWF has been demonstrated to preserve binaural cues of the speech content [16, 50]. Nevertheless, the spectral cues of the remaining noise are not preserved [154]. Therefore, it is possible to introduce a partial noise estimation (PNE) parameter $0 \leq \eta \leq 1$ to allow a certain amount of the original noise to be present in the output of the MWF [16, 154]. The resulting MWF with PNE are:

$$W_L = (1 - \eta) \frac{R_{nn}^{-1} R_{xx} e_L}{\mu + \text{Tr}\{R_{nn}^{-1} R_{xx}\}} + \eta e_L \quad \text{and} \quad W_R = (1 - \eta) \frac{R_{nn}^{-1} R_{xx} e_R}{\mu + \text{Tr}\{R_{nn}^{-1} R_{xx}\}} + \eta e_R \quad (3.8)$$

$\eta = 1$ corresponds to the case where no noise reduction is applied, and $\eta = 0$ to the previous MWF with no PNE.

3.1.3 Insertion of an additional remote microphone

The more specific case of a HA application using a RM offers several advantages in the implementation of a MWF algorithm. The high SNR in the RM signal allows a robust VAD implementation. In order to be independent of any absolute threshold, a VAD based on variations of the energy envelope of the remote signal was implemented. The noise correlation matrix R_{nn} is continuously updated as a smoothed value over time and only noise periods lasting over approximately 1 second are taken into account to ensure a reliable estimate. The high SNR remote signal can be used as a fifth microphone signal in the MWF computation as illustrated in Fig. 3.2. In the literature [154], the use of an additional signal was proven to

increase noise reduction performances and the preservation of noise binaural cues. Therefore a lower value of η can be used.

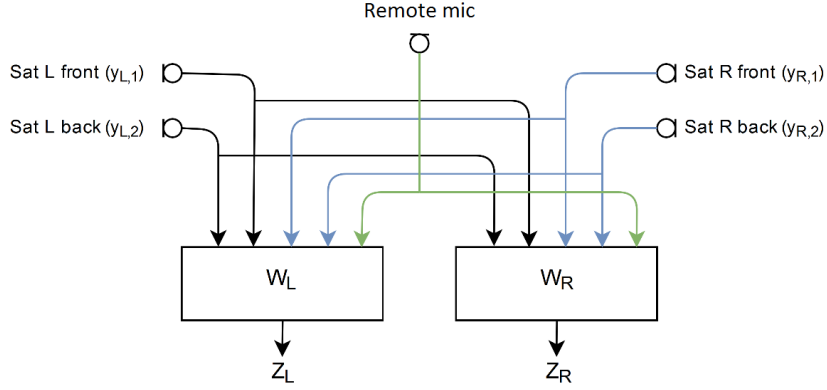


Figure 3.2: Multi-channel Wiener Filter of a 2x2 microphones binaural hearing device with additional remote signal.

3.1.4 Implementation and tuning

The tuning of the different parameters is important for an efficient and robust implementation of the MWF. The correlation matrix is smoothed over time for an efficient implementation with the parameter λ in the following manner:

$$R_{yy,k} = \lambda * R_{yy,k-1} + (1 - \lambda) * y \cdot y^H \quad (3.9)$$

where k is the frame index. λ should be small enough to ensure reactivity to the evolution of the noise in time, but large enough to ensure stability and a correct evaluation of the expected value. λ is empirically set to 0.95. $\mu \geq 0$ is another important parameter to set the trade-off between noise reduction and speech distortion. $\mu = 1$ corresponds to the case where the MMSE criterion is obtained. For $\mu \geq 1$ noise reduction will be increased, at the expense of speech distortion. $\mu = 0$ is the trivial case of no noise reduction. PNE is included to preserve the binaural cues of the remaining noise component. $\eta = 0$ leads to disturbing spatial motions of the noise and remaining reverberation. η should be set to the minimum value that preserves perceptible noise binaural cues. Empirically, $\eta = 0.1$ gave a satisfying perceptual trade-off. In this implementation, it is important to note that R_{xx} should be set to be positive semi-definite. Thus this requires to compute the eigen-values and eigen-vectors for each frame and each bin. Finally, it is desirable to have a control over the amount of reverberation that is kept so that only the direct sound and early-reflections would remain (or at least an approximation). Tweaking the previous variables allows an empirical tuning. An example of the performance of this algorithm is pictured with the corresponding spectrograms in Fig. 3.3. The upper spectrograms corresponds to a male speech in babble noise with a SNR of 10 dB. The lower

spectrogram depicts the same scenario with a more challenging SNR of 0 dB. Comparing the left and right panels, it can be seen that the MWF processing successfully removes a large amount of the noise in both cases. In the case of the SNR of 0 dB, the remaining noise during speech segments is substantially larger, as could be expected. However such low SNR correspond to extreme scenarios in the context of HAs. In such conditions, intelligibility should be prioritized over attempts of improving externalization with additional ERs. The next section aims at evaluating the performance of MWF with objective metrics.

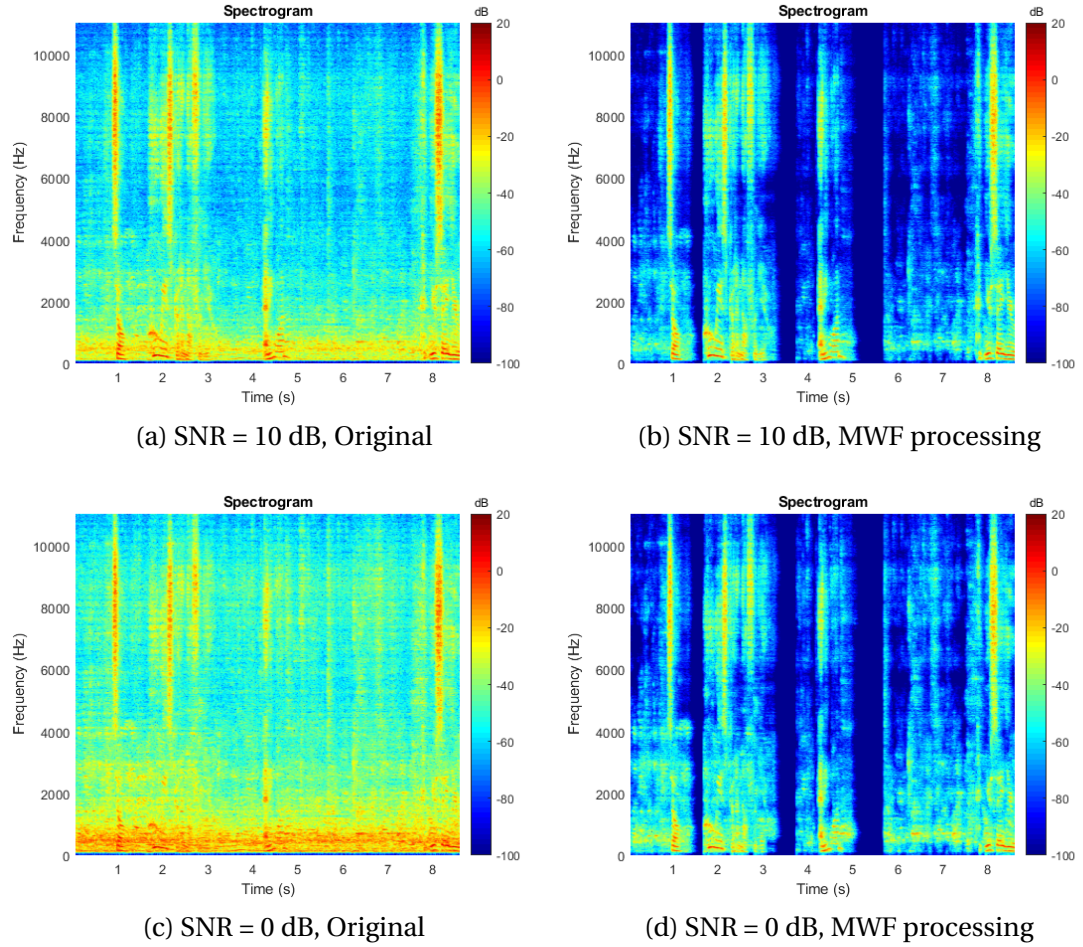


Figure 3.3: Male speech in babble noise, effect of the MWF processing for two SNR settings (SNR = 0 or 10 dB).

3.1.5 Objective assessment of the MWF method

In order to evaluate the reported method, different objective measures have been investigated. These measures aim to assess the efficiency of the MWF and the MWF with RM in terms of noise reduction as well as preservation of the speech quality. Therefore several metrics were used:

- Segmental signal-to-noise-ratio (segSNR) [70]: It takes as input the target and target+masker signals, which are divided into segments. Segment energies and SNRs are then computed. And finally, the mean over the local segmental SNRs (dB) is returned.
- Hearing aid speech quality index (HASQI) [82]: It evaluates speech quality in HAs. It is based on an auditory model that allows to include changes due to hearing loss in the evaluation, which is of special interest in this context.
- Perceptual Evaluation of Speech Quality (PESQ) [141]: It was developed to assess the voice quality perceived by humans in telecommunication applications. A reference and the tested signal are first temporally aligned, then the speech signal is analyzed sample-by-sample. The result models mean opinion scores, as used in subjective tests, on a scale from 1 (bad) to 5 (excellent).

The stimuli were obtained by convolving anechoic recordings with impulse responses from the position of the speaker to the HA microphones. The impulse responses used are part of an extensive database recorded at EPFL with a KEMAR manikin. The BRIRs of this database were recorded in a classroom environment, with a male speaker at a distance of 2 meters in front (0°) of the manikin. Three different SNRs (10 dB, 0 dB, and -5 dB) were simulated using binaural recordings of babble noise made in the same room. The results are reported in Tab. 3.1.

Measure	SNR = 10 dB			SNR = 0 dB			SNR = -5 dB		
	Orig.	MWF	MWFR	Orig.	MWF	MWFR	Orig.	MWF	MWFR
Seg SNR	-3.88	-1.46	-0.66	-5	-1.96	-1.59	-5.52	-2.87	-2.53
PESQ	1.17	1.36	1.78	1.11	1.18	1.54	1.08	1.1	1.36
HASQI	0.48	0.48	0.7	0.32	0.39	0.66	0.15	0.32	0.58

Table 3.1: Scores obtained with the objective metrics, with the original recording, the Multi-channel Wiener filter, and the Multi-channel Wiener filter with RM.

Generally, the performance of the MWF process is good in terms of SNR improvement as well as speech quality. It shows that the RM significantly improves the performance of the MWF regarding both noise reduction and audio quality. Indeed the algorithm benefits from the high SNR additional information available from the RM. Informal listening tests suggest that the processing sounds smooth and that no musical noise (particular distortion often caused by noise reductions algorithms) or other annoying artefacts can be noticed.

Spatial cue preservation is another crucial performance that should be evaluated. Indeed it is important to preserve spatial information of the ERs and the remaining direct sound to provide the feeling of a single speech source correctly located in space. Informal listening tests tend to suggest that MWF with PNE seems to preserve spatial cues of the signal in a satisfying way.

Despite the performance, MWF requires a large computational cost, which is not ideal in a HA or hearables context. The main issue comes from the exchange of information between the two hearing devices, as the 4 (or 5) signals from the 4 (or 5) microphones are streamed at each side. Furthermore, the MWF as implemented here involves the computation of eigen values and vectors of a 4x4 (or 5x5) matrix in the computation of R_{xx} . Finally, the process also includes a matrix inversion, for the computation of R_{nn}^{-1} , nevertheless this computation can be done punctually to reduce the computational cost (e.g. one time every 2-3 seconds is enough in case of a noise which properties vary quite slowly). Therefore, real-time implementation is compromised, especially with the constraints of HAs. Alternative methods with lower computational cost are thus described in the following sections.

3.2 Multi-band envelope processing

3.2.1 Principle

With a view to achieving a similar purpose of including ERs in the binaural spatialization algorithm by removing the noise and late reverberation of the signal coming in the satellite microphone, an original alternative method was proposed and implemented.

The RM speech signal has a significantly greater SNR than the signal captured by the satellite microphone. The energy envelopes for each frequency bin are computed in real-time on the remote signal. The principle is to apply a gain related to those frequency-dependent energy envelopes in the RM signal, on the satellite signal so as to filter out part of noise and reverberation, and preserve only the speech component. The processing can be applied in the FFT bins or in the Bark bands [182]. Thus for each Bark band m , the smoothed long-term energy of the RM signal E_{RM} can be computed using a recursive averager such that:

$$E_{RM}(k, m) = \lambda_m E_{RM}(k-1, m) + (1 - \lambda_m) |s_{RM}(k, m)|^2 \quad (3.10)$$

where s_{RM} is the short-time Fourier transform (STFT) of the RM signal, k is the sample time, and λ_m is a smoothing factor tuned to obtain the desired low-pass effect. Nevertheless, due to the different microphone constructions and specifications, the proximity effect, the acoustics of the room and the beamforming applied to the remote signal, the latter is significantly different than the hearing instrument signal, as shown in Fig. 3.4.

Therefore it is not suitable to apply a uniform coefficient for all frequency bins. Thus, a mapping should be defined for weighting the coefficients associated with the long-term energy of the different frequency bands of the RM signal, and one should modulate the gains applied to the satellite signal adequately. The gain $G_\alpha(k, m)$ is computed for each Bark band as:

$$G_\alpha(k, m) = \alpha(m) E_{RM}(k, m) \quad (3.11)$$

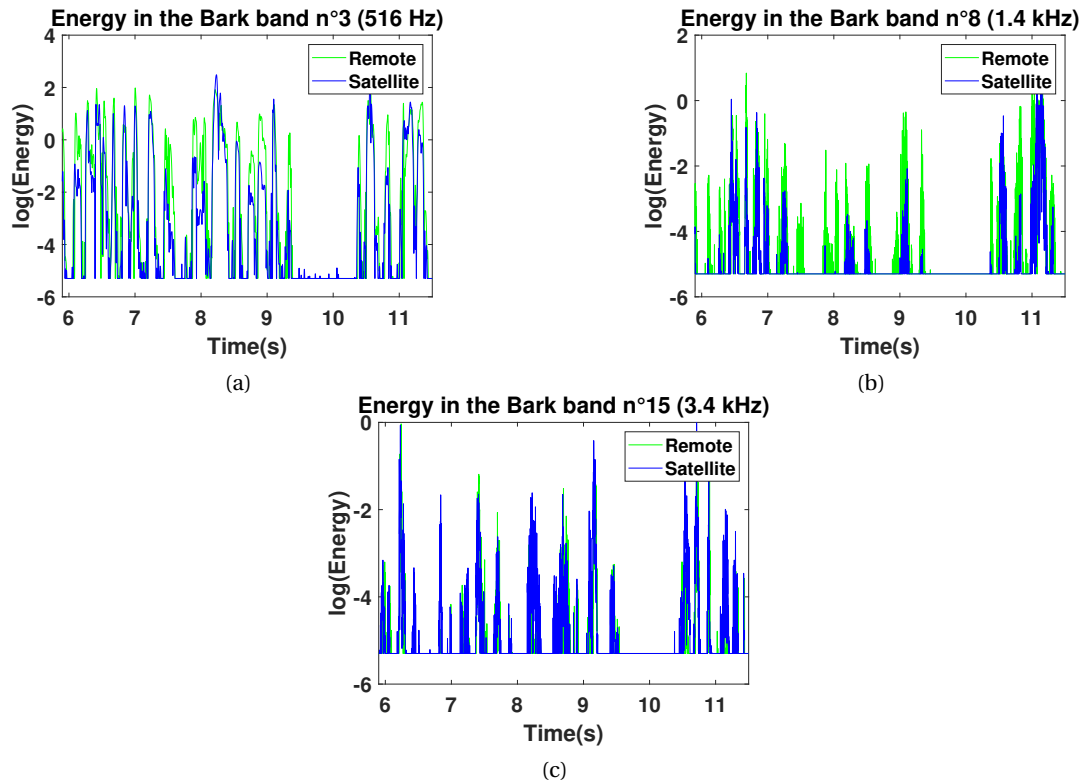


Figure 3.4: Energy comparison between the satellite and the RM signals in different Bark bands.

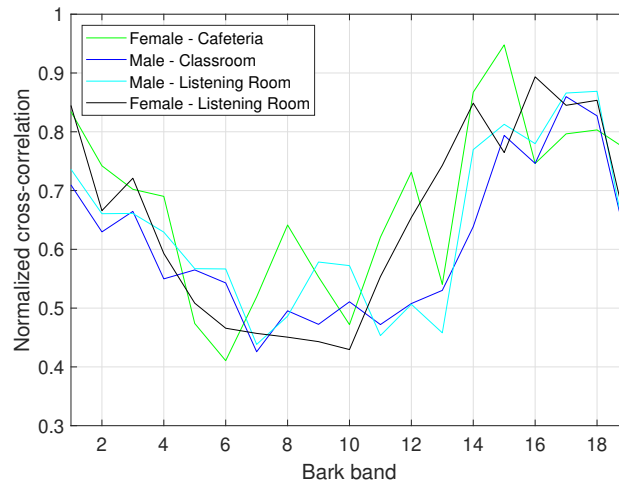


Figure 3.5: Normalized cross-correlation between the envelopes of the signals picked up by the remote and satellite microphones for various rooms and speaker (male/female).

where α is the frequency-dependent mapping coefficient tuned empirically for each Bark band. Finally, the "cleaned" output STFT $s_{ER}(k, m)$ is computed from the STFT of the signal picked by the HAs $s_{HA}(k, m)$ as:

$$s_{ER}(k, m) = G_{\alpha}(k, m) s_{HA}(k, m) \quad (3.12)$$

Looking at the cross-correlation between the envelopes of the remote and the satellite signals, as depicted in Fig. 3.5, it is possible to define sub band groups according to the degree of similarity to simplify the tuning. The tuning was achieved empirically, aiming for a trade-off between transparency and noise reduction. Nevertheless, no static optimal mapping can be chosen, as the correlation highly depends on the spectral characteristics of the speech, the position of the RM and the room acoustics.

An example of the performance of this algorithm is pictured with the corresponding spectrograms in two conditions in Fig. 3.6. The upper spectrograms correspond to a scenario with a male speaker in babble noise with a SNR of 10 dB, while the lower ones correspond to the same scenario with a SNR of 0 dB. It can be observed that the remaining babble noise is non-negligible, especially in the lower frequencies. The RM signal, despite the larger SNR, indeed contains a part of the babble noise. The residual noise is substantial, in particular between speech segments, in comparison to the MWF. This suggest that the method would benefit from a type of VAD to further reduce the noise of those segments. The method still achieve its denoising purpose at a much lower computational cost than the MWF.

3.2.2 Advantages and outlook

The method offers several advantages. It provides the ability to control the amount of ERs, by adjusting the rate of the smoothing of the envelope computation, with only one parameter. The resulting denoising is smooth and robust, and does not induce any kind of artefact such as musical noise. The same gains are applied on both ears, hence potential binaural cues distortions are small. Finally, the computational cost is minimal, as no complex computation or streaming between the two hearing instruments is necessary. In further works, optimization and evaluation of the method should be investigated as well as comparison with state-of-the-art methods such as MWF, as done in [40].

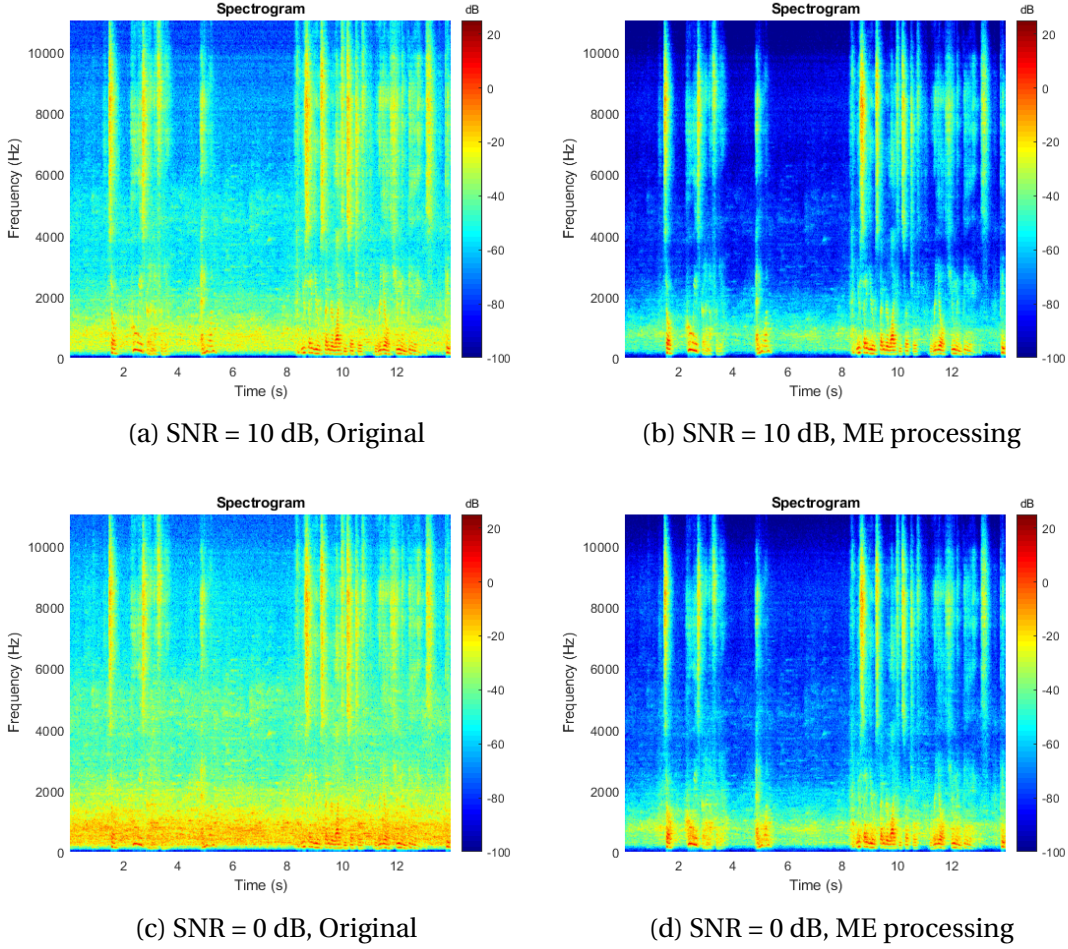


Figure 3.6: Male speech in babble noise, effect of the Multi-band envelope processing for two SNR settings (SNR = 0 or 10 dB).

3.3 Coherence-based method

This method aims at extracting ERs, by removing the noise and late reflections from the HAS microphone signal. It was proposed by Courtois et al. in [39]. This method implies the use of three blocks, named the Envelope Gain (EG), the Coherence Gain (CG) and the Gain Processing (GP) block respectively, which are structured as described in Fig. 3.7.

Envelope Gain block

This step relies on the estimation of an overall gain (frequency-independent) derived from the envelope of the RM signal. The goal of this block is thus to emphasize time segments which might include mostly the direct sound and ERs, while reducing the time periods which contain

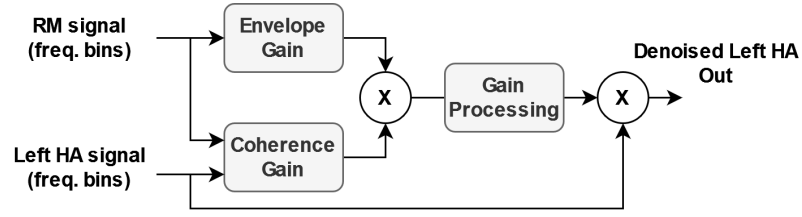


Figure 3.7: Simplified structure of the coherence-based method for ERs extraction, for the left HA.

only noise and late reverberation. The assumption is made that a decrease in energy in the RM signal should correspond to the start of the reverberation contribution in the signal captured by the HAs' microphones. Conversely, a rise of energy might correspond to an instantaneous onset of speech, and thus corresponds to time periods where the direct sound and ERs should be predominant. The derivative of the envelope is used to detect instantaneous variations of the energy. A time envelope is applied based on the detected variations in energy, and is comprised of three tunable components: the attack time, the ER time and the release time. The attack time represents the required time to reach the maximum gain value of 1. This enables to avoid abrupt changes in amplitude and partially remove the direct sound onset. The ER time represents the period to extract ERs after an offset of the speech, and is thus useful to tune the desired amount of ERs. The release time is the required time to reach the minimum gain, which also helps at avoiding abrupt sound cuts.

Coherence Gain block

In this block, the main assumption, is that time-frequency segments associated with a high coherence between the RM signal and the HA signal correspond to segments of the HA signal where the direct sound and ERs are the most important. Conversely, the time-frequency periods yielding low coherence values might be associated with segments for which noise and late reverberation are predominant in the HA signal. With this purpose, this CG block computes a frequency-dependent gain mapped to the coherence between the RM and HA signals, which aims at preserving the HA signal during segments where the two signals are similar, and reduce the HA signal while the two signals have a low coherence.

The coherence between the two signals is computed as:

$$C_{RH}(k, m) = \frac{\Gamma_{RH}(k, m)}{E_R(k, m)E_H(k, m)} \quad (3.13)$$

where k is the time sample index, m is the frequency bin index, Γ_{RH} is the cross-correlation between both signals, E_H and E_R are the long-term energies in the HA and RM signal respectively.

Γ_{HR} is computed using an exponential smoothing in time, such as:

$$\Gamma_{RH}(k, m) = \alpha \Gamma_{HR}(k-1, m) + (1 - \alpha) s_R(k, m) s_H^*(k, m) \quad (3.14)$$

where α is the smoothing factor, and s_R and s_H are the RM and HA STFTs.

A mapping function is then used, so that a gain value is computed based on the coherence between the two signals. To define this mapping, it is necessary to look at the distribution of the coherence during speech-only segments and noise-only segments. Fig. 3.8 depicts the logarithmic distribution obtained from a database with two speakers (female or male), different rooms and SNRs.

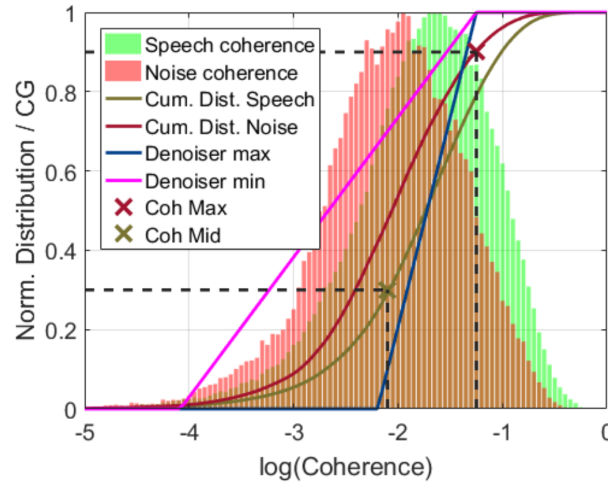


Figure 3.8: Definition of a mapping function used to compute a gain, based on the coherence of the two signals and their cumulative distributions (cum. dist.) [37].

The overlap between the two distributions is substantial. However, the distributions are still distinguishable from each other, and higher coherence values are indeed associated with speech-only segments. The noise-only segments are associated with lower coherence, as the noise included in the RM signal should indeed be low, as the signal is optimized to pick up mostly the voice from the speaker. Thus it is possible to define a linear mapping as shown in Fig. 3.8, so that time-frequency segments exhibiting high coherence are emphasized with higher gains, and lower coherence segments are reduced to a certain lower gain. The definition of this linear mapping can be made based on two points from the cumulative distributions of the speech and noise distributions ("Coh Mid" and "Coh Max" in Fig. 3.8). A more steep slope yields more noise reduction ("denoiser max") but might lead to sound artefacts such as musical noise. On the contrary, a more progressive slope will lead to less noise reduction ("denoiser min"), but the sound quality might be better preserved. Hence a trade-off has to be defined. As no clear difference in coherence distributions was found between the different frequency bin bands [37], a single mapping function can be used across frequency bands.

Gain Processing block

This block multiplies the gains obtained in the two previous steps, with the EG and CG blocks. From this result, the Gain Processing block computes the final output frequency-dependent gains. In this block it is possible to achieve a smoothing across frequency bands, to minimize the abrupt gain changes. This minimizes musical noise type artefacts, but might also reduce the denoising performance. This block can also serve to define the minimum gain introduced during noise only-segments. Indeed it is not desirable to provide a gain of 0 which would create an annoying scattering effect and would isolate the listener too much from the other sounds of the environment.

Performance

In [42], the coherence-based method was compared with several objective metrics against the two versions of the MWF method (with and without RM). In particular the same metrics as described above were used: SegSNR, PESQ and HASQI. The scores were computed by taking the difference (Δ) between the score of the input and the output of the algorithm. The simulation included several SNR settings from -5 dB to 30 dB, with diffuse babble noise in a reverberant room (classroom at EPFL). The results of these objective metrics in function of the SNR are shown in Fig. 3.9.

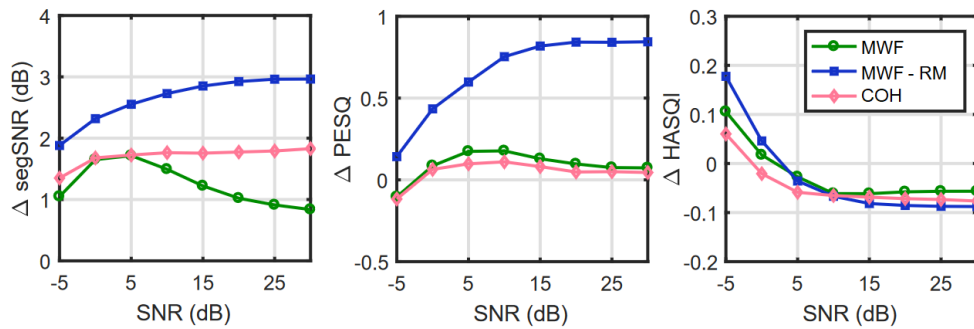


Figure 3.9: Objective metrics as a function of the input SNR, for the performance comparison between the MFW, the MFW with RM (MFW - RM) and the coherence-based method (COH). The measures were made with diffuse babble noise in a reverberant room (classroom at EPFL) (taken from [42]).

In general the MWF with RM gave the best results in terms of speech quality, especially regarding the Δ SegSNR and Δ PESQ. The Δ PESQ indicates that both the coherence-based algorithm and the MWF without RM preserve the speech quality while the MWF with RM enhances it. The large improvement in Δ SegSNR for the MWF with RM is explained as the clean RM signal is partially introduced in the signal with this method. In general, the performance of the coherence-based method was fairly comparable to the performance of

the MWF without RM. The Δ HASQI indicates a slight degradation at high SNRs and a small improvement at low SNRs, with a similar tendency for the three algorithms. This last result might be contradictory with the first one, showcasing the limitation of objective metrics. Hence, subjective listening studies should be undertaken to complete the evaluation.

3.4 Partitioned convolution

While the three previous methods described aimed at extracting ERs from the RM signal by removing noise and late reverberation from the microphones placed on the HAs, the method described in this section aims at synthesizing artificial ERs that can be superimposed to the direct sound.

In the context of an offline implementation, convolution with a binaural room impulse response (BRIR) can be performed to simulate the auditory cues of a virtual source at a precise location. However, the computational cost of convolution in the time domain is too high for an application such as HAs. Multiplication in the frequency domain cannot be applied directly either, as it would lead to a too large latency and is not compatible with the frame-based real-time processing of HAs, using 128-samples frames (5.8 ms at 22.05 kHz).

Partitioned convolution is a more efficient way to perform convolution in real-time application. It was implemented in this work both offline and online.

3.4.1 Uniform partitioned convolution

Partitioned convolution was first introduced for general purpose by Stockham [78]. Then it was applied to DSP and audio by Kulp [96]. The real-time implementation of this technique as been described in [157] in the context of ambiphonics surround sound. A real-time implementation is also described in detail in [5]. It relies on the partition of the impulse response into a series of smaller blocks. Uniform partitioned convolution corresponds to the case when the size of the blocks is constant. An example of partitioning in P blocks is depicted in Fig. 3.10.

Those blocks can be seen as separate impulse responses or sub-filters that can be run in parallel. More precisely, it is possible to use the FFT of each block with appropriate delays.

A complete implementation is depicted in Fig. 3.11 for a simple case with a number of partitions $P = 3$. The impulse response is split into blocks of K points. Each block is zero-padded to the length L (power of 2 closer to $2 \cdot K$) and FFT is applied. $S_1, S_2, S_3 \dots$ are the frequency domain filters corresponding to the blocks of the impulse response, which can be processed offline beforehand. This algorithm was implemented both offline and in real time in Simulink. In the framework of the algorithm used in this thesis, the choice of blocks with $K = 128$ is ideal for the implementation, as the current implementation of the spatialization with generic HRTFs is also based on an implementation using 128-samples frames.

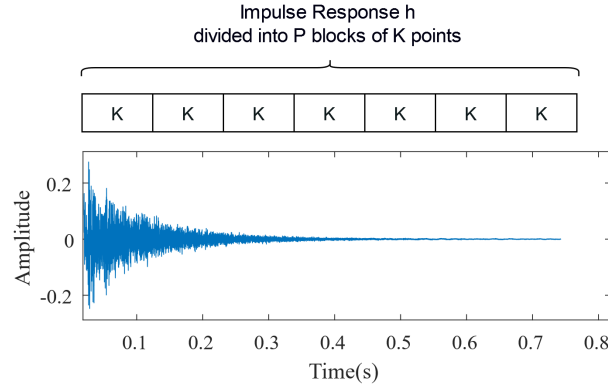


Figure 3.10: Impulse response partition in the case of a uniform partition.

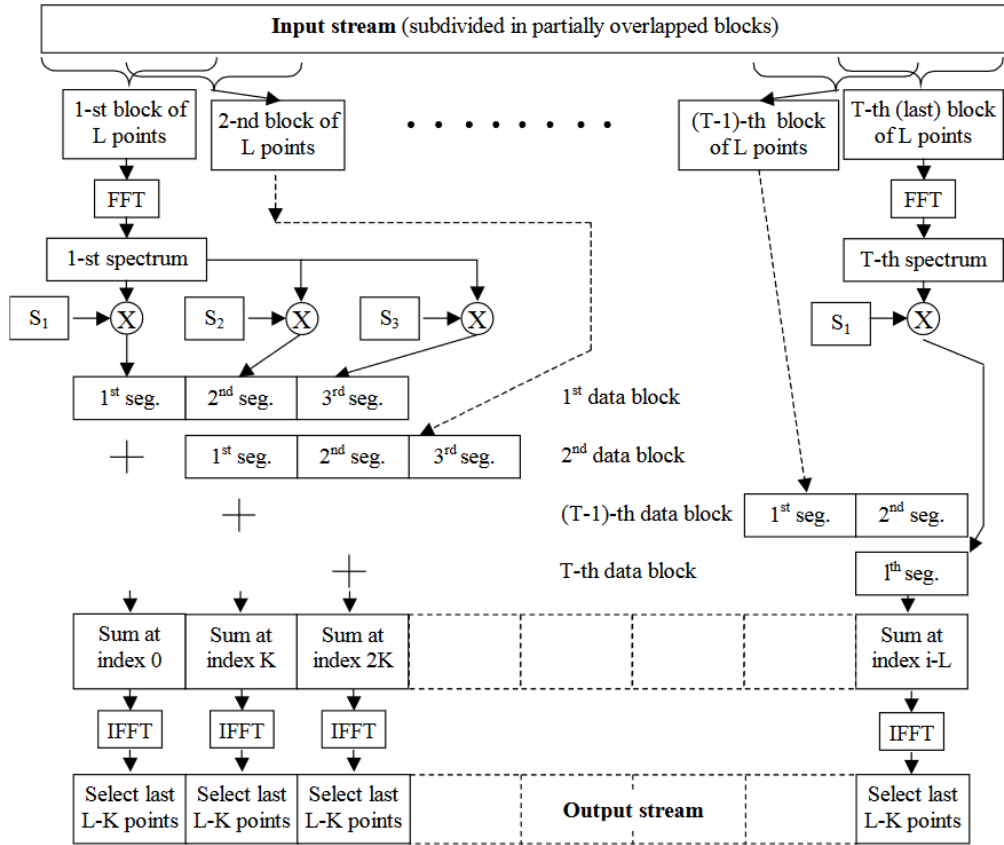


Figure 3.11: Graphic representation of the real-time implementation of the partitioned convolution algorithm (taken from [5]).

Only one FFT is applied for each block of input data, and a single IFFT is needed following the frequency domain summation. Those steps are already included in the current anechoic

spatial processing implementation which served as a starting point for these investigations, therefore no additional FFT computation is performed. The resulting latency from the whole processing is L points (256 samples, i.e. 11.6-ms at 22.05 kHz), which is acceptable in the context of binaural audio. The computational load is rather modest owing to the fast memory access available, as the data structures processed here are small.

For the application of interest, the first 5.8 ms of the impulse response are set to 0 (128 first points at 22050 Hz). This avoids duplication of the direct sound which is already provided by the generic minimum-phase HRTFs.

3.4.2 Non-uniform partitioned convolution

Another method consists in using a non-uniform partition of the impulse response, in order to optimize the computational efficiency. This is usually done by using shorter blocks for the beginning of the impulse response, which allows to achieve low latency, and progressively longer blocks toward the end of the impulse response for optimal computational efficiency. An example is depicted in Fig. 3.12. The implementation optimization can be made with several strategies depending on the type of operational system [11]. Nevertheless, the frame-

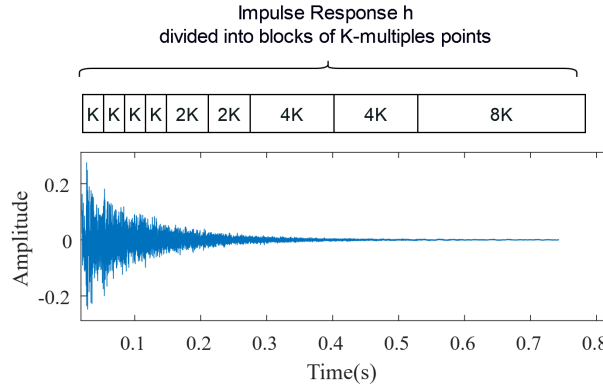


Figure 3.12: Impulse response partition in the case of a non-uniform partition.

based implementation of the algorithms used in this work encourages to work with a uniform partition, for which the size of the block equals the size of the frame, which corresponds to 128 samples at 22050 kHz, i.e. 5.8 ms. Consequently, the implementation with uniform partition implemented during this thesis takes advantage of the FFT processing already performed in the anechoic spatialization by using the same frames from the RM signal. Also, the frame size latency remains small in comparison to threshold of detection values, i.e. around 60 ms as reported in [26, 104]. Moreover, the synthesis only aims at adding ERs to the direct sound, so only rather short versions of the impulse responses are necessary and can be covered with a few blocks. For example, only four blocks are necessary to yield an impulse response of 23.2 ms, which might already be sufficient for the purpose of externalization [159]. Thus, so

far, nothing would justify the resort to non-uniform partitioned convolution to optimize the computation in this case.

3.5 Discussion

3.5.1 Performance

Although being a promising candidate for extracting ERs from the HAs microphone signal, the Multi-channel Wiener Filter with RM method is computationally heavy, and the implementation would not be compatible for the real-time rendering with small wearable devices such as HAs or hearables.

While being less efficient in noise reduction than the MWF, the coherence-based method still allows to denoise the signal at a lower computational cost. It can be noted that the performance of this method was significantly degraded when a delay was simulated between the RM and the HA microphone [42]. Indeed the RM and HA signal might reach the HA at different times, and thus not be temporally aligned. The RM signal delay is dependent on the digital communication protocol, while the HA signal delay is mainly defined by the time of flight between the speaker and the listener. The use of objective metrics is limited for the appreciation of speech sound quality, and thus subjective evaluation should also be conducted to compare the different methods further. As described, this processing is independent in each device. The ITD should not be affected, but it is possible that the ILD could be altered. Thus, it is necessary to ensure that binaural cues are preserved with an additional block linking the two binaural devices.

The original approach of the Multi-band envelope processing proposed in this section provides several advantages. Its implementation is rather simple, the denoising is efficient and the resulting sound can be qualified as quite "natural". Nevertheless, other methods were prioritized during the course of the developments reported in this thesis, therefore further developments should be conducted to define a tuning strategy, and to compare its performance against other state-of-the-art methods such as the MWF.

The use of partitioned convolution seems quite adapted in this context. First, the definition of the BRIRs length is accurate as it can be defined directly by the truncation of the BRIRs used for the synthesis. The frame-based implementation of the baseline spatialization algorithm also gives the necessary time-frequency transformation, so no additional FFT computation is necessary. As the purpose is only to provide a short amount of ERs in the synthesis, only a couple of 5.8-ms (128 samples at 22050 kHz) blocks is necessary to cover the BRIR length necessary to provide externalization. Minimizing the necessary memory usage is indeed crucial in wearable device implementations.

3.5.2 Perceptual considerations

When using partitioned convolution alone, i.e. only artificially generated ERs, no sounds from the environment are available anymore to the listener, which might be problematic for spatial awareness. This issue is addressed in Chapter 5, where several algorithms that aim at providing ERs and a certain amount of spatial awareness are proposed, including the partitioned convolution approach.

In the case of artificial ER, it is likely that the simulated reflections might not match the ones that would be occurring in the room where the listener is located. Begault [12] found that synthetic reverberation was enough to improve externalization but could decrease the performance in localization. However, Zahorik et al. [179] suggested that an accurate information in the reflections might not be necessary to yield an accurate performance in localization, including distance evaluation. According the authors, this might be explained by the robustness of the precedence effect [105]. Nevertheless these conclusions were observed mainly with the use of individualized HRTFs, contrarily to the study in [12] which used non-individualized HRTFs.

Moreover, the acoustic properties of room used to record the truncated BRIRs stored in the device may not match those of the room where the listener is located during playback. This can raise questions about the effect on the perceived externalization because of the previously mentioned room divergence effect [170]. However, the BRIRs used for this purpose could be selected to match an average room representing most of the use cases. For example, if it is to be used for attending classes (where the teacher uses a RM), it would be suitable to use a pair of average classroom BRIRs in the devices to yield room congruence in most cases. Methods aiming at extracting ERs from the HAs' microphone have the advantage to directly introduce the actual reflections from the room where the listener is located, yielding an automatic room congruence in any situation.

In the methods providing ERs extracted from the wearable device's microphones signal, it can be argued that the ERs cues are more individualized compared to the artificial ERs generated with partitioned convolution. However, Leclère et al. suggested in [97] that individualized BRIRs may not be necessary for auditory externalization. Zahorik suggested that individualized HRTFs might also be unnecessary for auditory distance estimation [176, 175], as long as natural reverberation was provided.

While the partitioned convolution approach can provide exactly the desired BRIR length to the listener, the other methods relying on ERs extraction from the HAs microphones are less accurate in terms of the amount of ERs that are introduced. It is possible to tune them by adjusting the mapping functions in high SNR scenarios.

For the methods aiming at extracting ERs from the HA microphones' signals, it can be argued that the filtered sound may still contain part of the direct sound. For those three methods, the algorithms include processing in time-domain related to the exponential averaging applied

on the signals. For the multi-band envelope processing and the coherence-based method, this is achieved directly in the computation of the envelopes. In the MWF approach, this is performed during the computation of the correlation matrices, which are smoothed over time with exponential averaging. With an appropriate tuning of the time constants, it is possible to partially remove the direct sound. In particular, it is relevant to use different time-constants for the attack and the decay to obtain a more accurate control of the removal of the direct sound and the time length of ERs. The tuning can be achieved with clean recordings (SNR > 30 dB). For the case of partitioned convolution, it is easier to provide only ERs and remove the direct sound. It is possible to set to zeros the first samples (5.8 ms) of the BRIRs used in the processing. This corresponds to the first frame (128-samples at 22.05 kHz) in the implementation. This avoids the superimposition of the direct sound from the HA's microphone signal to the one generated with the baseline spatialization algorithm. This was also found to be the more "natural sounding" method during informal listening session.

The advantages and drawbacks of the considered methods are summarized in Tab. 3.2.

Method	Advantages	Drawbacks
Multi-channel Wiener Filter	Best performance (state-of-the-art) Spatial cue preservation ERs matching the room	Too heavy CPU cost Part of the direct sound may remain
Multi-band Envelope proc.	Low CPU cost Little distortion of ILD ERs matching the room	Requires further tuning Lower performance compared to MWF Part of the direct sound may remain
Coherence-based proc.	Low CPU cost Performance / CPU cost ratio ERs matching the room	Additional block to preserve ILD Very sensible to latency Part of the direct sound may remain
Partitioned convolution	Very accurate control of ERs «Clean» and natural ERs Provide just ERs (no direct sound) Independent of the SNR	ERs not necessarily matching the room Memory (if several BRIRs stored)

Table 3.2: Summary of the main advantages and drawbacks of the investigated methods

While objective metrics provide a base to discuss a choice of algorithm, they do not indicate the perceptual consequences of these approaches on auditory distance perception. The next two chapters report perceptual listening studies assessing the perception of auditory distance with various binaural spatialization methods including the superimposition of ERs generated with the coherence-based and partitioned convolution-based strategies described in this chapter.

4 Auditory distance perception in NH and HI listeners with various binaural rendering strategies

In Chapter 2, we have seen that the auditory system enables listeners to estimate the location of auditory events resorting notably to individual binaural cues and room reflections reaching the eardrum. For HI listeners using HAs with RM systems, those cues are not available if the sound is presented diotically. Several methods to introduce ERs in the RM signal were presented in Chapter 3. It is hypothesized that superimposing ERs to the spatialized direct sound should help to perceive sounds as externalized and improve auditory distance perception. Nevertheless, it could be expected that a part of the cues might be distorted by the HA processing. This could make auditory distance perception a challenging task. Moreover, it is not clear whether HI listeners are able to use those cues or not. The study reported in this chapter^I, aims to evaluate the effect of several binaural rendering strategies on auditory distance perception (ADP) in HI and NH listeners. Within a multiple-stimulus listening paradigm, NH and moderate-to-highly-profound HI listeners were asked to rate the perceived distance of various stimuli presented over headphones, while visual cues were available.

4.1 Context and motivation

As mentioned in Chapter 2, remote wireless microphone systems are used with HAs to provide a more intelligible speech signal to HI listeners [17]. The voice of the speaker is picked up close to the mouth by a body-worn microphone and transmitted wirelessly to the HAs, usually diotically. While providing large improvements in speech intelligibility and reduced listening effort in challenging environments [172], this technology most commonly discards any spatial cue for localizing the source with the auditory system.

A real-time binaural localization and tracking algorithm for HAs and RM systems introduced

^IThis chapter is an adapted and extended version of the paper [68], associated with the article [41].

in [44] was described in Chapter 2. HAs process sound within short-time frames (typically less than 10 ms) so as to ensure low-delay rendering of the output signals. Therefore, short impulse responses containing mostly the direct sound with its binaural cues are used. When using simple generic head-related transfer functions (HRTFs), speaker localization experiments using the aforementioned feature revealed that most HA users had similar performance with natural and artificial spatial hearing, indicating that HI listeners keep their localization abilities in the frontal plane [43]. Nevertheless, in-head localization (lateralization) was reported to be experienced by the HI listeners, and mentioned as a limitation of this new feature. Externalization can be seen as a prerequisite for auditory distance perception.

As seen in Chapter 2, rendering a convincing externalized acoustic scene over headphones can be achieved by including the spatial cues of the filtering caused by the head, pinna and torso, as contained in the HRTFs, as well as reflections from in the environment [71, 13]. The resort to non-individualized HRTFs is known to yield a reduction of the degree of externalization [169]. Reflections have an important role in the perception of externalization [71], and more generally in the perceived distance of sound sources [23]. The length of the impulse response affects the degree of externalization, up to around 80 ms [34]. The direct-to-reverberant energy ratio (DRR) is used by the auditory system to estimate auditory distance [92].

The effect of hearing loss (HL) on the perception of externalization and auditory distance has been addressed in few studies in the literature. In [2], Akeyroyd et al. assessed the ability of elderly HI listeners to discriminate differences in the cues to the distance of speech stimuli. The HI listeners performed similarly as NH listeners when both DRR and level information were available. Nevertheless, they performed more poorly than NH subjects when the level cue was fixed and only the DRR was available. These results suggest difficulties to resort to DRR to evaluate distance in presbycusis subjects. Wide dynamic range compression (WDRC) [9] is one of the main processing performed by HAs, and is intrinsically non linear. Its goal is to provide a dynamic gain to make low-level sounds audible while not making higher-level sounds uncomfortable or painful, and compensate for loudness recruitment [152]. Thus, it is likely that WDRC alters the cues involved in a localization task. However, WDRC was found to have no significant influence on the performance in horizontal localization with HI listeners in [84]. In [3], WDRC did not have an effect on the performance of experienced HA users in a distance discrimination task, which can be explained by the accustomization of subjects to their HA settings. Moreover, in this study, only small compression was applied. This contrasts with the larger amounts of compression that can be used for severe or profound HI subjects in real fittings. The importance of the transparency of the amplification method regarding the distortion of localization auditory cues, notably level-related cues used for distance estimation, was underlined in [148]. In a localization task including both NH and HI subjects, fast compression, i.e. short attack time, was shown to increase the just-noticeable differences (JND) in ILD compared to linear compression [126]. In [18], HI listeners with mild-to-moderate HL and NH listeners evaluated the perceived externalization of binaurally synthesized speech sentences in a reverberant environment with available visual cues. It was suggested that HI listeners experience a contracted perception of externalization, since

stimuli were never rated as externalized or internalized as for NH listeners. The results showed that this observation was correlated with high-frequency HL. By varying the amount of head-related binaural information available to listeners during headphone reproduction, Ohl et al. [130] found that externalization perception abilities were less homogeneous in HI listeners compared to NH listeners, and that in general, HI listeners are less sensitive to changes in externalization. The influence of a RM mixed with the HA microphone signal on spatial perception was investigated in [145]. It was concluded that the gain of the RM should be reduced as much as possible to optimize localization, which is not desirable considering the intended purpose of the RM to improve speech intelligibility.

As it is not reasonable in practice to measure individual HRIR for every RM system user, it is desirable to investigate if a generic method could provide externalization to RM system users, and in particular HI listeners. Recently, a novel algorithm aiming at improving externalization in HAs was introduced in [81] and relies on the "structural binaural model" introduced in [24]. This model consists in combining separate individual filters associated with all the components of the HRIR (head, pinna, ear canal). The study evaluated the effect of various procedures for speech externalization presented through earphones to NH and moderate HI listeners. The addition of room reverberation brought a significant improvement in externalization compared to the anechoic HRIRs. The structural binaural model with simulated reverberation resulted in similar performance as the use of the listener's own HRIRs.

The present study aims at evaluating the perceived auditory distance obtained from several reproduction strategies in NH and HI listeners. In the context of RM systems it is of particular interest to investigate if the addition of early reflections (ERs) in the binaural synthesis of speech using generic HRIRs can provide a substantial improvement in the perception of auditory distance. Contrarily to the previous studies, HI listeners with a high degree of HL were included in the evaluation.

4.2 Methods

4.2.1 Participants

The subject panel consisted of 10 NH (med. age = 22 y.o., 5 female) and 20 HI (med. age = 31 y.o., 10 female) listeners. The mean of the pure-tone average (PTA) was 93 dB at the best ear (range 55-114 dB) for HI listeners. Among them, in regards of the HL grades as classified by the World Health Organization (WHO) [132] and reported in appendix Tab. A.1, two had a moderate HL, three had a severe HL and 15 had a profound HL (among which 11 had a PTA at the best ear higher than 100 dB HL). All HI listeners had congenital or pre-lingual HL. The audiograms of all participants are available in appendix A.1.

4.2.2 Experimental setup and measurement phase

The experimental setup was mounted in a classroom at EPFL ($V = 177 \text{ m}^3$, $RT_{60} = 530 \text{ ms}$). Three loudspeakers (Genelec 1029A), numbered from 1 to 3, were placed at 67 cm, 113 cm and 200 cm respectively from the listener's position. They were located at an azimuth of 30° on the right side of the listener, as shown in Fig. 4.1. This angle ensured that the loudspeakers were easily visible at the listener's position with no head movement required. They were placed with a slight increasing height, so that the further loudspeakers would not be masked by the closer ones.

The experiment started with a measurement phase. The subjects sat in the center of the room and had their head immobilized by mean of a chin rest. In-the-ear binaural microphones (The Sound Professionals MS-TFB-2) were used to measure their individual BRIRs corresponding to the three loudspeakers, as well as their headphone-to-ear impulse responses (HPIRs). For HI listeners, the real ear aided response (REAR) was measured for each ear using a probe microphone measurement unit (Aurical FreeFit) using a speech-shaped noise (SSN) stimulus played from the loudspeaker of the device at 65 dB SPL while they were wearing their HAs. Thereafter, during the experiment, HI listeners did not use their HAs, and all the stimuli were rendered through a pair of open headphones (Audeze LCD-2C). Additionally, a headphones amplifier (Lake People HPA RS 02) was used. These models were chosen to allow to render high output level with a low total harmonic distortion (THD) (THD = 0.4% measured at 120 dB). The custom WDRC settings from the participant's HAs were extracted from the audiologist fitting software (Phonak Target) and simulated in Matlab.

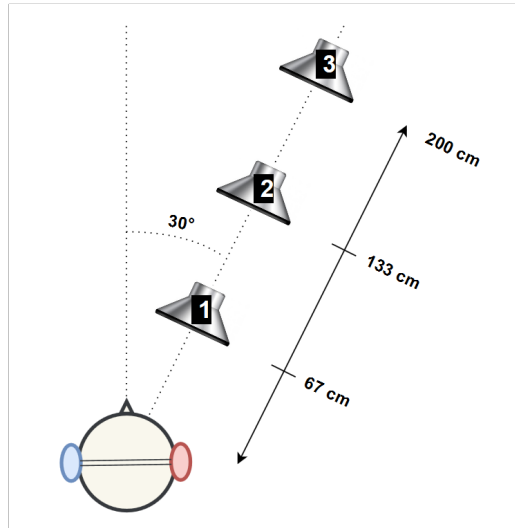


Figure 4.1: Schematic representation of the experimental setup mounted in a classroom.

4.2.3 Stimuli

The signals were 10-second speech sequences, derived from the concatenation of 5 short phonetically-balanced random sentences (French HINT database^{II}). They were processed with different methods described as follows:

- Reference (**Ref**): the speech sequence was convolved with custom BRIRs of Loudspeaker 3. The total length of the BRIRs is 740 ms. Those stimuli were expected to be perceived the furthest.
- Diotic (**Diotic**): the original speech sequence was reproduced diotically. Those stimuli were expected to be perceived the closest.
- Truncated to 60 ms (**ER60**): the speech sequence was convolved with a truncated version (60 ms) of custom BRIRs of Loudspeaker 3. This corresponds to the direct sound and ERs. A Hann falling half-window was applied on the last 1 ms of the truncated version.
- Generic (**Gen**): the speech sequence was filtered by generic minimum-phase 128-sample (5.8 ms) HRIRs, as measured on a KEMAR manikin (G.R.A.S, type 45BB) in an anechoic chamber with a source located at a distance of 2 m, and an azimuth of 30°. The ITD is simulated by a pure delay and corresponds for this source location to 210 μ s. The algorithm is described further in Section 2.3. This is comparable to a generic anechoic spatialization of a source at the position of Loudspeaker 3.
- Generic with ERs (**Gen/ER**): the stimuli were obtained from a mix of the **Gen** stimulus and additional ERs extracted from the HA microphones using the coherence-based algorithm described in Section 3.3. The HA microphone signals were generated with BRIRs of HAs worn by a KEMAR at the position of the listener, measured beforehand in the same listening room. The algorithm was tuned to yield comparable amount of ERs as the **ER60** stimuli.

The stimuli were compensated with the frequency response of the headphones measured for every subject. For HI listeners, all stimuli were processed with the WDRC settings obtained from the fitting software applied after the spatialization processing. Two cases were considered regarding the DRC, hence the HI listeners had to perform the test twice to experience both conditions. First case (WDRC-A), consists in applying the DRC after the spatial processing, which optimizes audibility for the listener. The second case (WDRC-B) consists in applying the DRC before the spatialization processing, which optimizes the accuracy of the binaural cues. In the case of the diotic reproduction, the WDRC was applied in the same way for both conditions. Half of the HI participants started with the first condition, while the other half started with the second condition.

^{II} Collège National d'Audioprothèse 2006

For NH participants, no WDRC (i.e. linear amplification) was performed. A tenth-order Butterworth low-pass filter with a cutoff frequency at 6.5 kHz was applied to prevent any effect from headphone positioning and to simulate the typical power hearing-aid bandwidth. All stimuli were normalized with the root mean square values to limit any potential effect of loudness differences in the perceived distance of the source. For NH listeners, the stimuli were rendered at 65 dB SPL. Based on the individual measurements of the REARs, the reproduction level varied between 67 and 120 dB SPL for HI listeners.

4.2.4 Task

Using a MUSHRA-type graphical user interface (GUI) displayed on a touch-pad, the participants were asked to evaluate the perceived auditory distance for the five stimuli presented simultaneously. The instruction consisted in answering the following question: "How far do you perceive each stimulus from your position?". They used a continuous scale displayed as a slider with the following markers: *Center of the head* (0), *Boundary of the head* (20), *At Loudspeaker 1* (40), *At Loudspeaker 2* (60), *At Loudspeaker 3* (80) and *Further than Loudspeaker 3* (80 to 100). The task was repeated over 4 to 6 runs for each participant, depending on the consistency of the ratings of each subject over the three last consecutive runs. The experiment was preceded by a training phase, in which the listeners were given the possibility to listen as much as they want to three versions of a speech sequence spatialized with their custom BRIRs corresponding to Loudspeakers 1, 2 and 3. This helped subjects to get accustomed to the task. Additionally, this served to ensure that their auditory spatial perception matched the visual location of the loudspeakers, as well as give them an *a priori* knowledge of the reproduction level used along the experiment.

4.3 Results

Auditory distance evaluations are reported in Fig. 4.2^{III} for NH listeners, and in Fig. 4.3 for HI listeners. For every subject, the three runs with the lowest variances on ratings of the reference(**Ref**) and **Diotic** stimuli were considered for the analysis. The first run was not taken into account and considered as training.

Data from MUSHRA-type tests are known to violate most of the assumptions associated with parametric tests such as the ANOVA [119], thus non-parametric statistics were used for the analysis of the auditory distance evaluations. Two HI subjects in the WDRC-A condition and one more HI subjects in the WDRC-B condition rated the diotic stimuli to be more externalized than the reference, thus they were excluded from the statistical analysis. Those three subjects had PTA at best ear of over 100 dB HL.

^{III}Note for all the boxplots displayed in this thesis: the line in the middle of the boxes corresponds to the median, while the bottom and top lines represent the 25th and 75th percentiles, respectively. The T-bars (whiskers) correspond to the minimum and maximum values within 1.5 times the interquartile range. The additional points are outliers.

4.3.1 Auditory distance perception in NH listeners

A Friedman test among repeated measures revealed significant differences in the perceived auditory distance between the five stimuli in NH listeners ($p < 0.001$, $\chi^2 = 114.96$). A large effect size was found using Kendal's W ($W = 0.958$).

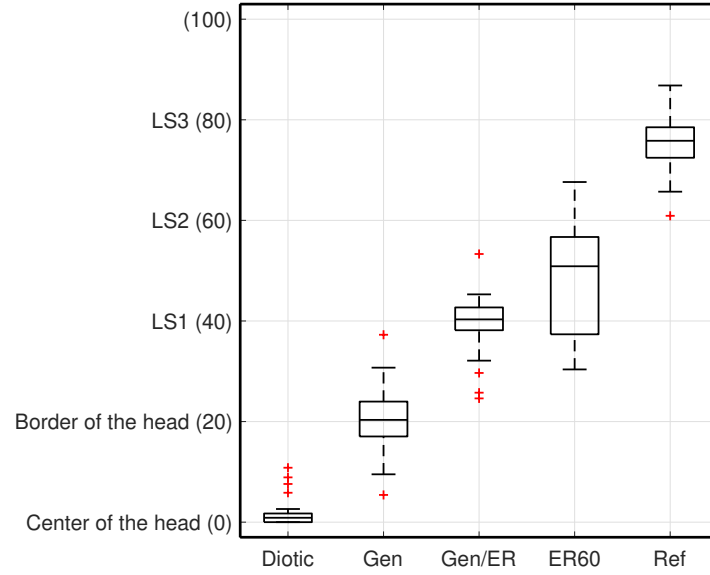


Figure 4.2: Auditory distance as evaluated by NH listeners on a continuous scale with the following markers: *Center of the head* (0), *Boundary of the head* (20), *At Loudspeaker 1* (40), *At Loudspeaker 2* (60), *At Loudspeaker 3* (80) and *Further than Loudspeaker 3* (80 to 100).

Post-hoc Wilcoxon signed-rank tests with Bonferroni correction were conducted (summarized in Tab. 4.1).

	Diotic	Gen	Gen/ER	ER60
Gen	0.143			
Gen/ER	< 0.001	0.015		
ER60	< 0.001	< 0.001	1.000	
Reference	< 0.001	< 0.001	< 0.001	0.015

Table 4.1: Perceived auditory distance (NH listeners), Wilcoxon signed-rank tests results (significant p -values are in blue).

The **Ref** stimuli were perceived further than the **Diotic** ($p < 0.001$), **Gen** ($p < 0.001$), **Gen/ER** ($p < 0.001$) and **ER60** ($p = 0.015$) stimuli. No significant difference was found between the **Diotic** and the **Gen** stimuli ($p = 0.143$). The **Diotic** stimuli were more internalized compared to the **ER60** ($p < 0.001$) and **Gen/ER** stimuli ($p < 0.001$). The **ER60** stimuli were perceived

significantly further than the **Gen** stimuli ($p < 0.001$), but not significantly different compared to the **Gen/ER** stimuli ($p = 1.000$). Finally the **Gen/ER** stimuli were perceived significantly further than the **Gen** stimuli ($p < 0.001$).

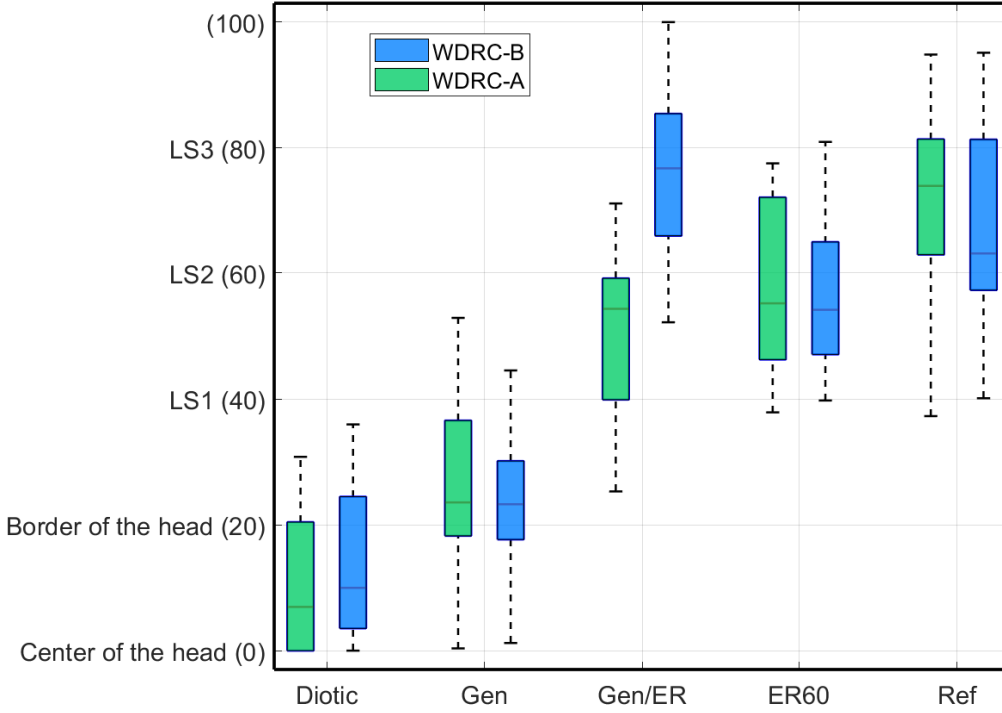


Figure 4.3: Auditory distance as evaluated by HI listeners on a continuous scale with the following markers: *Center of the head* (0), *Boundary of the head* (20), *At Loudspeaker 1* (40), *At Loudspeaker 2* (60), *At Loudspeaker 3* (80) and *Further than Loudspeaker 3* (80 to 100). WDRC-A is the case where the WDRC is performed after the spatial processing, and WDRC-B is the case where the WDRC is performed before.

4.3.2 Auditory distance perception in HI listeners

WDRC after spatial processing (WDRC-A)

A Friedman test among repeated measures showed that the evaluated auditory distance were significantly different between the five stimuli for the HI listeners ($p < 0.001$, $\chi^2 = 156.76$). A large effect size was found using Kendal's W ($W = 0.726$). *Post-hoc* Wilcoxon signed-rank tests with Bonferroni correction were performed (summarized in Tab. 4.2). The **Ref** stimuli were perceived significantly further away than the **Diotic** ($p < 0.001$), **Gen** ($p < 0.001$) and **Gen/ER** ($p < 0.001$) stimuli, but not from the **ER60** ($p = 0.074$) contrarily to NH listeners. Similarly as for the NH listeners, the **Diotic** stimuli were more internalized compared to the **ER60** ($p < 0.001$).

and **Gen/ER** ($p < 0.001$) stimuli, but not significantly different from the **Gen** stimuli ($p = 0.051$). The **ER60** stimuli were perceived significantly further than the **Gen** stimuli ($p < 0.001$), but not significantly different from the **Gen/ER** stimuli ($p = 0.634$). The perceived auditory distance of the **Gen/ER** stimuli was significantly higher than in the **Gen** stimuli ($p = 0.002$).

	Diotic	Gen	Gen/ER	ER60
Gen	0.051			
Gen/ER	< 0.001	0.002		
ER60	< 0.001	< 0.001	0.634	
Reference	< 0.001	< 0.001	< 0.001	0.074

Table 4.2: Perceived auditory distance (HI listeners, WDRC-A), Wilcoxon signed-rank tests results (significant p -values are in blue).

WDRC before spatial processing (WDRC-B)

A Friedman test among repeated measures was also conducted for this condition. It revealed that the evaluated auditory distance was significantly different between the stimuli ($p < 0.001$, $\chi^2 = 169.90$). A large effect size was found using Kendal's W ($W = 0.833$). *Post-hoc* Wilcoxon signed-rank tests with Bonferroni were conducted (summarized in Tab. 4.3). The **Ref** stimuli were perceived significantly further compared to the **Diotic** ($p < 0.001$) and **Gen** ($p < 0.001$) stimuli. However, no significant difference was found between the **Ref** and both the **Gen/ER** ($p = 0.36$) and **ER60** ($p = 1.000$) stimuli. The **Diotic** stimuli was perceived significantly closer compared to the **ER60** ($p < 0.001$) and **Gen/ER** ($p < 0.001$), but not significantly different in terms of auditory distance compared to the **Gen** ($p = 1.000$) stimuli. The **ER60** stimuli were perceived further compared to the **Gen** ($p < 0.001$) but closer than the **Gen/ER** ($p < 0.001$) stimuli. Finally, the **Gen/ER** stimuli were perceived further than the **Gen** ($p < 0.001$) stimuli.

	Diotic	Gen	Gen/ER	ER60
Gen	1.000			
Gen/ER	< 0.001	< 0.001		
ER60	< 0.001	< 0.001	< 0.001	
Reference	< 0.001	< 0.001	0.359	1.000

Table 4.3: Perceived auditory distance (HI listeners, WDRC-B), Wilcoxon signed-rank tests results (significant p -values are in blue).

4.3.3 Auditory distance perception and PTA

Spearman's correlations were evaluated between the auditory distance estimations and PTA. All the HI listeners are included in the evaluation. The results are reported in Tab. 4.4.

Stim.	WDRC-A		WDRC-B	
	Spearman's ρ	p	Spearman's ρ	p
Ref	-0.354	0.005	-0.091	0.489
Diotic	0.515	<0.001	0.648	< 0.001
ER60	-0.161	0.218	0.081	0.538
Gen	0.163	0.215	0.218	0.113
Gen/ER	0.070	0.597	0.168	0.210

Table 4.4: Spearman's correlations coefficient (ρ) and associated p -values for the perceived auditory distance and the PTA at the best ear of HI listeners in the WDRC-A and WDRC-B conditions. Significant correlations are in blue.

In the HI listeners with the WDRC-A condition, a significant moderate correlation was found between the ratings of the **Diotic** stimuli and the PTA at the best ear (Spearman's $\rho = 0.515$). This means that in the WDRC-A condition, the HI listener with larger HL tended to perceive the **Diotic** stimuli further away. In this group, a small (Spearman's $\rho = -0.354$) but significant correlation was also found between the auditory distance of the **Reference** and the PTA at best ear. This shows a small tendency of HI listeners with larger HL to perceive the **Reference** closer. No correlation was found between the PTA and the perceived auditory distance for the **ER60**, **Gen** and **Gen/ER** stimuli.

With the WDRC-B condition, a significant moderate correlation was found between the PTA at best ear and the perceived distance of the **Diotic** stimuli (Spearman's $\rho = 0.648$). No correlation was found between the PTA and the perceived auditory distance for the **Ref**, **ER60**, **Gen** and **Gen/ER** stimuli. In particular contrarily to the WDRC-A condition, the ratings of the **Ref** stimuli were not correlated with PTA at best ear.

The auditory distance perception as a function of the PTA at best ear for the stimuli with significant correlations are pictured in Fig. 4.4.

4.4 Discussion

4.4.1 Perceived auditory distance in NH and HI listeners

In NH listeners, the perception of auditory distance was as expected: the reference stimuli were perceived at the location of the Loudspeaker 3, and the diotic stimuli were perceived in the center of the head. Most HI listeners were also able to perceive differences in perceived auditory distance, and a similar tendency was observed in the ratings. Nevertheless, the reference

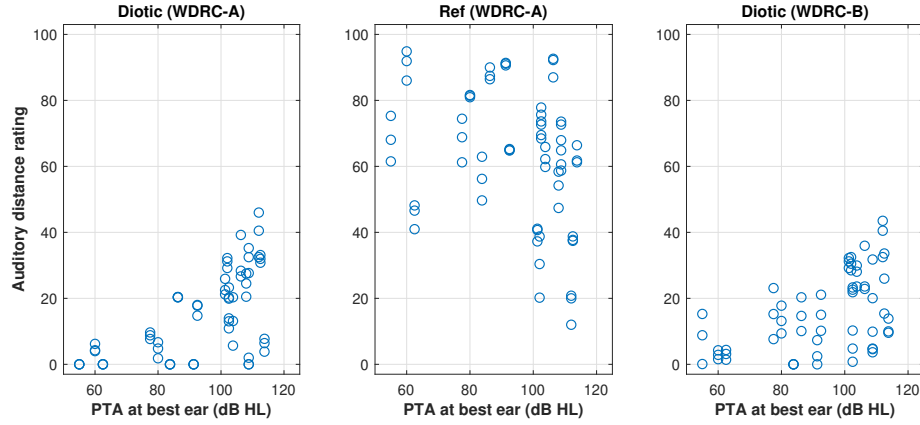


Figure 4.4: Perceived auditory distance as a function of the PTA at best ear for the three stimuli with significant correlation.

was generally perceived closer than in NH listeners and the diotic stimuli were rarely perceived as completely inside the head. Thus, HI listeners experienced a contracted perception of auditory distance, in accordance with the results obtained with mild-to-moderate HI listeners in [18]. In general, the results also revealed that moderate-to-highly-profound HI listeners reported less homogeneous results compared to NH listeners. This is in agreement with the conclusions drawn in [130] for mild-to-moderate HI listeners. Moreover, for HI listeners, a higher degree of HL was associated with increased distance perception of the diotic anchor in both WDRC conditions, and a decrease in the perceived distance of the reference when the WDRC was preceding the spatial processing. This suggest that the observed compression of the perception in distance might be more pronounced with stronger HL.

In NH listeners, the stimuli including ERs were perceived further compared to the stimuli using anechoic spatialization. This confirms that ERs can be sufficient to provide externalization [13] and auditory distance [23]. In HI listeners, the stimuli with ERs were also perceived further away than the diotic and anechoic spatialization. There was no significant difference in the perceived distance between the reference and the convolution with truncated BRIRs in HI listeners, while NH listeners rated the reference to be perceived further. This is in accordance with [3], in which the difficulty to resort to DRR to estimate auditory distance for HI listeners was pointed out. In the present study, as all stimuli were level-equalized, the HI listener could not take advantage of the absolute level cue.

Interestingly, for NH listeners, no difference was perceived in auditory distance between the stimuli using non-individualized HRTFs with superimposed generic ERs, compared to the stimuli made with truncated individualized BRIRs. It has been reported in several studies that non-individualized HRTFs can lead to a degradation of the perception of externalization [52, 87]. However, studies also suggested that the use of non-individualized BRIRs might not affect externalization [97] nor auditory distance perception [176, 175]. It can be hypothesized that the availability of the visual cue might have attenuated the difference between those two

stimuli for certain listeners. Indeed, the visual cue provided by the loudspeakers might have generated a visual capture effect for both NH and HI listeners. Moreover, the distributions in NH listeners suggest that the truncated individualized BRIRs was perceived with more variation compared to the other stimuli. It can be hypothesized that the BRIR truncation might have been perceived as less natural as it has no physical equivalent as suggested in [64]. This could have affected the distance ratings. In the same study, the authors suggest that it is possible to reduce the ER time with a decaying gain instead which may sound more natural. This could potentially yield different ratings.

4.4.2 Considerations on the RM system for HA application

The addition of ERs picked up from the HA microphones to a simple anechoic and generic spatialization method was sufficient to significantly enhance the perceived distance in both HI and NH listeners. This is of special interest in the context of RM systems, for which an anechoic spatialization method with non-individualized HRTFs has been already proposed in [44] (described in this thesis in Section 2.3).

With a binary approach to externalization, a threshold can be defined at 25% for which stimuli rated above are considered as externalized. The superimposition of ERs to the spatialized anechoic direct sound enabled to increase the externalization rate from 50% to 90% in the WDRC-A case, and from 51% to 92% WDRC-B case compared to the the spatialized anechoic direct sound alone. The methods investigated in Chapter 3, aiming at superimposing ERs to this direct sound, could therefore solve the problem of missing realism reported by users by substantially improving externalization. Similar conclusions were found in [81] for mild-to-moderate HI listeners and artificial ERs.

WDRC, because of its non-linear characteristic, could be expected to alter some important cues for auditory distance estimation such as the DRR or ILD. However, the present study does not show a significant degradation in perceived auditory distance due to the WDRC, in agreement with the results found by Akeroyd in [3]. When the WDRC was applied after the spatialization (WDRC-A), the correlation analysis suggest that the HI listeners with lower HL had a better estimation of the distance of the reference stimuli, suggesting that they could take advantage of the WDRC processing in their evaluation. This was not the case when the WDRC was applied before the binaural synthesis (WDRC-B), which yielded a reduction of the perceived distance of the reference for certain listeners. This suggests that the WDRC after spatialization, which optimizes the audibility of the signal, is more adequate for the HI listeners with severe-to-profound HL, than an accurate rendering of the binaural cues as provided in the case of the WDRC placed before the spatialization. As hypothesized by Akeroyd [3], this could be explained by the accustomization of the HI listeners to the gain processing provided by their personal HA settings.

The algorithm combining the superimposition of ERs as described in Section 3.3, and the direct sound spatialized with generic non-individualized HRTFs as described in Section 2.3

helped the HI listeners to perceive auditory distance. The ERs provided by the algorithm in this study were extracted from recordings in the same room with a KEMAR manikin, as HI listeners were not wearing their HAs during the test. This means that the ERs were not individualized contrarily to the targeted application in which the ERs would be extracted from the signal picked up by the HAs' microphones. However, despite being individualized, this signal might include some residual noise from the environment. Hence, the perception of distance could be different in those conditions.

4.4.3 Limitations of the study

In this study, the sound were played back through headphones rather than with the HI listeners' own HAs. Their personal amplification was simulated and the signal was low-passed with a cut-off frequency of 6.5 kHz to yield a similar bandwidth as found on most HAs. However, the sound they experienced was significantly different from the sound they are used to with their own devices. At the end of the experiment, twelve listeners reported that they preferred the sound rendered by their own HAs over the headphones rendering. This could be due in part to the acoustic coupling difference between the open headphones used for the experiments compared to the in-ear coupling usually obtained with the receivers of HAs.

The study presented in this chapter only included one azimuth direction for the sound sources, which were all synthesized to be coming from 30°. The laterality of the sound source is known to affect auditory externalization [81, 20]. This angle was chosen to allow the three loudspeakers to be visible from the listener's position, and as a compromise between a completely frontal angle (0°), which is the more challenging to externalize, and a too lateral angle that might be already well externalized. The laterality of the sound source also affects auditory distance perception. In particular, Brungart et al. [25] found that distance evaluations for lateral sounds were more accurate than for frontal sounds. Hence, further investigations on this topic could include several azimuths for the auditory distance estimation.

In this experiment, the visual cues were always available to the listener. Vision affects the perception of distance and can help listeners to estimate auditory distance [31, 177]. As previously discussed the presence of visual references can also steer the perceived distance to those locations, in particular with the effect of visual capture [121]. Hence, it is likely that the listeners in this experiment took advantage of the visual cues in their distance evaluation and that it influenced their ratings. It would be interesting to conduct another study with similar stimuli and panel of listeners, that would rather focus on auditory externalization with eyes closed, as done in [97].

4.5 Conclusion and perspectives

This study showed that moderate-to-highly-profound HI listeners have a contracted perception of auditory distance compared to NH listeners, as was previously observed in mild-

to-moderate HI listeners. Previous studies also showed that mild-to-moderate HI listeners perceive auditory distance less homogeneously, which was found in this study as well.

The addition of ERs to a generic and anechoic spatialization was shown to substantially improve the perception of auditory distance, and increase the externalization rates, for moderate-to-highly-profound HI aided listeners. This is a promising conclusion in the context of RM systems, for which anechoic spatialization was already addressed [44]. In addition, the WDRC processing did not prevent the perception of auditory distance for most aided severe-to-profound HI listeners. Their performance was closer to NH listeners when the WDRC was placed after the spatialization, which could be explained as they are accustomed to their own gain compression settings.

Testing different azimuths, in particular more frontal (e.g. 0°) and more lateral (e.g. $>60^\circ$), as well as conditions where the visual cue is not available, could provide additional knowledge in how HI listeners perceive auditory distance. Optimization of speech intelligibility remains the main purpose of RM systems. It is also important that the listener has access to sounds of their surrounding, so as to allow spatial awareness. Hence, subsequent studies should include the potential effect of additional ERs on speech intelligibility and spatial awareness. This is addressed in the next chapter.

5 Auditory distance perception of a RM signal using low computational cost algorithms

In some real-time binaural sound applications with RM systems, technical constraints might not allow for processing the exact spatialization with complete BRIRs that is necessary for the correct perception of auditory distance. This is the case with miniaturized hearables. This potentially results in the perception of virtual sources inside the head rather than externalized, and thus not perceived at the appropriate distance. The study presented in this chapter^I aims to assess three sound spatialization strategies for improving the perception of externalization. While it can be expected that an externalized sound image of the RM signal should provide a more natural perception of the auditory scene to the listener, it is possible that the perception of other sounds of the environment could be affected. Being able to be aware of surroundings in space, referred to as "spatial awareness" in this thesis, is a crucial feature in this type of application. Consequently, the goal of the tested algorithms is to provide externalization and distance perception in the presence of a visual cue, while preserving spatial awareness. Those algorithms are designed to be implementable on wearable devices, using low computational power and little memory. These are based on the superimposition of early reflections (ERs) with methods described in Chapter 3 to a generic direct sound spatialized with non-individual HRTFs as described in Section 2.3. These algorithms were evaluated with NH listeners, in terms of distance perception as well as spatial awareness of a surrounding event.

5.1 Context and motivation

While the previous chapter addressed the HA use case, hearables can also be used in combination with a RM system. As in HA, the main goal of the RM process is usually to optimize speech intelligibility in a challenging auditory situation such as a noisy environment or when the distance between the speaker and the listener is large. In most cases, the voice of a speaker

^IThis chapter is an extended version of the work presented in [67]

is transmitted directly and diotically to the hearables. Therefore the perceived location of the speech source does not match its physical location in the environment.

In the context of HAs, where RM systems are often used, sound localization and spatialization have been topics of interest in several studies. In [145], it was shown that localization performance was better with lower gain on the RM signal. Methods to estimate the localization of the sound source have been proposed in [44, 55]. In particular, while successfully providing sound localization to the listeners, the spatialization method proposed by Courtois et al. [44] and described in Section 2.3 usually fails to provide an externalized sound image, i.e. sounds are perceived inside the head rather than surrounding the listener. In [71, 23], the authors have shown that ERs contribute largely to auditory distance perception and sound externalization. In the context of RM systems, several studies have demonstrated that the superimposition of ERs to the RM signal can significantly improve the perception of auditory distance. Those studies used ERs that were either extracted from the hearing device microphones signals as in Chapter 4 [41], or artificially synthesized [80].

It is expected that rendering an externalized speech signal from the speaker wearing the RM would help getting a more natural perception of the environment. However, it can be hypothesized that it might have consequences on how other surrounding sounds are perceived. Indeed, when only the speech source is played to the listener, they may not be able to hear other sounds of the environment. The listener's ability to be aware of themselves and of their surroundings in space is referred to as "spatial awareness" in this thesis. This is a crucial ability, as many situations require the listener to be informed of how people and objects move in the environment. If only the RM signal is played back to the listener, they may feel dissociated from their surroundings and would not be able to hear other information such as another speaker. Safety concerns are also raised, as competing sounds may not be heard by the listener, e.g. an alarm or any unexpected event.

In this study, three different algorithms devoted to hearables with RM systems were subjectively evaluated by a panel of NH listeners. The first algorithm is the baseline algorithm, which is the RM signal to which is superimposed the ambient signal from the hearables microphones, with a certain gain. This method will be referred to as **RMM** for "Remote Microphone + ambient Microphone". Two other algorithms are introduced, based on methods described in Chapter 3. Their goal is to introduce ERs in order to improve externalization and the perception of auditory distance. The first of those two methods uses the coherence-based method from Section 3.3, and is referred as **ERC** for "Early Reflections extraction and Cleaning". The second method is based on partitioned convolution as described in Section 3.4, and is referred to as **PConv** in this chapter.

As mentioned, an important factor in the context of hearables is the ability to be aware of oneself and surroundings in space [144]. Hence, a certain amount of ambient sound has to be reproduced in combination with the ERs in the algorithms. The performance of the three algorithms was evaluated both in terms of auditory distance perception and spatial awareness.

5.2 Description of the algorithms

This section aims at providing a brief description of the three algorithms evaluated in this study, which are based on some of the methods described in Chapter 3.

5.2.1 Remote Microphone + ambient Microphone (RMM)

This is the baseline implementation, as depicted in Fig. 5.1.

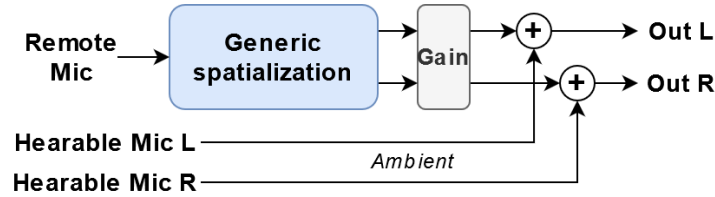


Figure 5.1: Block diagram of the **RMM** algorithm.

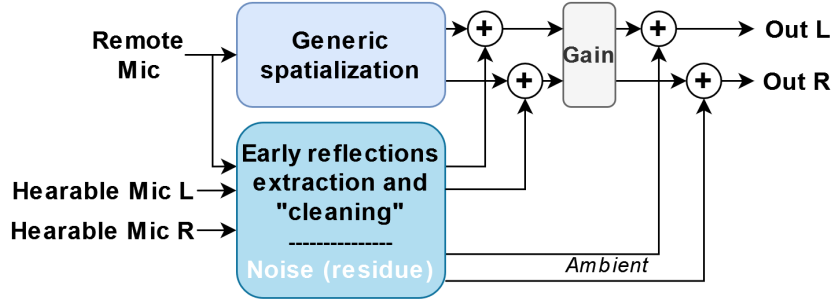
This algorithm is composed first of a spatialized direct sound to which is superimposed the raw signal captured by the hearables' microphones. The method used to spatialize the direct sound with non-individualized HRTFs was introduced by Courtois et al. [44] and is described in this thesis in Section 2.3. As a reminder, the spatialization uses minimum-phase 128-sample (5.8 ms) non-individualized HRIRs, and the ITD is simulated by a pure delay. The input from the hearable microphones is superimposed with a certain gain to the main output. This gain depends on the estimated signal-to-noise ratio (SNR): high gain for high SNR and conversely. This implementation is designed for achieving an optimal speech intelligibility.

5.2.2 Early Reflections extraction and Cleaning (ERC)

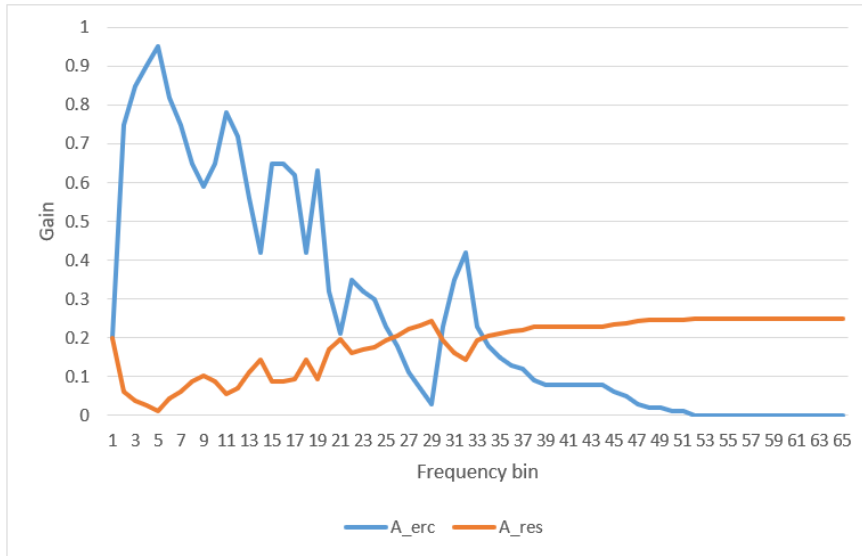
The first goal of this algorithm, which principle is depicted in Fig. 5.2, is to add ERs picked up by the hearable microphone signals to the aforementioned generic spatialization. This is done using the coherence-based algorithm described in Chapter 2, which is designed to clean the ambient noise and extract ERs from the hearable microphone signals.

Limiting the algorithm to the addition of the extracted ERs would not be optimal for spatial awareness. The remaining residual noise should only include very weak and distorted ambient information. Hence, a complementary filter is introduced, to bring back some of the ambient sound (called here: residue) with a lower gain. Based on the filter computed by the **ERC** module, a new filter is calculated as:

$$A_{res}(\omega) = (1 - A_{erc}(\omega)) * G_{res}, \quad (5.1)$$

Figure 5.2: Block diagram of the **ERC** algorithm.

where A_{res} is a coefficient of the residue filter, A_{erc} is a coefficient of the **ERC** filter, ω is the frequency index and G_{res} is the general gain (between 0 and 1) applied to the residue part. A gain of 1 would lead to an original restitution of the signal (unprocessed). The frequency resolution used is the 65 first frequency bins. This enables to easily tune the algorithm. An example of filter values with its complementary residue filter at some random frame (frequency domain) is depicted in Fig. 5.3. G_{res} was arbitrarily fixed to 0.25 for the purpose of the example.

Figure 5.3: Example of some complementary filter gains at a random frame (frequency bins), A_{erc} and A_{res} , with $G_{res} = 0.25$.

The RM signal, the ERs and the ambient sound can be tuned with independent gains, allowing reaching a trade-off between speech intelligibility, auditory distance perception and spatial awareness.

5.2.3 Partitioned Convolution (PConv)

This algorithm consists in introducing synthesized ERs as depicted in Fig. 5.4. The ERs are generated by using a uniform partitioned convolution algorithm [62], as implemented in [157]. The method was described in Section 3.4. As a reminder, it consists in partitioning an impulse response into a series of smaller blocks. Those blocks can be thought of as independent impulse responses or subfilters that can be used in parallel in a real-time processing, using FFTs and re-combined with appropriate delays.

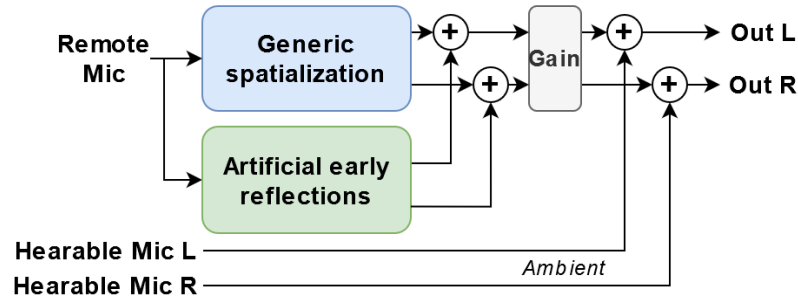


Figure 5.4: Block diagram of the **PConv** algorithm.

The RM signal was convolved with a 50-ms truncated pair of binaural room impulse responses (BRIRs). BRIRs measured in a classroom were used for this study, regardless of the room used for the experiment. Indeed, in a real usecase, BRIRs cannot be measured in every room the hearable user walks in. A pair of BRIRs corresponding to a source at 0° and a distance of 2 m was used in the experiments. This limits the required computational cost and memory usage. To enable spatial awareness, an additional signal from the hearable microphones is superimposed. This implementation allows tuning independently the gain of the main speech (RM signal), the ERs and the ambient sound.

5.3 Experiment 1: Auditory distance perception

The goal of this experiment was to study if the additional ERs introduced in the **ERC** and **PConv** algorithms improve the perception of auditory distance.

5.3.1 Setup

The experimental setup was installed in a listening room (volume = 125 m^3 , $RT_{60} = 0.17 \text{ s}$). Three loudspeakers (Genelec 1029A), were aligned along the azimuth of 30° on the right side of the listener, at a distance of 67 cm, 113 cm and 200 cm respectively from the listener's position, as depicted in Fig. 5.5. They were numbered from 1 to 3. During the test, the listener was sitting at the center of the room and had their head immobilized by means of a chin rest.

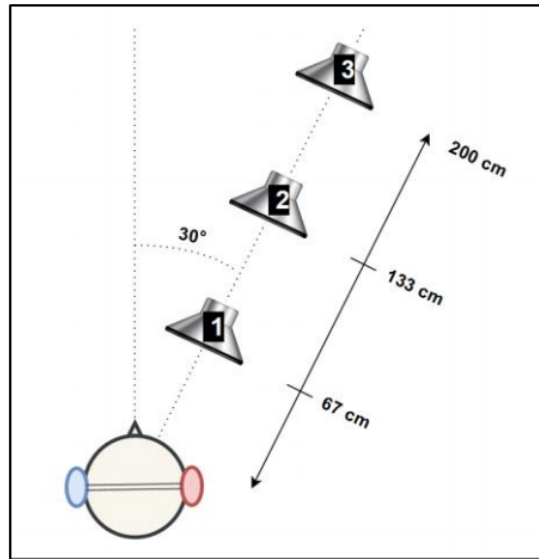


Figure 5.5: Schematic representation of the setup for Experiment 1, mounted in a listening room.

5.3.2 Stimuli

Five stimuli were evaluated in this experiment. One was a diotic reproduction, one was the reference, convolved with the individual BRIRs of loudspeaker 3, and the three remaining stimuli corresponded to the **RMM**, **ERC**, and **PConv** algorithms. The stimuli consisted of 15-second speech sequences, obtained by concatenating short phonetically-balanced random sentences from the French HINT database^{II}.

In the baseline algorithm (**RMM**), the spatialized speech was set 10 dB louder than the ambient sound. When adding ERs to the direct signal, the direct-to-reverberant ratio (DRR) is an important parameter regarding distance perception [178]. For this part of the experiment, the DRR was fixed to 5 dB for the **ERC** and **PConv** algorithms based on preliminary informal tests. The **PConv** algorithm uses a 50-ms truncated version of the BRIRs in the listening room (the total length is 360 ms). The **ERC** algorithm was tuned to provide the same ER time. The speech counterpart, made of the direct sound and ERs for the **ERC** and **PConv** algorithms, and the direct sound only for the **RMM** algorithm, were level-equalized between stimuli. This was to ensure that level would not be a determining cue in the auditory distance evaluation.

Similarly as in the experiment presented in Chapter 4, the input for the **ERC** algorithm were generated from recordings made beforehand with a KEMAR manikin at the same position as the listener. Hence, the ERs provided by the algorithm were not individualized contrarily to a real case where the signal would be picked up by the hearable's microphones. All stimuli were presented at a level of 65 dB SPL. For all participants, the stimuli were compensated with

^{II} Collège National d'Audioprothèse 2006

their individual headphone-to-ear impulse responses (HPIRs) and low-pass filtered (cut-off frequency = 6.5 kHz).

5.3.3 Procedure

For each participant, individual BRIRs corresponding to the three loudspeakers, as well as HPIRs were measured using a pair of in-the-ear binaural microphones (Sound Professionals MS-TFB-2). During the experiment, all stimuli were played through a pair of open headphones (Audeze LCD-2C) driven by a headphones amplifier (Lake People HPA RS 02).

Using a MUSHRA-like^{III} graphical user interface (GUI) displayed on a touch-pad, the participants were asked to rate the auditory distance perceived for the five stimuli. The stimuli were available simultaneously and it was possible for the listener to cycle through the stimuli and listen to them as many times as they wanted. They were instructed to answer the question: "How far do you perceive each stimulus from your position?". They used a continuous scale displayed as a slider with the following markers: Center of the head (0), Boundary of the head (20), At Loudspeaker 1 (40), At Loudspeaker 2 (60), At Loudspeaker 3 (80) and Further than Loudspeaker 3 (80 to 100). The task was repeated over 4 runs.

The experiment was preceded by a short training phase in which each listener could listen to speech sound examples processed with their individual BRIRs of loudspeakers 1, 2 and 3. This helped the participants to get accustomed with the task and gave them an a priori knowledge of the reproduction level used in the experiment. This also served to ensure that their auditory impression matched the visual location of the loudspeakers.

The allocation of the different tested stimuli to the "play" buttons was randomized across runs to avoid any bias due to the order of presentation of the various stimuli. It was also randomized across participants, i.e. every participant experienced different allocations during the experiments.

5.3.4 Results

Auditory distance ratings are reported in Fig. 5.6. For each listener, only the last three runs are considered. The first run was considered as training and therefore not taken into account. Three subjects (out of 25) were not retained in the results because they did not perceive the reference as externalized, or perceived the diotic stimuli as the furthest sound.

^{III}ITU-R recommendation BS.1534-3, 2015.

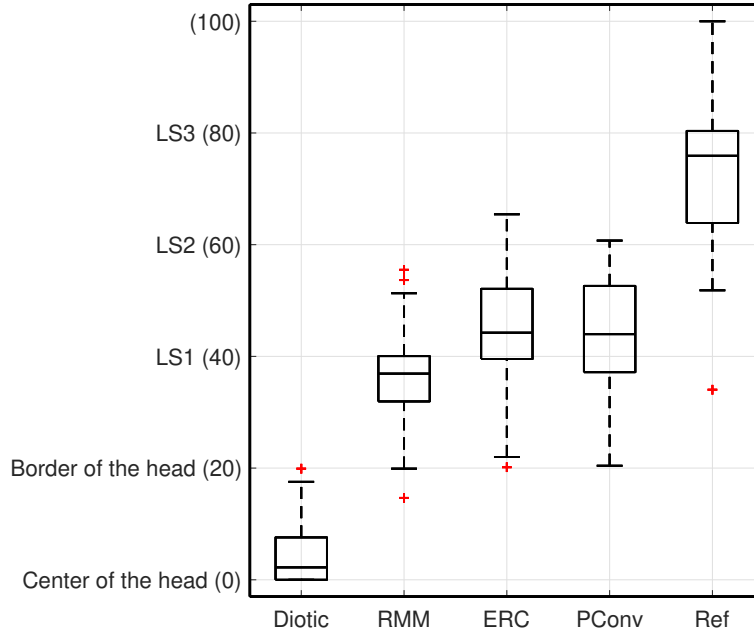


Figure 5.6: Auditory distance estimations evaluated on a continuous scale with the following markers: Inside of the head (0), Border of the head (20), At Loudspeaker (LS) 1 (40), At LS 2 (60), At LS 3 (80) and Further than LS 3 (80 to 100).

A Friedman test among repeated measures revealed significant differences in the perceived auditory distance between the five stimuli, $\chi^2(4) = 226.65$, $p < 0.001$. A large effect size was found using Kendal's W ($W = 0.859$). *Post-hoc* Wilcoxon signed-rank tests with Bonferroni correction were conducted. The results are reported in Tab. 5.1. The reference was perceived

	Diotic	RMM	ERC	PConv
RMM	< 0.001			
ERC	< 0.001	0.001		
PConv	< 0.001	0.001	1.000	
Reference	< 0.001	< 0.001	< 0.001	< 0.001

Table 5.1: Perceived auditory distance, Wilcoxon signed-rank tests results (significant p -values are in blue).

significantly further than the diotic, **RMM**, **ERC** and **PConv** stimuli. The diotic stimuli were perceived significantly closer than the reference, **RMM**, **ERC** and **PConv** stimuli. No significant difference was observed in the distance evaluation between the **ERC** and **PConv** stimuli. Both the **ERC** and **PConv** stimuli were perceived as further away compared to the **RMM** stimuli.

5.3.5 Discussion

It should be noted that the partitioned convolution-based algorithm used BRIRs from a classroom with a significantly different and denser reflection content than the ones of the listening room in which the test was conducted. It has been shown that room congruence contributes to auditory externalization [170]. Nevertheless, the BRIRs were truncated to 50 ms in the case of the partitioned convolution-based algorithm, and tuned to yield the same ERs time in the case of the coherence-based algorithm. Thus, this could have reduced the perceptual difference between the two stimuli.

Additionally, the BRIRs used to generate the ERs in the partitioned convolution-based algorithm corresponded to a source located at 0° . Despite this, the partitioned convolution-based algorithm gave a similar perceived distance compared to the coherence-based algorithm that used the BRIRs of the listening room and reflections corresponding to a source at 30° . This is promising, as this means that it might not be necessary to store too many BRIRs in the hearables' device memory to provide externalization and cover the localization in the horizontal plan. Informal listening sessions suggest that the precedence effect might allow to use the 0° BRIRs for rather frontal sources, i.e. below around $\pm 40^\circ$, and still benefit from the improvement in auditory distance perception. However for larger azimuths, in particular above around $\pm 60^\circ$, an image split might be perceived and could lead to a collapse in perceived auditory distance. However, this was only observed informally, and was not addressed in this study. This would be interesting to investigate in further researches.

This is in agreement with Begault [12] who found that it was not necessary to provide accurate information in reverberation to obtain improvement in externalization, which could be obtained with synthetic reverberation and non-individualized HRTFs. However, this degraded the performance in localization. These observations are interesting to compare with the results of Zahorik & al. [179], which suggested that the accurate information in the reflections might not be necessary to yield an accurate performance in localization, including distance judgements. However, this was observed with individualized HRTFs in the latter. This is also consistent with the precedence effect.

It is also possible that any small difference was compensated by the availability of the visual cue and the effect of visual capture. Even though the participants were clearly instructed that they could use the scale continuously and perceive sounds at intermediate positions between the loudspeakers, the visualisation of the loudspeaker might also have steered the perception of sound sources to the reference points of the scale. While vision has been shown to improve the accuracy of distance perception with real sources [177], this might be different in the case of the three algorithms tested in this experiment. Indeed, the listeners were only provided with a simplified spatialization with approximate cues for externalization and distance perception, which might encourage visual capture.

This study showed that the superimposition of ERs using the partitioned convolution-based algorithm or the coherence-based algorithm improves the perception of auditory distance

in NH listeners compared to the baseline algorithm. This could be expected as the baseline algorithm only allows to introduce reverberation with a lower gain.

5.4 Experiment 2: Spatial awareness

The second experiment aimed at studying if the **ERC** and **PConv** algorithms affect the ability of the listener to detect surrounding sound events, compared to the baseline **RMM** strategy.

5.4.1 Setup

The test took place in the same room and with the same hardware as in Experiment 1. A complex auditory situation was presented over headphones.

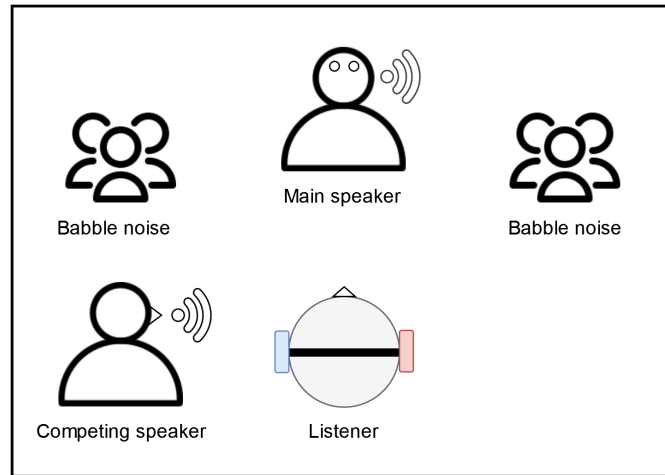


Figure 5.7: Auditory scene for Experiment 2, simulated over headphones using binaural synthesis.

5.4.2 Stimuli

The stimuli consisted of complex auditory scenes including a main speaker in the front (0°), conversational noise with a SNR of 0 dB at the listener's position, and a competing speaker located laterally compared to the listener (90°). This auditory scene is schematically depicted in Fig. 5.7. Binaural synthesis was used to simulate the main and competing speaker, using BRIRs for a source at 2 m, 0° in a listening room and 2 m, 90° in an anechoic room, respectively. The speech sequences for those two speakers were generated from short sentences extracted from the French HINT database. Binaural recordings from an internal database were used for the babble noise. Those recordings had been made with a KEMAR manikin at the position of the listener, and four loudspeakers in the corners of the same listening room, which were playing babble noise. The stimuli were processed using the full implementation of the **PConv** and **ERC** algorithms as well as the **RMM** strategy. A low SNR was chosen in order to ensure that

the task would be challenging for NH listeners. The **ERC** and the **PConv** algorithms were each presented with two settings: either with a DRR of 5 dB or 2 dB. The **RMM** algorithm is used with the same setting as in Experiment 1: the RM signal was amplified by 10 dB compared to the ambient sound. Those algorithms are denoted: **ERC5**, **ERC2**, **PConv5**, **PConv2** and **RMM** respectively. All stimuli were presented at a level of 65 dB SPL, and the spatialized speech was level-equalized between stimuli as in Experiment 1.

5.4.3 Procedure

The listener was asked to perform a dual-task. The listener had to repeat one of the short consecutive sentences pronounced by the main speaker (the one followed by a “beep”). Simultaneously, they had to pay attention to the continuous speech of the competing speaker and click a button on the screen when one segment was presented backwards (time-reversed). As shown in the example depicted in Fig. 5.8, the time-reversed segment was necessarily overlapping with the sentences of the main speaker. For any sequence, the sentence to repeat could be anywhere between the 3rd and the 6th sentence. The listener was not provided with these clues. The time-reversed segment in the competing speaker could happen at any time before the sentence to repeat. The reaction time was measured as the time it took for the participant

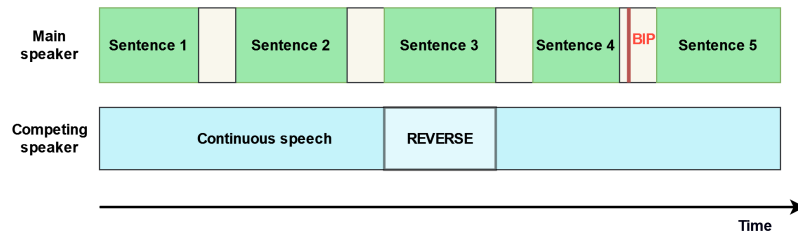


Figure 5.8: Example of stimuli structure in time, for the main and competing speakers (Experiment 2).

to click, from the moment when the time-reversed section started. The first task was described to them as the main (or priority) task to avoid a situation where the listener would focus only on the competing speaker which would have not been relevant for the intended investigation of this experiment. The test included 10 repetitions for each algorithm and parameterization. Thus, the listeners had to evaluate a total of 50 stimuli. The order of the sentences and the order of the algorithms applied to each sentence were randomized across participants. A short training with 5 runs preceded the test in order to familiarize the listener with the task and the stimuli.

5.4.4 Results

The number of times a time-reverse speech signal was correctly detected was analyzed. Detection before it was presented or after the sentence to repeat was considered as a non-detection.

Boxplots for the scores of the various algorithms are reported in Fig. 5.9.

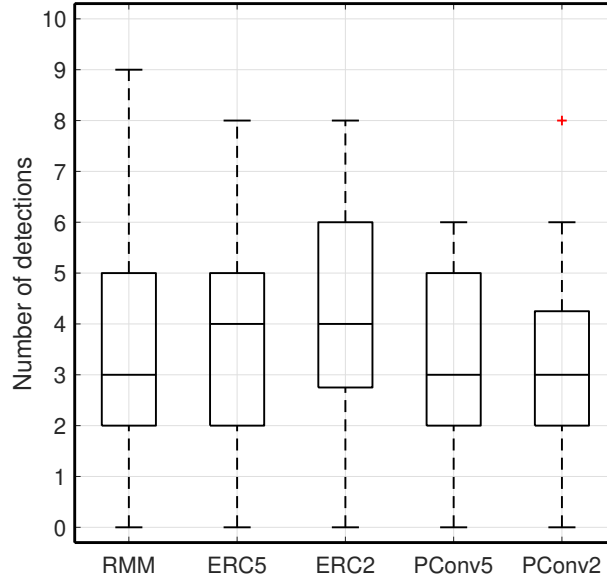


Figure 5.9: Scores for the number of detections of the "reverse" speech in the competing speaker.

One of the 25 subjects, who systematically detected the speech-reverse and was removed from the analysis to ensure normality of the distribution. Normality was assessed using a Shapiro-Wilk test ($p = 0.075; 0.173; 0.093; 0.119; 0.247$ for **RMM**, **ERC5**, **ERC2**, **PConv5**, **PConv2**, respectively) indicating no significant deviation from a normal distribution. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(4) = 5.781, p = 0.762$. Using a repeated measures ANOVA, no significant difference was found in the scores of the various algorithms regarding the number of time-reverse speech detections, $F(4, 92) = 1.571, p = 0.189$.

The results of the reaction time are also reported in Fig. 5.10. Normality was checked using a Shapiro-Wilk test ($p = 0.151; 0.817; 0.236; 0.659; 0.103$ for **RMM**, **ERC5**, **ERC2**, **PConv5**, **PConv2**, respectively). The Mauchly's Test of Sphericity indicates that sphericity was not violated, $\chi^2(4) = 3.121, p = 0.237$. No difference was found between the stimuli with a repeated measures ANOVA, $F(4, 92) = 1.125, p = 0.348$. This indicates that the reaction time was not different between the different stimuli.

The number of sentences presented between sentences to repeat ranged from 2 to 5 and was balanced across sequences for every participant. This could yield a "tension" effect affecting the performance in detection. To assess the possibility of such an effect, the repeated measure ANOVA included the interaction between the tension (number of filling sentences) and the detection scores. No significant interaction was found, $F(4, 192) = 0.726, p = 0.575$.

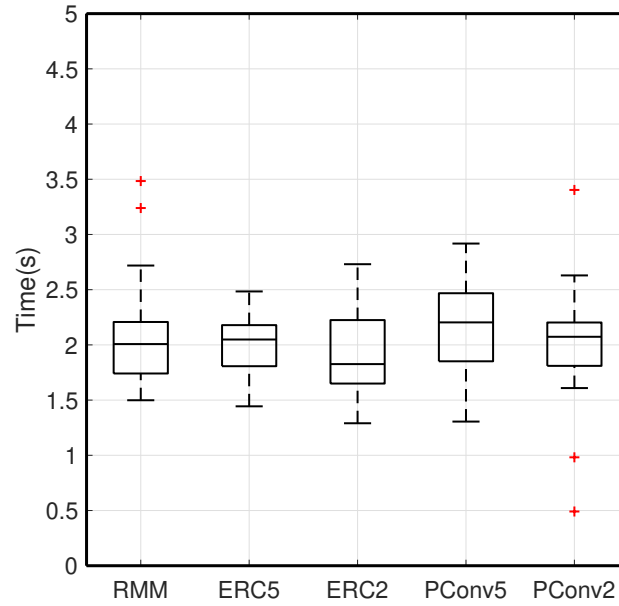


Figure 5.10: Reaction time in the speech reverse detection task.

The speech intelligibility task was intended only as a control task, which had a high success rate for every participant and algorithm. The rare occurrences of failure in repeating the target sentence were orally reported by listeners as moments when they tried to focus all their attention on the competing speaker. Hence results are only reported in appendix A.2.

5.4.5 Discussion

This second experiment was conducted with a low SNR of 0 dB. This value was chosen after informal pre-tests, so that the task would be challenging enough for NH listeners and avoid ceiling or floor effects in the scores. It appears that the task was sufficiently challenging while not being too simple, as only one subject did detect the speech-reverses systematically and no subject was unable to detect the reverses. The results suggest that the two proposed algorithms that introduce ERs in the RM signal might not affect the ability of NH listeners to detect surrounding auditory events compared to the baseline algorithm in a challenging auditory scenario. However, the proposed test design remains fairly experimental at this point. It could be argued that the listener's ability to detect speech-reverses could have been influenced by other factors, such as the chosen strategy to complete the task. For example, certain subjects orally reported that they tried to focus on the competing speaker during the short silences between sentences of the main speaker. However this particular strategy had been anticipated as the reversed segment was always overlapping a sentence of the main speaker.

The study focused on the case of hearables and NH listeners. A similar study could be conducted with aided HI listeners. It can be expected that the task might be much more challenging for this population, even with higher SNRs. A speech intelligibility test could be conducted with HI listeners and the algorithms that introduce ERs in the spatial rendering. Indeed, studies have suggested that ERs might contribute to improve speech intelligibility [107], even possibly for HI listeners [6]. However the effect might be dependent on the noise level [150].

5.5 Conclusion and perspectives

In this work, two algorithms were proposed to improve the perception of auditory distance in hearables with RM, while preserving spatial awareness and speech intelligibility. The algorithms are based on the superimposition of ERs from either the coherence-based method described in Section 3.3 or from the partitioned convolution-based method described in Section 3.4, to a direct sound spatialized with generic HRTFs as proposed in [44] and described in Section 2.3. Both algorithms include a strategy to provide some sounds from the environment, and allow to independently tune the gain applied to the RM signal, the ERs and the ambient sound. This enables to look for a trade-off between speech intelligibility, auditory distance perception and spatial awareness. Future works could investigate the tuning of the algorithms to reach this trade-off, as a function of the SNR in particular.

Two experiments were conducted and showed that a significant improvement in the perceived auditory distance was obtained with both algorithms aiming at providing ERs compared to the baseline algorithm which served as a starting point for this thesis. The results of the second experiment suggest that the proposed DSP strategies do not affect spatial awareness for NH listeners.

To follow up, similar tests could be conducted with HI subjects, to assess the generalization of the results to the specific case of HI listeners wearing HAs. Those tests could focus on speech intelligibility further, as ERs are known to contribute to speech intelligibility.

In the experiments described in this chapter as well as in Chapter 4.3, the listeners kept their head still during play back. As discussed in Section 2.2.5, head movements, enabled by the availability of head-tracking in binaural reproduction, provide numerous perceptual advantages. In particular, externalization could be improved, in addition to an enhancement in plausibility and the resolution of front-back confusions. The next chapter is dedicated to the investigation of a head-tracking algorithm designed to be compatible with HAs and hearables.

6 Design of a head-tracking algorithm using two 3-axis accelerometers

As seen in Chapter 2, head-tracking can provide several perceptual advantages in the context of binaural synthesis. With the miniaturization of low-power accelerometers and gyroscopes, it is most likely that future HAs will be equipped with this type of sensors. In particular, accelerometers are already being implemented in HAs for fall detection. This chapter^I describes the design and implementation of a head-tracking algorithm aiming at estimating the azimuth position and relying solely on two 3-axis accelerometers.

6.1 Motivation to include head-tracking in binaural communication devices

The introduction of head-tracking in the binaural rendering presents several advantages regarding auditory spatial perception. With the improved spatialization technique for RM systems, the listener might be able to localize the sound source. However, if the listener rotates the head to a certain azimuth, the simulated direction is shifted by the same angle and becomes erroneous. The first motivation to include head-tracking is the possibility to provide dynamic binaural synthesis, which is key to improve plausibility as well as reduce significantly the occurrence of front-back confusions as explained in Section 2.2.5. Consequently, the availability of head-tracking can improve the localization performance and accuracy by substantially reducing the occurrences of reversals when head movements can be performed [155, 128]. In addition, Begault et al. [13] found that head-tracking coupled with head movement improved the localization accuracy, in particular with a more pronounced effect in the case of non-individualized HRTFs. Head-tracking, especially when combined with large head movements, has been shown to substantially improve auditory externalization in several studies [20, 74, 98]. A detailed review of the literature is available in Section 7.1.2,

^IThis chapter is an extended version of the article [69].

hence it is not repeated here.

6.2 Equipment

6.2.1 Reference tracking

To obtain a reliable estimation of the Euler angles, a complete IMU/AHRS device (NGIMU^{II}) was used. The data are retrieved on Simulink either through USB COM Port communication or WiFi. The device allows to access to many real-time data. Notably, two of these devices were used at the early stage of the algorithm described in this section, for which both accelerometer data and Euler angles could be retrieved simultaneously. A 3D representation from the device's software is pictured in Fig. 6.1.

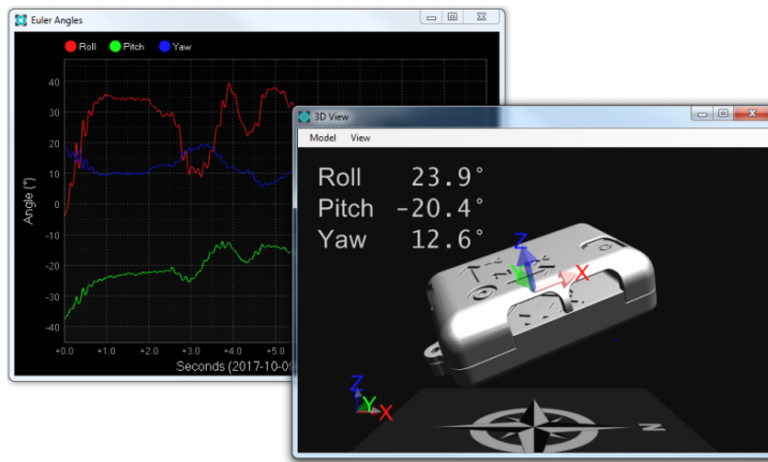


Figure 6.1: Head-tracker software with real-time Euler angles tracking and 3D visualization of the tracker.

6.2.2 Prototype

Sonova AG provided a prototype for the next steps of development. This prototype is pictured in Fig. 6.2, and consisted of two HAs with a DSP board fixed on a head-band. The accelerometer data could be retrieved through USB directly on a dedicated Simulink model. Additionally audio from an audio interface (RME Babyface Pro Fs) could be played directly through the HAs' receivers to render the binaural synthesis.

^{II}<https://x-io.co.uk/ngimu/>



Figure 6.2: Prototype HAs and audio interface.

6.2.3 Turntable

For the early stages of development of the algorithm, a turntable was acquired. The turntable was controlled via serial communication (USB COM port) and a dedicated Matlab script to generate rotations in the horizontal plane. The setup is schematized in Fig. 6.3.

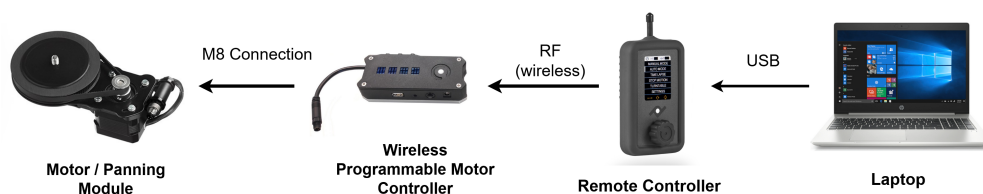


Figure 6.3: Setup for the control of the turntable.

First a circular plate was mounted on the motor, on which were placed two of the above mentioned IMU/AHRS sensor as pictured in Fig. 6.4. This allowed to make the first tests for the algorithm, in a situation where the axes of the sensors compared to the horizontal plane are known, as the devices are fixed horizontally on the table. Then, an artificial head (Schoeps KFM 6) was mounted on the motor, as pictured in Fig. 6.5 which allowed to test the condition with motions in the horizontal plane only, but the initial position of the sensor is unknown.

The early phases of the implementation are not described here for the sake of conciseness, but the use of this turntable was necessary for the preliminary trials of the algorithm under controlled conditions, before using measurements with real head motions from real participants.

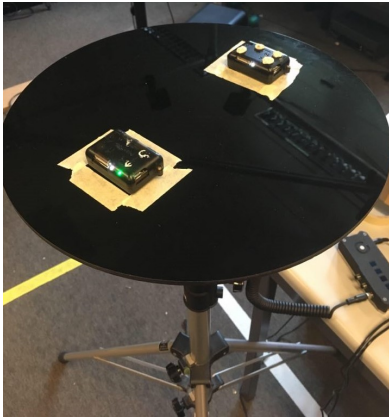


Figure 6.4: Two IMU/AHRS devices mounted on a turntable.



Figure 6.5: Artificial head equipped with the prototype, and mounted on the motor of the turntable.

6.3 Context of the implementation

Bilateral motion sensors worn on the head are typically found in hearables and HAs. Accelerometers have recently been integrated in those devices as they combine a miniature size and a low power consumption [19]. Multiple functionalities are developed to allow to detect an abrupt fall of a user, tracking sudden peaks in the vertical dimension of an accelerometer signal, and confirming the fall by comparing the left and right acceleration patterns [28, 139]. More complex systems are designed to evaluate the type of motion (immobile, walking, running) in order to steer audio processing algorithms aiming at enhancing speech intelligibility and/or spatial awareness [60, 160]. Voss et al. [161] showed significant improvement in speech understanding, sound localization as well as preference in HI listener with such steering approach. Being able to track the head motions of a user is of high interest since this would allow e.g. controlling the beam of a directional microphone system. In the context of spatial audio rendering with wearable devices, the knowledge of the orientation of the head of the user in the azimuth plan enables to achieve dynamic spatial rendering, i.e. the spatial filtering can be adapted to remain valid according to the listener's motions. Such processing can greatly improve the listener's experience of spatial audio rendering by increasing plausibility and reducing typical spatial perception artefacts such as front-back confusions [13]. Previously developed binaural spatialization algorithms aim at improving the perception of auditory externalization and distance in NH and HI listeners [41]. In particular they are optimized for low computational cost, to be compatible with wearable devices [67]. The addition of head-tracking in this context is motivated by the potential increase in auditory externalization demonstrated in several studies [74, 99].

The orientation of an object, and a human head in particular, can be determined by its Euler angles. Each angle, named the roll, pitch and yaw, corresponds to the rotation from an initial position, around the x-, y- and z-axis respectively, as pictured in Fig. 6.6.

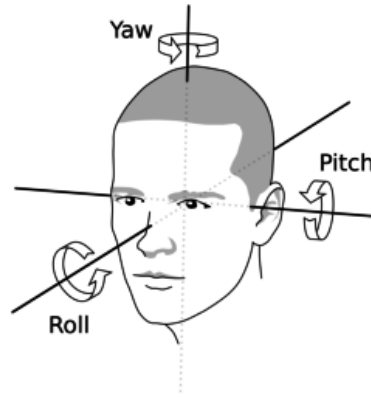


Figure 6.6: Human head angular motions, with the yaw, pitch and roll (image from [112]).

Euler angles are expressed as a sequence of those three rotations around the rigid body's local coordinate axes. One of the most common method for tracking each of these angles relies on inertial sensors, and is usually achieved using at least a combination of one or more three-axis accelerometers and one or more three-axis gyroscopes. The gyroscope measures the sensor's angular velocity, while the accelerometer measures the external specific force acting on the sensor, i.e. the earth's gravity and the sensor's acceleration. A magnetometer can also be used in combination, to help find orientation using the earth's magnetic field. Devices using such sensors are usually referred to as inertial measurement units (IMUs). However, IMUs are typically prone to integration drift. Errors due to the measurement of acceleration and angular velocity can progressively be integrated and the accumulation leads to larger errors in the estimation of the velocity and the angle. Some devices combine an IMU to an on-board processing system to run sensor fusion algorithms and aim at reducing this type of drift [180, 90]. Such devices are usually referred to as Attitude and Heading Reference Systems (AHRSs). Drift from the gyroscopes integration can then be compensated by the reference vectors: gravity and the earth magnetic field. With all these devices, eventually in combination with machine learning algorithms, a reliable drift-free instantaneous estimation of the head orientation can be obtained [56, 1, 21].

Nevertheless, not every device can be equipped with all the necessary sensors above mentioned and few solutions have been proposed to track head movements relying solely on bilateral accelerometers so far. Some devices, such as hearing aids (HAs) for which battery life and compactness are key factors, are only equipped with a single three-axis accelerometer. Such an accelerometer can be sufficient to determine the roll and pitch of the object to which it is attached, as detailed in [134], but does not allow estimating the yaw alone. An algorithm

aiming at estimating the yaw rate of a vehicle using two 1-axis accelerometers installed on the center-line of the vehicle was presented in [35]. Nevertheless the proposed algorithm relies on a model and assumptions associated with the vehicle tracking use case, which are not applicable for tracking the motions of a human head.

In [89], an algorithm is proposed to determine the yaw and roll angular rotation rates of a spinning rigid object using only two linear three-axis accelerometers. The method was found to achieve satisfactory performance when the rotational motions were rather short (< 2 s) and fast: $140^\circ/\text{s}$ to $1200^\circ/\text{s}$. The proposed method is dependant on the noise of the sensors and relies on a fixed threshold to prevent drifting due to this noise. It is also specified that the angular accelerations need to be lower than the accelerometer sensing range in order to not saturate the sensor. For example, in [89], this corresponded to $1200^\circ/\text{s}$ with the accelerometers used. In the use case of head motions typical head motions rarely exceed peak values of $350^\circ/\text{s}$ in the case of healthy subjects [142].

This chapter describes a real-time method aiming at estimating the relative yaw position of a human head using only two 3-axis accelerometers such as the ones found on wearable devices. Such solution using only accelerometers may provide significant power consumption savings over solutions relying on the use of a gyroscope.

6.4 Description of the algorithm

This section describes the algorithm used for the yaw position of a human head using two 3-axis accelerometers. The base concept of the algorithm described here is directly inspired by the algorithm described in [89]. The principle of this base algorithm is to combine the data from the left and right accelerometers in order to estimate the angular rotation speeds independently of any translation, and decrease the effect of sensor noise. By integrating these rotation speeds, and calibrating the algorithm applied in the context of tracking head motions, it is possible to infer the yaw of the head wearing the sensors.

6.4.1 Description of the model

Two accelerometers A1 and A2 are mounted on the opposite sides of a head, at the position of the ears. When testing the algorithm described in this paper, they were placed above the ears. The assumption is made that each of those accelerometers is considered to be fixed on the head, and cannot move independently from each other nor from the head. Their coordinate system $(x, y, z, \omega_x, \omega_y, \omega_z)$ is defined as shown in Fig. 6.7.

In this algorithm, it is also assumed that both accelerometers are aligned with the center of the head, on the y-axis of rotation.

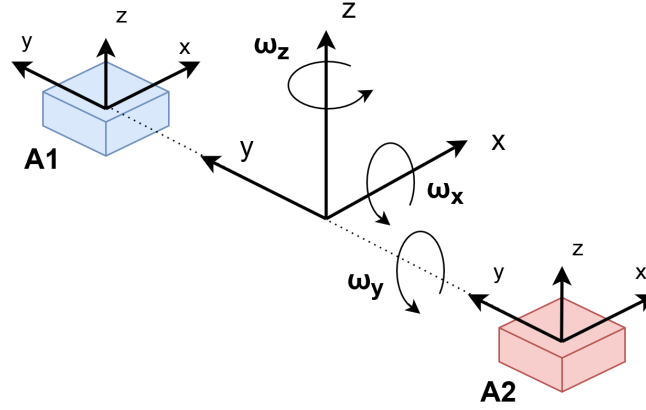


Figure 6.7: Accelerometers setup. The accelerometers A1 and A2 are spaced by a distance D and supposed to be aligned together with the center of the system on the y -axis. .

6.4.2 Base algorithm

For this part of the algorithm, the assumption is made that both x -axes of the accelerometers are parallel to each other and that both z -axes are parallel too. Of course, in practical cases where accelerometers are placed on a human head, this is not the case, but the issue is discussed further in Section 6.4.3.

Computation of the rotational speed magnitude

The angular speed around the x -axis (roll velocity) ω_x , and the angular speed around the z -axis (yaw velocity) ω_z , can be computed by first calculating the radial and tangential accelerations.

The radial acceleration, as measured by the accelerometer A1, can be computed as:

$$a_{y1} = -\Omega_r^2 d_1 \quad (6.1)$$

where d_1 is the distance between the center of the head and the accelerometer A1 and Ω_r is the angular speed. Similarly, the radial acceleration measured by the accelerometer A2 is computed as:

$$a_{y2} = \Omega_r^2 d_2 \quad (6.2)$$

where d_2 is the distance between the center of the head and the accelerometer A2. Then, by taking the difference between the two measurements:

$$a_{y1} - a_{y2} = \Omega_r^2 (d_2 + d_1) = \Omega_r^2 D \quad (6.3)$$

where $D = d_1 + d_2$ is the fixed distance between the two accelerometers. Hence, the angular

speed can be determined as:

$$\Omega_r = \sqrt{(|\frac{a_{y1} - a_{y2}}{D}|)} \quad (6.4)$$

Computation of the angular velocities

To deduce the angular velocities, it is first necessary to compute the angular accelerations. The tangential acceleration measured at the accelerometer A1 corresponds to:

$$a_{x1} = -\alpha_z d_1 \quad (6.5)$$

where α_z is the angular acceleration around the z-axis. The tangential acceleration measured at accelerometer A2 is:

$$a_{x2} = \alpha_z d_2 \quad (6.6)$$

Then, the difference between the two measurements leads to:

$$a_{x2} - a_{x1} = \alpha_z (d_2 + d_1) \quad (6.7)$$

Similarly, it is possible to deduce that:

$$a_{z2} - a_{z1} = \alpha_x (d_2 + d_1) \quad (6.8)$$

Hence, the angular accelerations around the z-axis and x-axis can be calculated from the x-axis and z-axis accelerations respectively:

$$\alpha_z = \frac{a_{x2} - a_{x1}}{D} \quad (6.9)$$

$$\alpha_x = \frac{a_{z2} - a_{z1}}{D} \quad (6.10)$$

The rotation speed magnitude Ω_r is obtained from Eq. 6.4. The accelerometer signals can be summarized as a sum of the ideal accelerometer signals and a noise component. The calculated Ω_r therefore contains noise, meaning that even in the absence of movement, $\Omega_r > 0$. Integrating the noise of Ω_r would lead to errors of rotation estimation, as it would imply to integrate a constant value in the next steps. In order to limit these errors, any value of Ω_r that is below a threshold Ω_T is set to 0.

$$\Omega_r = \begin{cases} \sqrt{(|\frac{a_{y1} - a_{y2}}{D}|)} & \text{if } \Omega_r \geq \Omega_T \\ 0 & \text{otherwise} \end{cases} \quad (6.11)$$

The integration of α_z and α_x enables determining ω_z and ω_x . Thus, the relative magnitude and

direction of the rotation speeds are calculated from the integrals of the tangential accelerations, and can therefore be calculated from an initial speed constant:

$$\omega_z = \int_{t_{init}}^t \alpha_z dt \quad (6.12)$$

$$\omega_x = \int_{t_{init}}^t \alpha_x dt \quad (6.13)$$

where t_{init} is an initial time instant at which the speed is known, e.g. $\omega_x(t_{init}) = \omega_z(t_{init}) = 0$.

Computation of the yaw estimation

In the setup described in Fig. 6.7, the accelerometers A1 and A2 are located on the rotation y-axis. Consequently, the total angular rotation rate ω_{total} that can be measured by this setup is invariant under a rotation around the y-axis, and given by:

$$\omega_{total} = \sqrt{\omega_z^2 + \omega_x^2} \quad (6.14)$$

Using the total rotation rate magnitude calculated in Eq. 6.14 and the relative magnitude and direction of rotation obtained from Eq. 6.11, the yaw velocity yaw can be determined as:

$$yaw = \Omega_r \frac{\omega_z}{\omega_{total}} \quad (6.15)$$

which is integrated in order to obtain the yaw :

$$yaw = \int_{t_{init}}^t \Omega_r \frac{\omega_z}{\omega_{total}} dt \quad (6.16)$$

where t_{init} is a time instant where the orientation of the user's head relative to a fixed point in the environment is known, e.g. the user is facing forward at 0° at the start.

6.4.3 Upgrades of the algorithm for the head tracking application

Estimation of the threshold Ω_T

Ω_r is the square root of the absolute value of a difference between two variables, thus it is always positive, even when the system or the head of the user does not move. The mean value of Ω_r is then defined by the mean and standard deviation of the noise of the sensors a_{y1} and a_{y2} . A non-null Ω_r implies a non-null rotation speed. In order to ensure that Ω_r is null when there is no movement, a threshold Ω_T is used to force any smaller value of Ω_r to be set to 0. i.e. when a steady-state is detected. This value should depend on $\overline{\Omega_r}$, the mean of Ω_r and its standard deviation $\sigma(\Omega_r)$. As the noise is likely to evolve between measures, and during the estimation, it is necessary to define Ω_T dynamically so that the value can be optimally set in every situation. The priority is to have a robust computation, as having an underestimation of

Ω_T would lead to integrate noise in the next step when $\Omega_r > \Omega_T$. It is also important to have a reactive computation of Ω_T that adapts to the evolution of Ω_r with a small delay, in order to avoid integrating noise for too many samples nor missing too much of the motion if Ω_T is temporarily overestimated. A satisfying method for defining and updating Ω_T was defined using the following computation. First let K_{Ω_T} be defined as the following function:

$$K_{\Omega_T}(n, Nb, Nc) = \overline{\Omega_r}(n - Nb : n - Nc) + \sigma(\Omega_r(n - Nb : n - Nc))\gamma \quad (6.17)$$

where γ is a parameter of the algorithm which value can be set empirically and Nb is the fixed size of the sliding buffer used for this computation. Nc is used to select the first samples of the buffer and can be set in relation to Nb , e.g. $Nc = \frac{3}{4}Nb$. The following conditions are verified at every sample n :

$$\begin{cases} \sigma(\Omega_r(n - Nb : n)) < S_{\Omega_r} \\ \Omega_r(n) > K_{\Omega_T}(n, Nb, Nc) \\ clock \bmod(r_{\Omega_T}) = 0 \end{cases} \quad (6.18)$$

where S_{Ω_r} is the tolerance on the standard deviation that can be set empirically as an absolute value. Finally $clock$ is the running sample counter and r_{Ω_T} is the refresh rate in samples used to update $\Omega_T(n)$, defined such as:

$$\begin{cases} \Omega_T(n) = K_{\Omega_T}(n, Nb, 0) & \text{if all conditions (6.18) true} \\ \Omega_T(n) = \Omega_T(n - 1) & \text{otherwise} \end{cases} \quad (6.19)$$

An example of the results obtained for Ω_T with this method is displayed in Fig. 6.8

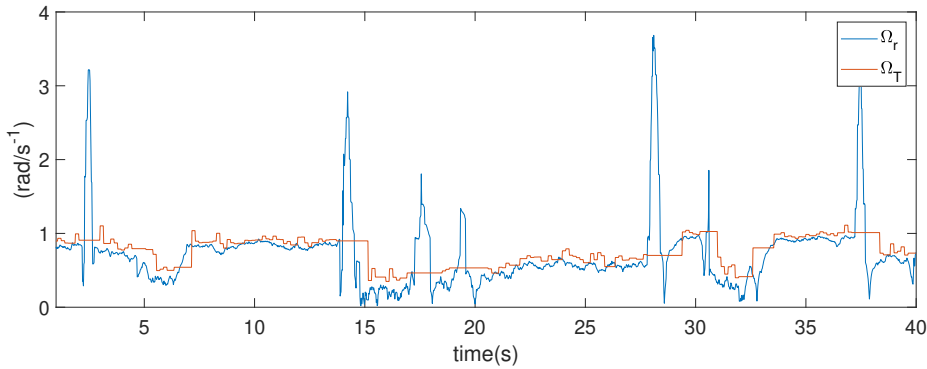


Figure 6.8: Example of the computation of Ω_T in relation to the evolution of Ω_r .

Integration

In practice, the sensor data acquisition is discrete. The various stages of integration could therefore be programmed with either a cumulative sum or with the trapezoidal rule. Both approaches were compared and differences were not significative. The simpler cumulative sum was therefore chosen for the integration, and Eq. 6.12 and Eq. 6.13 therefore become:

$$\omega_z(n) = \frac{1}{f_s} \sum_{n_0}^n \alpha_z(k) + \omega_z(n_0 - 1) \quad (6.20)$$

$$\omega_x(n) = \frac{1}{f_s} \sum_{n_0}^n \alpha_x(k) + \omega_x(n_0 - 1) \quad (6.21)$$

where k varies between the first sample of the buffer n_0 and the last sample n , and f_s is the sampling frequency (50 Hz). Similarly, Eq. 6.16 in discrete time becomes:

$$yaw(n) = \frac{1}{f_s} \sum_{n_0}^n yaw(k) + yaw(n_0 - 1) \quad (6.22)$$

These integration steps introduce some errors, because of the noise of the sensors and the discrete approximations. Consequently, the two stages of integrals should be reset whenever possible. For Eq. 6.20 and Eq. 6.21, this means every time the rotation speed of the system Ω_r is null. For Eq. 6.22 this means every time the rotation speed Ω_r is null and the system receives information about the position compared to a fixed point in the room, otherwise the last position before the steady-state is always kept in a buffer.

In the current implementation of the system, the reset of the integrals in Eq. 6.20 and Eq. 6.21 is performed with the following method:

$$\begin{cases} \omega_z(n_0 - 1) \text{ is set to } 0 & \text{if } \Omega_r(n - N_{Init}\Omega_T + 1 : n) = 0 \\ \omega_z(n_0 - 1) & \text{unchanged otherwise} \end{cases} \quad (6.23)$$

$$\begin{cases} \omega_x(n_0 - 1) \text{ is set to } 0 & \text{if } \Omega_r(n - N_{Init}\Omega_T + 1 : n) = 0 \\ \omega_x(n_0 - 1) & \text{unchanged otherwise} \end{cases} \quad (6.24)$$

This means that the integral is reset when no movement is detected during $N_{Init}\Omega_T / f_s$ seconds. The integral of Eq. 6.22 is only reset at the beginning of the yaw estimation.

Sensor denoising using an adaptive smoother

Eq. (6.11), Eq. (6.9) and Eq. (6.10) include the difference of the accelerations measured by each sensor on the x-, y- and z-axis respectively. Let A_Y , A_X and A_Z be defined as:

$$A_Y = a_{y1} - a_{y2} \quad (6.25)$$

$$A_X = a_{x1} - a_{x2} \quad (6.26)$$

$$A_Z = a_{z1} - a_{z2} \quad (6.27)$$

The axis label will be discarded in the following as the concepts developed in this section are independent of the axis.

As mentioned in section 6.4.2, the sensor signals are noisy. Thus the available signals are the noisy counterparts \tilde{a}_1 and \tilde{a}_2 of the true accelerations a_1 and a_2 . Assuming an additive noise model, the noise components v_1 and v_2 are superposed to a_1 and a_2 respectively such as:

$$\tilde{A} = \tilde{a}_1 - \tilde{a}_2 = (a_1 + v_1) - (a_2 + v_2) = A + v \quad (6.28)$$

Because the sensor noises v_1 and v_2 are uncorrelated, $v = v_1 - v_2$ can be viewed as an average noise signal. Using \tilde{A} leads to a 3dB SNR improvement compared to a method with a single sensor that would rely on a unique signal, e.g. \tilde{a}_1 . Nevertheless, the signal \tilde{A} has to be further denoised before being processed in the integration stages and to compute Ω_T .

Two other reasonable assumptions about the acceleration signal A and the noise signal v can be made. First A and v are orthogonal since they are uncorrelated, i.e. $\mathbb{E}[Av] = \mathbb{E}[A]\mathbb{E}[v]$ with v zero mean, i.e. $\mathbb{E}[v] = 0$. The variance of \tilde{A} is:

$$\sigma_{\tilde{A}}^2 = \mathbb{E}[(\tilde{A} - \mathbb{E}[\tilde{A}])^2] \quad (6.29)$$

$$= \mathbb{E}[(A + v - \mathbb{E}[A])^2] \quad (6.30)$$

$$= \mathbb{E}[(A - \mathbb{E}[A])^2] + 2\mathbb{E}[A]\mathbb{E}[v] + \mathbb{E}[v^2] \quad (6.31)$$

$$= \sigma_A^2 + \sigma_v^2 \quad (6.32)$$

Second, v is stationary and its power spectral density is relatively uniformly distributed over frequencies. On the other hand, A is a non-stationary band-limited signal with a short-term power spectral density $\Phi_A(f, n)$ upper bounded by a maximal frequency $f_B(n)$ that can vary over time, depending on the head motion. The motion can be qualitatively characterized by two different motion modes, "slow" and "fast". In the slowest mode, the head is either moving

and/or accelerating very slowly or not moving at all. In this case, $\Phi_A(f, n)$ is band-limited at $f_B(n) = f_0$, which means that ν is the dominant signal beyond f_0 as qualitatively pictured in Fig. 6.9(a). In the fastest mode, the head is moving quickly and/or accelerating strongly. In this case, $\Phi_A(f, n)$ is smeared over a broader frequency range, delimited by a frequency $f_B(n) = f_1$ with $f_1 \gg f_0$ as qualitatively depicted in Fig. 6.9(b). A is the dominant signal up to f_1 . In particular, compared to slow mode, $\Phi_\nu(f)$ is negligible compared to $\Phi_A(f, n)$ between f_0 and f_1 .

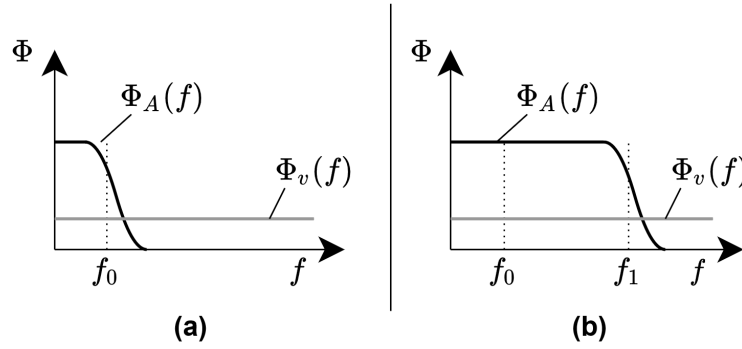


Figure 6.9: Power spectral density as a function of frequency in slow (a) and fast (b) mode.

Based on these assumptions, one can expect a good estimation \hat{A} of A using a low-cost smoothing of \tilde{A} made of a first order recursive filter with cut-off frequency f_0 in slow mode and cut-off frequency f_1 in fast mode. The effective cut-off frequency f_c can be shifted between f_0 and f_1 as a function of the observed mode, i.e. continuously between the "slow" and "fast" modes. This can be achieved with the following non-linear filtering operations. The signal is left untouched below f_0 . A thresholding mechanism is applied between f_0 and f_1 . If sufficient power is detected up to f_c , the band from f_0 to f_c is passed through, while the spectral band beyond f_c is filtered out. The heuristic behind this approach is the following. On one hand, if the detected power between f_0 and f_c cannot be produced by the high frequency components of ν alone, some participation of A is necessary. On the other hand, beyond f_c , the high frequency components of ν are sufficient to explain the amount of power detected.

Such a smoothing has to be transparent while the user is moving (fast mode), in order to accurately track the variations during those segments (f_c close to f_1). Nevertheless, when the user is nearly static (slow mode), it is desirable to have a more radical smoothing (f_c close to f_0), in particular because one of the goals is to reduce the standard deviation of Ω_r as much as possible so that the computation of Ω_T can be optimized to the lowest possible value. Indeed, overestimating Ω_T means that slow motions will not be detected, whereas underestimating Ω_T means that some noise will be integrated, and thus the final estimation of the yaw will drift.

The estimate $\hat{A}(n)$ of A is obtained via recursive averaging:

$$\hat{A}(n) = \lambda(n)\tilde{A}(n) + (1 - \lambda)\hat{A}(n - 1) \quad (6.33)$$

where $\lambda(n)$ is the time variable smoothing coefficient that must be adaptive controlled to guarantee a good approximation of $A(n)$ independently of the context (slow or fast mode). $\lambda(n)$ can be derived from $f_c(n)$ [113]:

$$\lambda(n) = \lambda(f_c(n)) = c(n) + \sqrt{c^2(n) - 2c(n)} \quad (6.34)$$

$$c(n) = \cos(2\pi f_c(n)/f_s) - 1 \quad (6.35)$$

with f_s the sampling frequency. $\lambda(n)$ is bounded by $\lambda_0 = \lambda(f_0)$ and $\lambda_1 = \lambda(f_1)$. To use the above equation, the adaptive control of $\lambda(n)$ requires the signal $\tilde{A}(n)$ to be transformed into the frequency domain to compute an estimate $\hat{f}_B(n)$ from the evaluation of $\Phi_{\tilde{A}}(f, n)$. However, to keep the complexity low, it is desirable to realize the adaption control in the time domain. $\lambda(n)$ can be adapted as a function of the time constant $\tau_c(n)$:

$$\lambda(n) = 1 - e^{\frac{-1}{\tau_c(n)f_s}} \quad (6.36)$$

In slow mode, there is only little power beyond f_0 . Consequently, the change between the current and the previous observed samples:

$$\Delta\tilde{A}(n) := \tilde{A}(n) - \tilde{A}(n - 1) \quad (6.37)$$

cannot be large. Indeed, in slow mode, the acceleration can be expected to be approximately constant, i.e.:

$$\Delta A(n) := A(n) - A(n - 1) \approx 0 \quad (6.38)$$

This means that the observed change is approximately the noise variation between two successive samples:

$$\Delta\tilde{A}(n) \approx \Delta v(n) := v(n) - v(n - 1). \quad (6.39)$$

Replacing $A(n - 1)$ by its estimated $\hat{A}(n - 1)$, a distance function $d(n)$ can be defined as:

$$d(n) := |\tilde{A}(n) - \hat{A}(n - 1)| \quad (6.40)$$

such that in slow mode, the magnitude of $\Delta v(n)$ can be approximated using this distance:

$$|\Delta v(n)| \approx |d(n)| \quad (6.41)$$

In fast mode, the acceleration signal contains power in the high frequencies, allowing $A(n)$ to be very different from $A(n - 1)$. Hence, the approximation of Eq. 6.41 is not valid anymore.

Slow and fast modes can thus be distinguished from each other by comparing the distance function $d(n)$ to a threshold $l_0 \sim \sigma_v$. Practically, $\lambda(n)$ is mapped from the distance $d(n)$ using a piece-wise linear curve as pictured in Fig. 6.10 and implemented using the following equation:

$$\lambda(n) = \min(\lambda_1, \max(\lambda_0, (d(n) - l_0)\eta + \lambda_0)) \quad (6.42)$$

or

$$\lambda(n) = \min(\lambda_1, \max(\lambda_0, (d(n))\eta + \lambda_h)) \quad (6.43)$$

with

$$\eta = \frac{\lambda_1 - \lambda_0}{l_1 - l_0} \quad (6.44)$$

$$\lambda_h = \frac{\lambda_0 l_1 - \lambda_1 l_0}{l_1 - l_0} \quad (6.45)$$

Other mapping functions (e.g. arctangent, hyperbolic tangent, sigmoid function, etc), could also be used to provide a smoother control of lambda, eventually with larger computational costs.

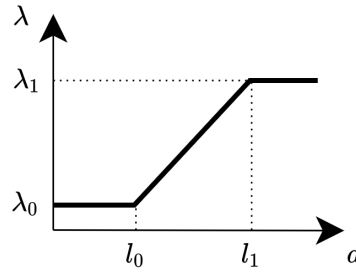


Figure 6.10: Mapping function used to define $\lambda(n)$.

Finally, Ω_r , α_z and α_x are obtained with the following equations:

$$\Omega_r = \begin{cases} \sqrt{|\frac{\hat{A}_y}{D}|} & \text{if } \Omega_r \geq \Omega_T \\ 0 & \text{otherwise} \end{cases} \quad (6.46)$$

$$\alpha_z = \frac{\hat{A}_X}{D} \quad (6.47)$$

$$\alpha_x = \frac{\hat{A}_Z}{D} \quad (6.48)$$

Initialization of the headtracking algorithm

Because of the human head morphology, when using two sensors placed at the surface of a human head, the axes of the accelerometers may not be aligned with the axis described in the algorithm. The y-axis should be aligned with the interaural axis, the z-axis should be vertical, and the x-axis should be in the forward direction. In order to correct for this, a multi-stage rotation has been implemented. A 12-stages calibration of multiple sensors was proposed in [133]. The method proposed here requires only two stages of calibration.

This requires a calibration step for all users. First, once the algorithm is running, they need to remain upright, looking in front of them, and steady. This is the first steady-state, and helps doing an initial alignment of the z-axis by applying a roll and a pitch correction. The users then need to lower their head, i.e. doing a pitch movement without any roll nor yaw movement, and keep their head down steadily for several seconds. This step helps calculate the individual yaw of each accelerometer.

• *Steady states detection for the calibration*

An automatic detection of the calibration states has been developed. It detects, in the first few seconds of the tracking algorithm, the time interval when the head of the user is steady. The goal is to trigger the correct computations for each of the two steady states.

For this purpose, a sliding buffer of $N_{SteadySTD}$ samples ($N_{SteadySTD}/f_s$ seconds) for each acceleration data is used. The acceleration element exhibiting the highest standard deviation is selected during the first $N_{SteadySTD}$ samples to define a threshold S_{cal} based on this standard deviation and a factor δ determined empirically.

$$S_{cal} = \max(\sigma(a_{x1}), \sigma(a_{x2}), \sigma(a_{y1}), \sigma(a_{y2}), \sigma(a_{z1}), \sigma(a_{z2}))\delta \quad (6.49)$$

After S_{cal} is determined, if the standard deviation of the last $N_{SteadySTD}$ samples remains below this threshold for every acceleration, the head is considered in steady state:

$$\begin{cases} \sigma(a_{x1}) \\ \sigma(a_{x2}) \\ \sigma(a_{y1}) \\ \sigma(a_{y2}) \\ \sigma(a_{z1}) \\ \sigma(a_{z2}) \end{cases} < S_{cal} \text{ during } N_{SteadySTD} \text{ samples} \quad (6.50)$$

A counter is incremented for each new steady state detected, enabling to trigger the appropriate computation corresponding to each of the two first steady states.

• **Compensation of the initial pitch, roll and yaw of each sensor**

The first step consists in computing the initial pitch and roll angle for each accelerometer individually. This can be done at the start, as soon as the first steady state is detected. For each sensor ($j = 1$ or $j = 2$ for the accelerometers A1 or A2), the following equations, for which the demonstration can be found in [134], are used for the pitch and the roll, respectively:

$$\theta_j = \frac{180}{\pi} \arctan \left(\frac{a_{xj}}{\sqrt{a_{y,j}^2 + a_{zj}^2}} \right) \quad (6.51)$$

$$\phi_j = \frac{180}{\pi} \arctan \left(\frac{a_{yj}}{\sqrt{a_{x,j}^2 + a_{zj}^2}} \right) \quad (6.52)$$

For each sensor, the following rotation matrices are then used to compensate for the initial pitch and roll respectively:

$$R_{y,j}(\theta_j) = \begin{pmatrix} \cos(\theta_j) & 0 & \sin(\theta_j) \\ 0 & 1 & 0 \\ -\sin(\theta_j) & 0 & \cos(\theta_j) \end{pmatrix} \quad (6.53)$$

$$R_{x,j}(\phi_j) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\phi_j) & -\sin(\phi_j) \\ 0 & \sin(\phi_j) & \cos(\phi_j) \end{pmatrix} \quad (6.54)$$

Once the first steady-state is detected, the means $\bar{a}_{x1,steady1}$, $\bar{a}_{y1,steady1}$, $\bar{a}_{z1,steady1}$ and $\bar{a}_{x2,steady1}$, $\bar{a}_{y2,steady1}$, $\bar{a}_{z2,steady1}$ for each of the sensors are calculated. The means are computed over $N_{SteadySTD}$ samples of smoothed sensor data. For this specific part of the algorithm, the sensor data is smoothed using a basic exponential smoothing, independent from the adaptive smoother described in Section 6.4.3. A fixed smoothing factor is used, with a value $\lambda_R = 0.05$ ($\tau = 0.39s$ with $f_s = 50Hz$), and the smoothed accelerations for each axis i ($i = x, y$ or z) and each sensor j are computed as:

$$a_{ij,lp}(n) = \lambda_R a_i(n) + (1 - \lambda_R) a_{ij,lp}(n - 1) \quad (6.55)$$

The roll and pitch value of both sensors are then computed using the means for each sensor.

However, this compensates only the roll and the pitch of each sensor. It does not correct their potential initial azimuth. If the azimuth is not corrected, when the head of the user remains

in the horizontal plane during a movement, the accelerations on the y-axis and on the x-axis will not be correct. When the head of the user is not in the horizontal plane, a difference of acceleration between the respective y-axis of the two sensors is generated, even if the user is not moving.

In order to correct the yaw ψ of each sensor, a second steady-state is required. When users rotate their head so as to have a non-null new pitch, but no new roll nor yaw, if the yaw of one of the sensor is non-null, the y acceleration of this sensor will also be non-null. Therefore when a second steady-state is detected, the algorithm assumes the user moved only in pitch (around the y-axis). It computes the means $\bar{a}_{x1,steady2}$, $\bar{a}_{y1,steady2}$, $\bar{a}_{z1,steady2}$ and $\bar{a}_{x2,steady2}$, $\bar{a}_{y2,steady2}$, $\bar{a}_{z2,steady2}$ of each of the sensors on the last $N_{SteadySTD}$ samples, and tests all possible yaw values for each sensor between -30° and $+30^\circ$ by steps of 0.1° . Potential adjusted values of the $\bar{a}_{y1,steady2}$ and $\bar{a}_{y2,steady2}$ are tested for each of the possible yaw values, such as:

$$\psi_1 = \operatorname{argmin}_{\psi} \bar{a}_{x1,steady2} \sin(\psi) + \bar{a}_{y1,steady2} \cos(\psi) \quad (6.56)$$

and

$$\psi_2 = \operatorname{argmin}_{\psi} \bar{a}_{x2,steady2} \sin(\psi) + \bar{a}_{y2,steady2} \cos(\psi) \quad (6.57)$$

From these values, the matrix associated with the initial yaw of each sensor is computed as:

$$R_{z,j}(\psi_j) = \begin{pmatrix} \cos(\psi_j) & -\sin(\psi_j) & 0 \\ \sin(\psi_j) & \cos(\psi_j) & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (6.58)$$

In order to increase the precision of the estimations, it is necessary to run this estimation in a loop: the roll, pitch, and yaw angles are estimated, then the rotation matrix $R_{z,j}(\psi_j)R_{y,j}(\theta_j)R_{x,j}(\phi_j)$ is applied to the $\bar{a}_{steady1}$ and $\bar{a}_{steady2}$ values, and then roll, pitch, and yaw correction are re-estimated and added to the initial roll, pitch, and yaw estimations in a loop until:

$$\begin{cases} \bar{a}_{x1,steady1} \\ \bar{a}_{y1,steady1} \\ \bar{a}_{x2,steady1} \\ \bar{a}_{y2,steady1} \\ \bar{a}_{x1,steady2} \\ \bar{a}_{y1,steady2} \\ \bar{a}_{x2,steady2} \\ \bar{a}_{y2,steady2} \end{cases} < tol \quad (6.59)$$

where tol is fixed empirically based on the noise of the sensors. Alternatively, the computation of the means can be achieved until more than N_{loop} loops took place.

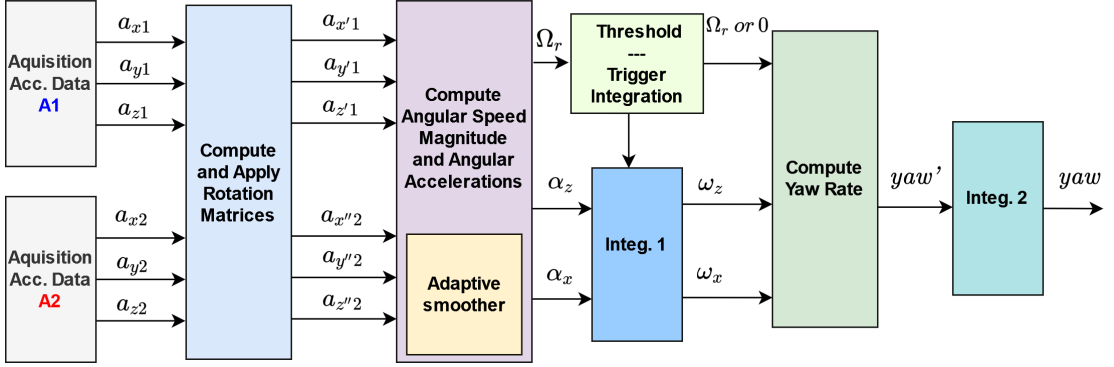




Figure 6.11: Simplified block diagram of the real-time implementation, showing the main steps and data. $(a_{x'1}, a_{y'1}, a_{z'1})$ and $(a_{x''2}, a_{y''2}, a_{z''2})$ denote the accelerations of the sensors 1 and 2 after applying their respective rotation matrices.

6.4.4 Real-time implementation

The real-time version of the algorithm was implemented on Simulink. A simplified descriptive block diagram is depicted in Fig. 6.11. The sampling frequency was fixed to 50 Hz. The algorithm uses sliding block processing. Most of the processing is performed sample by sample, inside each block. The size of the block was therefore reduced to 1, keeping latency low. No overlapping blocks are necessary in the implemented algorithm. The steps of the real-time implementation are described below and serves as a simplified summary of the principle of the algorithm.

- First, the data from the accelerometers is retrieved by the **Acquisition Acc. Data** blocks. The acquisition and processing rate is fixed by this step, at 50 Hz.
- Rotation matrices for each of the sensor are computed in the block **Compute and Apply Rotation Matrices**. The smoothing factor used for this computation is internal to this block and is set to 0.05 ($\tau = 0.39s$). The accelerations transformed with the rotation matrices are provided to the next block, they are called $a_{x'1}$, $a_{y'1}$, $a_{z'1}$ and $a_{x''2}$, $a_{y''2}$, $a_{z''2}$ for each sensor respectively.
- In the next block **Compute Angular Speed Magnitude and Angular Accelerations**, the angular accelerations α_z and α_x as well as Ω_r are computed. The block also includes the **Adaptive Smoother**, which behaviour is described in Section 6.4.3, and outputs a smoothed Ω_r .
- Ω_T is computed in the block **Thresholding - Trigger Integration** based on the computation of Ω_r . Then the output Ω_r is set to 0 if $\Omega_r < \Omega_T$. The same block triggers the first stage of integration.
- The first integration stage **Integ. 1** takes the angular accelerations α_z and α_x as input and outputs the angular velocities ω_z and ω_x . A 1-sample buffer is used in those

integrations for the initial condition of each new sample. If a steady state is detected, the integration is frozen, and initial conditions of the integration are reset to 0 (no motion, the velocity is null).

- The yaw rate is computed in the next block  **Compute Yaw Rate** based on Ω_R , ω_Z and ω_X .
- Finally, the yaw position is estimated using a second integration stage  **Integ. 2**. Similarly as for the previous integration, a 1-sample buffer is used for the initial conditions. No computation is achieved during steady states, and the last position is kept in the buffer as initial condition for when the steady state is over and the next motion has to be evaluated.

To obtain a reference yaw as a target performance, an additional IMU/AHRS device is used^{III}. A scope is used at the end of the chain to display results in real-time, as well to assess the performance in comparison to the reference tracking obtained with the IMU/AHRS device.

The yaw is not computed until the two first steady-states are detected and the rotation matrix is completely determined.

6.5 Evaluation of the algorithm

In order to evaluate the reliability of the algorithm, a database of head movements was first recorded, on which the algorithm was then tested.

6.5.1 Database

The aim of this database is to display a large and realistic range of head movements that can occur in conditions where the algorithm could be used. The database aims at acquiring realistic data, it was therefore chosen to measure data from subjects, doing a task where they would not limit their head movements to the horizontal plane.

Setup and task

The setup consisted of five loudspeakers located at -90° , -45° , 0° , 45° and 90° around the participant, at a distance of 2 m, as shown in Fig. 6.12.

For the data acquisition, HIs with embedded accelerometers were placed over the pinna of the participants. A headband is used to carry the small circuit boards used to process the sensors' data. The cable linking the HIs to the circuit board and the circuit board to the computer are chosen to be thin and flexible to avoid unwanted tension forces to be applied on the HIs. The

^{III}NGIMU, x-io Technologies Limited

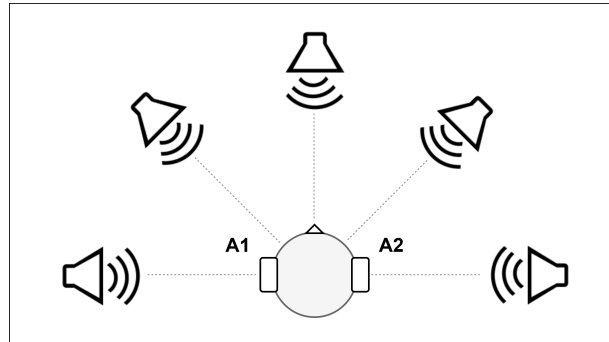


Figure 6.12: Schematic representation of the setup used for recording the database, the loudspeakers were placed respectively at -90° , -45° , -0° , $+45^\circ$ and $+90^\circ$.

device fitting is shown with a close-up shot in Fig. 6.14. The subject was also equipped with the IMU/AHRS device.

The acceleration data from both sensors was retrieved via USB. Euler angles measured with the IMU/AHRS device were recorded as well, and transmitted through a wireless WiFi connection and served as a reference measurement. All the data was acquired and recorded on a dedicated Simulink model.

The listeners were asked to point their head toward the loudspeaker which emitted a sound (a speech sample of the number visible on the loudspeaker). Then they had to wait in this position until a word was pronounced by the same loudspeaker. Subsequently they had to write down the word on a piece of paper on the table. Then they could move their head back up to the last loudspeaker emitting the sound, and wait for the next number and word. The total duration of the measure was about 2 min and included 12 target positions.

For this database, 11 participants were fitted with the prototype and the reference IMU/AHRS sensor, and followed the protocol mentioned above. The setup is pictured in Fig. 6.13. The instruction was given to the participants to turn their head as naturally as possible, as they would be attending a meeting with several speakers around them.

For the calibration, visual marks were placed in front of the subject and on the table for the first and second steady-states respectively. To ensure a precise movement from the subject, a visual feedback of the roll angle, as measured by the IMU/AHRS device, was provided to them on a screen. When the subject performed the second steady state, with a targeted pure pitch motion, they were instructed to keep this roll value between $\pm 2^\circ$, which was empirically found to be sufficient precision. The participants managed to perform the calibration with those constraints after a maximum of 3 trials.

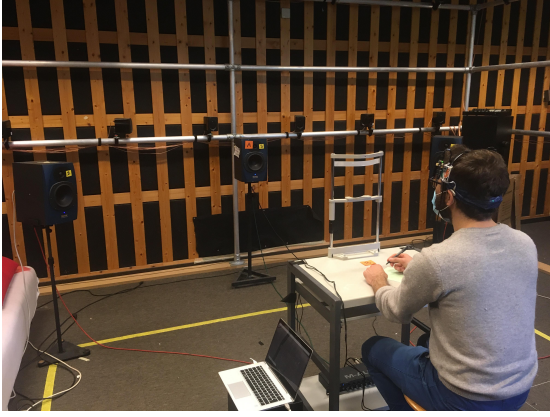


Figure 6.13: Setup for the acquisition of the database.

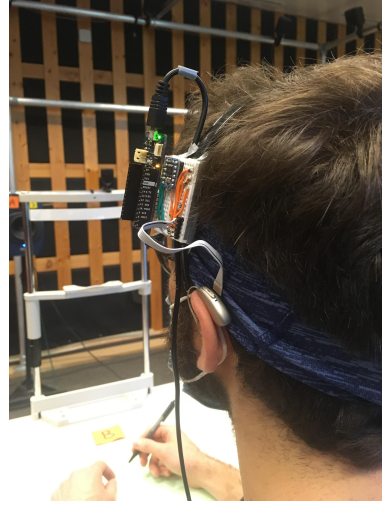


Figure 6.14: Fitting of the prototype with embedded accelerometers.

6.5.2 Evaluation Method

A batch evaluation was conducted using the measures and various settings of the algorithms. The goal was first to assess the effect of various parameters used in the algorithm and look at the best performance achievable for each measurement.

An error score was computed using the RMS error of the difference between the head-tracking reference yaw as obtained from the IMU/AHRS device and the estimation obtained from the above described the algorithm using two hearing devices. This corresponds to:

$$RMSE = \sqrt{\frac{1}{N_m} \sum_{k=1}^{N_m} (yaw_{est}(k) - yaw_{ref}(k))^2} \quad (6.60)$$

Where N_m is a length of the measurement in samples, yaw_{est} is the estimated yaw, and yaw_{ref} is the reference one.

The choice of the tested parameters and their range of interest were defined empirically and are shown in Tab. 6.1. The factorial design with those factors and levels leads to total of $n = 576$ combinations.

6.5.3 Results

During the measure, the USB cable was attached to the headband and could potentially apply some tension on the sensors if the participant involuntarily pulled the cables. Participants were

Parameter	Tested values
D	0.15 and 0.17 (m)
δ	2 and 4
γ	4, 5.5, and 7
Nb	16, 24, and 32 (samples)
Nc	$3/4Nb$ and $1/2Nb$ (samples)
S_{Ω_r}	0.07, 0.08, 0.09, and 0.10
r_{Ω_T}	$1/4Nb$ and $1/2Nb$ (samples)

Table 6.1: List of the tested parameters and their values..

wearing face masks, which made it complicated to have the hearing aid fixed behind the ear firmly enough. Consequently, HAs could sometimes move, in which case the calibration would not be adequate anymore, leading to a drift in the yaw estimation. Hence the occurrences of displacement had to be monitored both with feedback from the participants who felt that the device moved, and by scanning the acceleration data and detect unexpected behaviors, such as a sudden jump in acceleration while the participant is supposed to stand still (as can be checked with the reference Euler angles from the IMU/AHRS device). For that reason, in order to maximize the chances of obtaining at least one measure per participant with no hardware issues, two to three measures were recorded per participant. Out of the total of 30 measures, 19 were kept, for which no obvious hardware issue causing the sensors to move was detected.

The results of the RMSE obtained for those 19 measurements for all tested combinations of parameters are displayed in Fig.6.15.

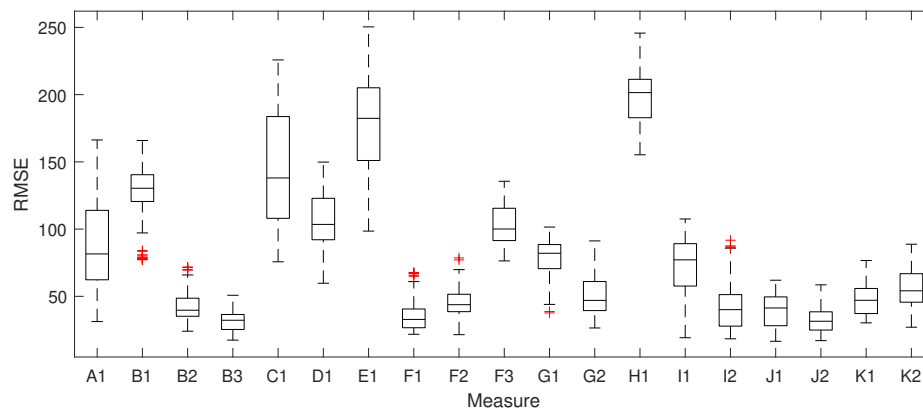


Figure 6.15: RMS Error as a function of the measurement. The letter corresponds to the anonymized participant, and the number is the repetition when the subject's recordings led to several usable measures. $n = 576$ combinations of the parameters were performed for each measure.

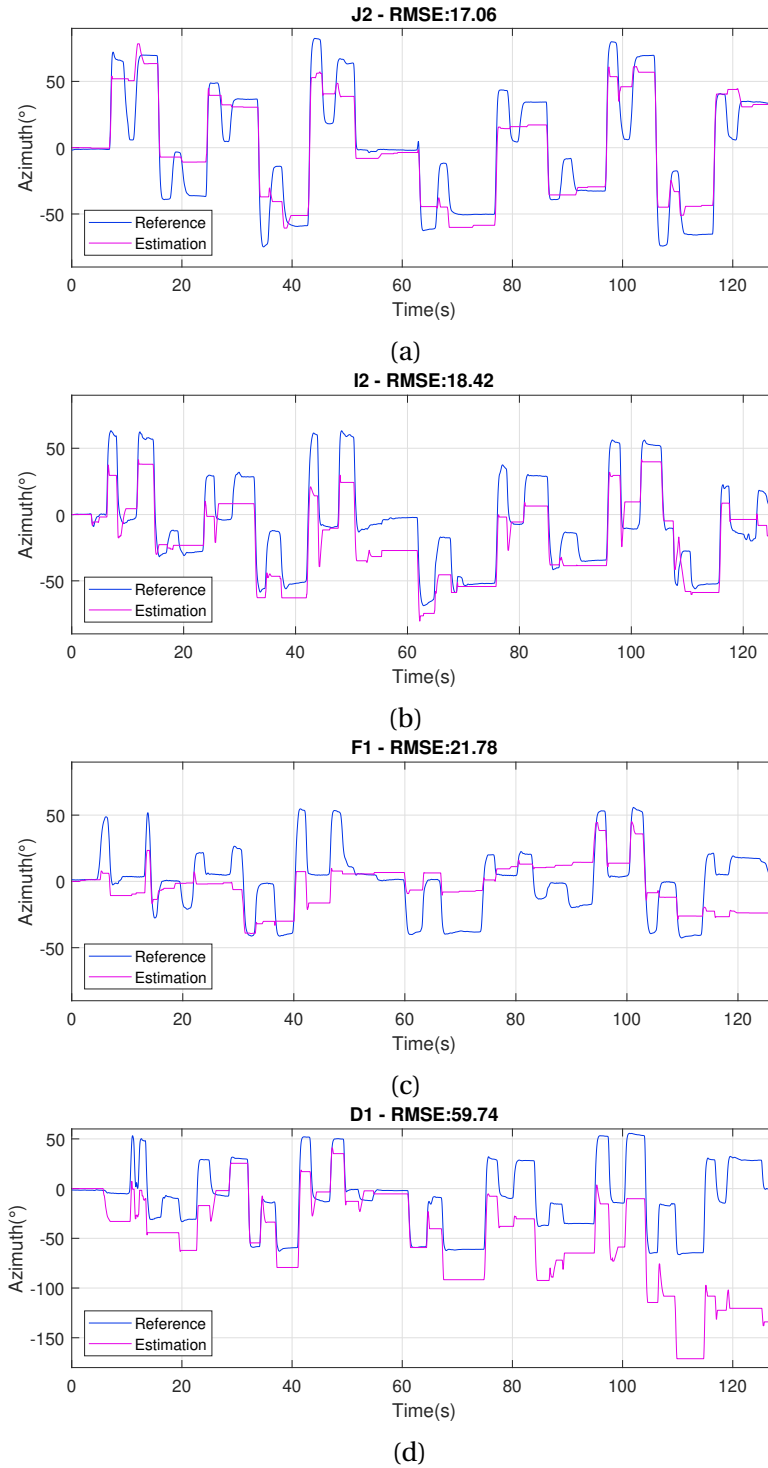


Figure 6.16: Examples of various degrees of performance of the algorithm. The plots show the estimations and the related references for several measures with different participants and their best setting. (a) Illustrates an example of a good performance of the algorithm. (b) Shows an example of an overall acceptable estimation with small errors yielding local mismatches. (c) Illustrates a scenario where most of the movements are not detected. (d) Shows the worst type of performance, where large errors yield a general drift of the estimation.

Examples of various yaw estimations on different measures are depicted in Fig. 6.16. First two measures, associated with lower RMSE values, are shown in Fig. 6.16(a) and 6.16(b). In Fig. 6.16(a), it can be seen that the main movements are detected in the correct direction and are well tracked while the subject is moving in azimuth. While the subject performs the writing task, the azimuth might not always be tracked. The same observations can be made in Fig. 6.16(b), while some local errors decrease the performance.

Fig. 6.16(c) depicts a case for which most of the movements are underestimated, and a part are not detected at all. This could be explained by head movements which are too slow. In this case, the acceleration values are too small and thus, a part or the totality of the movement is not properly detected and set to 0 (see Eq. 6.11) as the magnitude of the acceleration is too close to the magnitude of noise. Finally an example of a low performance of the algorithm is depicted in Fig. 6.16(d). In this case a large drift can be observed at the end of the measure. Some sign inversions sometimes occur at the end of the movement, which yields large errors. An hypothesis can be made in this case that the initial yaw was poorly estimated during the calibration, which results in an overestimation for a part of the movements in one direction, and underestimation for the other direction, resulting in a drift over time.

6.6 Limitations of the algorithm

6.6.1 Measure dependent performance

With a goal to assess the potential for an optimal setting which might yield a good performance on most measurements, a factorial analysis was conducted, as described in [111]. For seven of the measures (A1, B1, C1, D1, E1, F3, H1), it is considered that the errors are too large to be included in the data used to find a common set of parameters. It is hypothesized that sensors displacement might have occurred during these measurements, and that these displacements might not have been visible when checking the acceleration data. This could happen if a displacement occurred during a movement. In this case, the displacement-induced shift in acceleration would not be visible while the acceleration is changing because of the head movement. Another hypothesis is that the calibration phase was not achieved precisely enough, notably to compute the initial yaw of the hearing-aid behind the ear. Hence, it was decided to conduct the investigation for the possibility of an optimal parametrization with the twelve other measures.

The factorial analysis method described in [111] assumes that the dependent variable Y of the experiment is determined by the experimental conditions, and can be approximated by a polynomial function with second order interaction terms. An example is given below for three hypothetical parameters x_1 , x_2 and x_3 :

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + residual \quad (6.61)$$

b_0 corresponds to the mean value of all possible RMSE over all parameter combinations for a single measurement. The parameters and their two levels can be summarized in a factorial table, where each line is a unique combination of all parameters. Then a sign table is built from this table, for each column (associated with one parameter), a "-" is associated with the minimum value and a "+" is associated with the maximum value. The coefficients b_1, b_2, \dots are evaluated by either adding or subtracting the value of the response (in this case the RMSE value) for each line based on the signs in their corresponding columns from the factorial sign table. The result is divided by the total number of observations. Once the coefficients b_0, b_1, b_2, \dots are evaluated, the function describes qualitatively how the experimental variables and their interactions influence the response. This analysis was conducted for the twelve selected measurements and the seven parameters of Tab. 6.1 were used for the analysis. The minimum and maximum levels of the table are used for the two levels in the analysis. These parameters were combined in a full factorial design, leading to a total of $2^7 = 128$ combinations. A batch evaluation was run for all the measurements with all the 128 parameters combinations. The mean b_0 and the coefficients for the main effects of the seven parameters for each of the twelve selected measures are displayed in Fig. 6.17. The means and standard deviations across measures for the main effect of the seven parameters are reported in Tab. 6.2. The interactions are not displayed for conciseness.

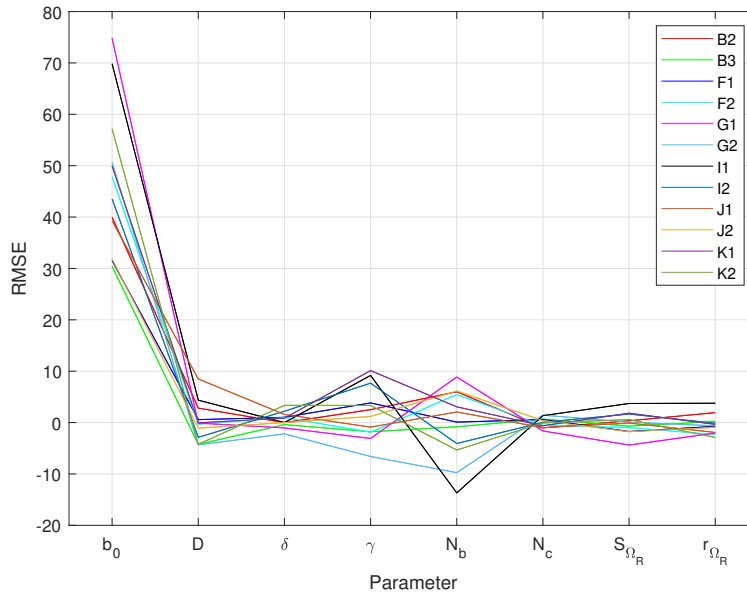


Figure 6.17: Mean b_0 and main effects for the seven parameters and twelve selected measurements obtained from the factorial analysis.

Those results indicate that the largest influence in the range of tested values is due to γ , Nb and D . It also shows that the interaction coefficient between γ and Nb is important too ($mean = 0.52, STD = 4.21$). However, no clear trend can be observed about which direction

Parameter	Mean coefficient	STD coefficient
D	-0.084	3.85
δ	0.53	1.47
γ	1.96	5.16
Nb	-0.17	6.85
Nc	-0.032	0.96
S_{Ω_r}	-0.12	2.07
r_{Ω_T}	-0.51	1.87

Table 6.2: Coefficients weights of the seven parameters used in the factorial analysis .

those parameters tend to affect the RMSE value, as the STD is large and the mean mostly centered around 0.

The results show that with this algorithm, a good performance can be achieved for certain of the measurements. The attempts to find a satisfying tuning that could work on every measures, revealed that it is difficult at this point to find a unique set of parameters which might improve the performance of all measures. In particular, Nb , γ and D have the largest influence in the range of tested values, but a change in certain direction in those values might decrease the performance for some measurements and reduce it for other measurements. This indicates that the algorithms could benefit from a dynamic tuning of those three parameters depending on some characteristics of the accelerations retrieved from the sensors.

6.6.2 Other limitations and related perspectives for improvement

The acquisition of the data revealed that the algorithm is sensitive to displacements of the sensors. Indeed, if one of the sensors is displaced after calibration, the initial orientation of the sensors as computed during initialization becomes erroneous. This means that a differential is generated between the two sensors, and this offset might be interpreted as a constant speed movement. The dynamic definition of Ω_T as described in Eq. 6.19 should prevent to drift during static parts. Nevertheless, the differential created by a displacement of the sensor might result in overestimation for certain movements and underestimation of other movements which will create an accumulation of error with time. The recording of the data with the prototype included several hardware constraints (cables in particular) which might have made the measures particularly sensitive to this kind of issue. Real HAs may be less prone to this kind of displacements, as they are placed behind the ears without being attached to other devices.

The integration errors that might occur during the estimations are likely to accumulate over time, and generate a shift of the overall estimation. The current implementation would benefit from an absolute reference in the environment where it is used, that could punctually help re-defining the initial condition of the second stage of integration. HAs are usually equipped

with two microphones per device, so it is for example imaginable to use direction-of-arrival (DOA) algorithms with sound emitted from a known position in the room.

In the base theory behind the equations of this algorithm, the sensors are supposed to be approximately aligned with the center of rotation of the head. In reality, this is dependent on the morphology of the head of the user. Further investigations and developments should study the influence of this parameter in the results.

6.7 Conclusion

In this chapter, a method was proposed to estimate the azimuth orientation of a human head in the horizontal plane, using only two 3-axis accelerometers. The method was evaluated with measures of realistic movements made by human subjects, and acquired with the accelerometers of HAs.

The results are promising and a good performance can be achieved for a part of the measures as long as certain parameters are set individually, as revealed by a factorial analysis and a batch analysis investigating the effects of several parameters. The algorithm is sensitive to sensors displacement or wrong estimation of the initial orientation of the sensors, which both yield poor results. With the current implementation, it is not possible to find a unique set of the fixed parameters that would yield good performance for every measurements. This encourages investigation of the possibility to tune dynamically some of them, based on the acceleration data.

The motivation behind the design of this algorithm remains to estimate the yaw of the head to provide dynamic binaural synthesis in the horizontal plane. Hence, an interesting question is the effect of such a non-optimal head-tracking performance, on spatial hearing. The next chapter aims at studying how head-tracking artefacts such as a large latency or an estimation mismatch can affect attributes of spatial hearing. In particular, the influence on externalization and localization in azimuth were studied.

7 Effects of head-tracking artefacts on externalization and localization

This chapter aims at investigating the effect of head-tracking artefacts such as a latency or an amplitude mismatch, on the potential improvement in externalization that might be brought by dynamic binaural rendering. These artefacts are the main ones that could be observed when developing the head-tracking algorithm described in Chapter 6. Additionally a subsequent experiment aimed at measuring to what extent those artefacts could affect auditory localization.

7.1 Introduction

7.1.1 Context and motivation

Head-tracking coupled with head movements allows achieving dynamic binaural synthesis. By retrieving the listener's head movements and position with a head tracking device, binaural impulse response filters can be selected accordingly while achieving real time convolution. This enables rendering a realistic and plausible auditory experience of sound sources of which the position remains valid regardless of the listener's orientation, as explained in Section 2.2.5. Additionally, head-tracking can help resolving the occurrence of front-back confusions [162] and thus improve the localization performance, as explained in Section 2.2.5. Moreover several studies have shown that head-tracking combined with head movements can improve the perception of externalization [20, 74, 98].

In Chapter 6, a head-tracking algorithm based on two 3-axis accelerometers was developed. The algorithm design was constrained by the limitations of hearing devices, and thus could not rely on gyroscopes or magnetometers that are conventionally used to ensure a robust estimation of the orientation of the head in azimuth. Hence, tracking artefacts are usually encountered with this type of algorithm. The goal of this chapter is to evaluate the effects of

such artefacts on the perception of a non-individualized binaural synthesis algorithm designed for wearable binaural communication devices. While the studies in Chapter 4 and Chapter 5 addressed auditory distance perception, this study focused on auditory externalization. The effect on the performance in localization was also assessed.

7.1.2 Contribution of head-tracking on spatial perception in literature

Head-tracking and auditory externalization

In [20], listeners were equipped with a head-tracker that could be active or not, and were asked to either keep their heads as stationary as they could or rotate their head between -15° and 15° . The participants were asked to make a binary decision for each stimulus, answering whether the stimulus was perceived outside or inside their head. The stimuli were generated using individualized HRTFs, so called "head-absent" transfer functions (HATFs) measured with a pair of microphones placed on a bar, and several mixes of the two conditions obtained with linear interpolation with a certain percentage of each transfer functions. The initial position of the sound source was between -25° and 25° . Six subjects took part in the experiment, the stimuli were 3 s long. They found that, in the case of individualized HRTFs, head movement without head-tracking drastically reduced the perceived externalization. However the combination of head-tracking with head movements did not significantly improve externalization compared to conditions with no head movements. Finally, the results also showed that head-tracking combined with head movement could improve externalization in the case of HATFs, suggesting that head-tracking brings a larger potential improvement in the case of non-individualized binaural synthesis.

Hendricks et al. [74] found that head movements combined with head-tracking could significantly improve the perception of externalization in non-individualized binaural synthesis, for frontal and rear sources in particular. The protocol consisted in achieving a controlled large motion for every stimulus, and included more time, i.e. 8 s, to achieve the movement compared to previous studies which used shorter times for every stimulus, i.e. less than 3 s. The participants rated externalization after completing the movement while they remained static. This was intended to ensure that potential improvements provided by the head movement was not only due to the lateralization, and remained after completion of the movement. They were using a six-level scale, and evaluated four conditions combining the availability or not of head-tracking and the performance or not of a large movement. It was found that head-tracking in combination with head movements substantially increased externalization compared to the conditions for which the listeners did not move their head, which contradicts the results of [20]. Head-tracking combined with head movements also enhanced externalization for lateral and frontal sources compared to the condition when the listener moves the head but the head-tracking is inactive, as found in [20]. The results also suggest that large head movements might be necessary.

In [98], a similar protocol was used, but the experiment also included source movements with

various trajectories while the head remained stationary, in addition to conditions where the source was static and various types of head movements were performed. The source was always frontal, and the listeners rated externalization using a four-level categorical scale. The main conclusion was that for both source and head movements, only large movements did increase the perception of externalization, whereas small movements did not affect it.

A few earlier studies suggested that head tracking combined with head movements might have a weak effect or no effect at all on the perception of externalization [166, 13]. Nevertheless, in [13] lateral and frontal azimuths were used and mixed in the analysis, hence the potential improvement brought by head-tracking and head movement might have been lost in the results. Moreover, the head movements were probably too short to allow the listener to take advantage of them, as suggested by the authors. The same limitation can be mentioned in [166] where only a very moderate effect of head-movement on externalization was found. Other early studies have suggested on the contrary that head movement might help in the perception of externalization, but the data was not quantitative enough to provide a strong statement [108, 83, 125].

In [101], non-individual BRIRs were measured in a listening room and truncated to different lengths (from 2.5 ms to 120 ms). Speech and music signals were convolved with those BRIRs, and the resulting binaural signals were presented over headphones to eight participants, who were performing or not a large head motion. The results suggest that the improvement in perceived externalization with head movements was smaller for longer BRIR lengths. The study concluded that head movements combined with head-tracking can significantly improve the perceived externalization for virtual sound sources synthesized with short BRIRs, for frontal sound sources in particular. In their study, this corresponded to BRIRs that were shorter than 10 ms. For longer BRIRs, they found that head movements had no influence on the perceived externalization of virtual sound sources. Such stimuli may indeed already be well externalized if enough early reflections (ERs) are available (see Section 2.6.5).

Head-tracking and localization

Wallach [162] suggested that dynamic cues associated with head movements, such as ITDs and ILDs were necessary to resolve the front-back confusion, also called reversals, in the auditory localization of a sound source. Fixing the head of a listener has been shown to yield a large increase in front-back confusions [171].

Multiple studies showed that when listeners are free to move their heads, they are more accurate at locating a sound source than when their heads is fixed or constrained, mostly owing to the front-back resolution [155, 128, 135, 76]. In [13] it was found that head movements coupled with head tracking improves localization performance of virtual sources compared to static rendering (after compensation of the reversals), with a larger improvement for non-individualized HRTFs.

Impact of head-tracking artefacts on spatial perception

In the literature, the studies assessing the effect of head-tracking artefacts on the perception of binaural spatialization mainly focused on the effect of head-tracking latency. In [26], two studies about the effect of head-tracking latency in the context of virtual audio display were conducted. The first one investigated the impact of head-tracking latency on the localization of broadband sounds. They found an increase of localization errors for brief sounds with latency values larger than 70 ms and an increase in the time required to locate a continuous sound source with latencies larger than 90 ms. The results suggest that head-tracking latency values lower than 60 ms might be acceptable for most virtual audio applications. In the second study, they found that some listeners were able to detect latency values of 60–70 ms for isolated sounds. In a second part of this experiment, the delayed target sounds were presented in conjunction with a reference tone with minimal possible latency, that was co-located with the virtual sound source. In this case, their detection thresholds were approximately 25 ms lower. Hence, they suggest a latency of 30 ms or less should be difficult to detect even in complex virtual auditory scenarios. In [104], only a single source was used, and the values for detection of latency had a mean and standard deviations of about 100 ms and 30 ms respectively (pooled threshold values), and the minimum detected was around 50 ms. The nature of the stimulus, as well as the reverberant vs. anechoic condition did not affect those results. In [153], the stimuli consisted of whether a single frontal sound source or a complex sound scene including five sources. The study aimed at investigating to which extent the spatial stability of sound sources in binaural reproduction was influenced by head-tracking latency. The results support that with an increase in latency, the source instability was more audible with the single source. In this case, the threshold was 10 ms lower compared to the complex sound scene. In [143], the effect of head-tracking artefacts on localization performance was investigated with an anechoic binaural spatialization. It was found that the average localization accuracy was not significantly degraded until the system latency reached a threshold of 96 ms and the tracking update rate was decreased down to a threshold of 10 Hz. The evaluation of the effect of head-tracking latency on externalization was found in only one study by Wenzel [164]. The author found that latency up to 500 ms did not affect externalization.

The above described studies mainly concerned latency and its effect on the perceived stability of a sound source and accuracy in localization. In addition to latency, the study of this chapter deals with the effect of a head-tracking estimation mismatch. The effect of such artefacts on auditory externalization is addressed in this chapter. In addition, the influence on the performance in azimuth localization is also addressed.

7.1.3 Goal

The present study evaluates to what extent head-tracking artefacts potentially happening in low-cost head-tracking algorithms, might affect the spatial perception. In particular, this study describes two experiments, which aim to assess the effect of those artefacts on auditory

externalization and performance in localization respectively. The binaural rendering used in this experiment is also adapted to the constraints of this type of devices, with the use of non-individualized and low-cost spatialization algorithms. The results were expected to show the advantages of the head-tracking algorithm described in Chapter 6 for binaural synthesis in the context of RM systems for wearable devices.

7.2 Spatialization algorithm and simulated artefacts

7.2.1 Binaural synthesis algorithm

The algorithm used in this experiment is designed to spatialize a RM signal. It consists in superimposing synthesized non-individual ERs to a direct sound generated with anechoic generic HRTFs as depicted in Fig. 7.1.

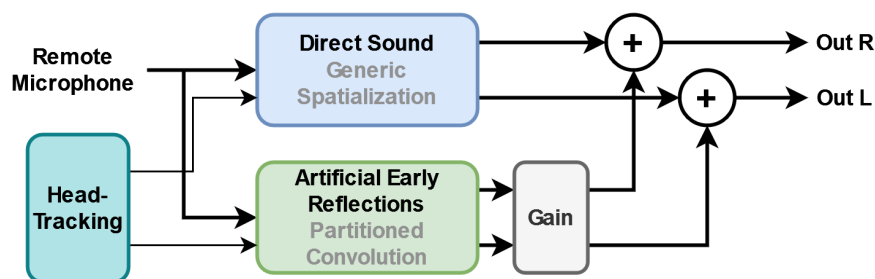


Figure 7.1: Block diagram of the non-individualized spatialization algorithm used in this experiment.

The spatialized direct sound is generated using non-individual HRTFS approximated by minimum-phase filter as described in Section 2.3. The ERs are synthesized in real-time by using a uniform partitioned convolution algorithm [62], as implemented in [157]. More details about the partitioned convolution algorithm implemented during this thesis can be found in Section 3.4. The RM signal was convolved with a 10-ms truncated pair of binaural room impulse responses (BRIRs). This short length is chosen to limit computational cost and memory usage. Originally, it was planned to test other length values, but informal pre-tests suggested that a single value could be tested. Pairs of BRIRs corresponding to each orientation in the azimuth plan with a step of 10° and a distance of 2 m were used in this experiment. The head-tracking device enables to retrieve the yaw orientation of the head and spatialize the sound accordingly, i.e. to select the correct minimum-phase filter and pure delay value in the direct sound, and the correct pair of BRIRs for the artificial ERs. The independent gain between the direct sound and the ERs allows to tune independently each part to achieve any desired direct-to-reverberant ratio (DRR).

7.2.2 Head-tracking conditions

Three main head-tracking conditions are described in this section: a reference, as well as two different types of artefacts. These artefacts were simulated from a reference tracking in order to obtain replicable artefacts across stimuli and subjects.

Reference

When all the necessary sensors, i.e. gyroscopes, accelerometers and magnetometers are available, a reliable and accurate estimation of the head yaw position can be achieved. For this, a low-latency IMU/AHRS device (NGIMU¹) was used to obtain a reference estimation. The sampling rate was set to 50 Hz for the data acquisition from the device to a dedicated Simulink model.

Latency

The first type of artefact consists of a simple delay compared to the reference. When listening to a sound source in real life, there is no latency between the movements of the listener and the consequent changes in the sound reaching their eardrums. However, most virtual display systems introduce a certain amount of delay due to the inherent latency of the tracking device itself, the communication delay between that device and the audio display, the selection of the appropriate HRTFs, and the subsequent audio processing. In this study, the latency was simulated by applying a delay on the reference estimation. After informal pre-test sessions, it was decided to test a single and large latency value of 400 ms in the experiments. This value is larger than any potential latency occurring in actual HAs with RM system implementations. An example measurement of the reference and the related latency simulation is depicted for a simple motion in Fig. 7.2.

Amplitude mismatch

In Chapter 6, a head-tracking algorithm relying solely on the use of two 3-axis accelerometers was developed. One of the main limitations from this algorithm comes from the difficulty in tracking slow parts of the motion, as the low values of accelerations have amplitude too small to be distinguished from the noise of the accelerometers. Hence, the algorithm relies on freezing the computation when such values of accelerations are reached to avoid drifting associated with the integration of noise. This results in parts of the motion being missed and an underestimation of the absolute values of the yaw estimation. As using the actual HAs prototype with the embedded accelerometers and the developed algorithm would lead to excessive unpredictability and variation in the artefacts experienced by the listeners, it was decided to simulate an artefact mimicking the behavior of the developed algorithm from the reference estimation obtained with the IMU/AHRS device.

¹<https://x-io.co.uk/ngimu/>

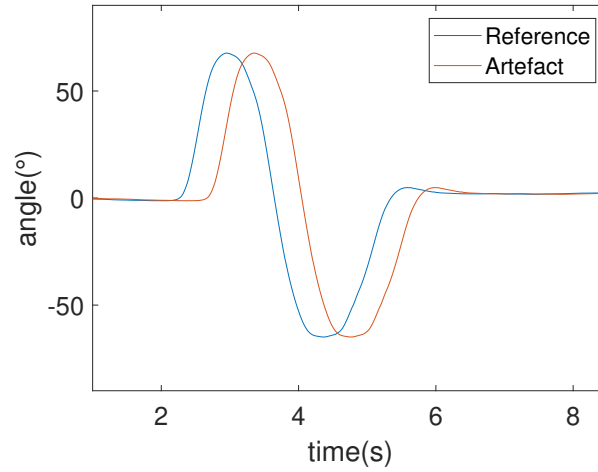


Figure 7.2: Example of the simulation of the latency artefact, in this case with a delay of 400 ms.

A satisfying method yielding to very close results was implemented using the following algorithmic computation of the artefact version yaw_{est} from the reference estimation yaw_{ref} :

Simulation of the amplitude mismatch

```

 $v(n) = yaw_{ref}(n) - yaw_{ref}(n-1)$ 
 $v_{diff}(n) = |v(n)| - |v(n-1)|$ 
if  $sgn(v_{diff}(n)) \neq sgn(v_{diff}(n-1))$  &  $v_{diff}(n) < 0$  then
     $v_{max} = v(n)$ 
    if  $|v_{max}| > Thr$  then
         $Thr = \gamma_v |v_{max}|$ 
    end if
end if
if  $|v(n)| \geq Thr$  then
     $yaw_{est}(n) = yaw_{est}(n-1) + v(n)$ 
else
     $yaw_{est}(n) = yaw_{est}(n-1)$ 
end if

```

The value of γ_v was set empirically to 0.8 which leads to similar range of errors as the algorithm described in Chapter 6. An example of the resulting artefact estimation is depicted in comparison to the reference it was computed from in Fig. 7.3.

7.3 Participants

30 naive listeners took part in both experiments (15 female, 15 male, avg age = 23.9 y.o.). All the listener were self-reported as having NH.

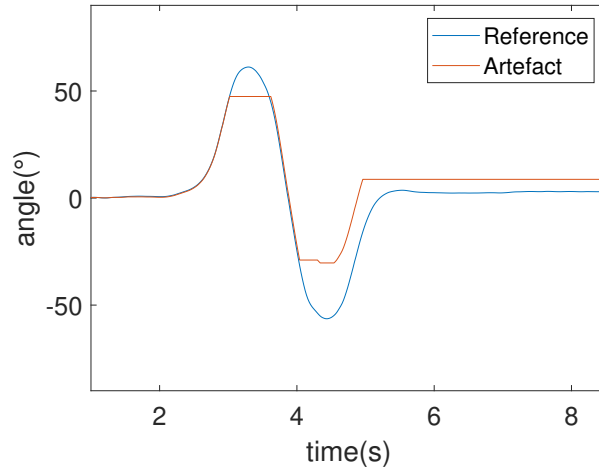


Figure 7.3: Example of the simulation of the amplitude mismatch artefact.

For each participant, the two experiments were performed on different days, in order to limit fatigue. Half of the participants started with the externalization experiment, while the other half started with the localization experiment. A minimum of one week between the two experiments was ensured for each participant, to limit the effect of the training with the non-individualized HRTFs and BRIRs used in the binaural synthesis. 25 subjects had a gap of one or two weeks between the experiments, 2 had a gap of about one month and 3 had a gap of about two months.

7.4 Externalization experiment protocol

7.4.1 Setup

The experimental setup was installed in a listening room (volume = 125 m³, RT₆₀ = 0.17 s). All stimuli were played through a pair of open headphones (Audeze LCD-2C) driven by a headphones amplifier (Lake People HPA RS 02). A low latency audio interface (RME Babyface Pro Fs) was used to playback the sounds from the Simulink implementation of the spatialization algorithm. The IMU/AHRS device was mounted on the top of the headphones in order to track the head motions. To increase stability and remove the chance of missing bit in the data transmission, USB COM Port communication was preferred over WiFi. The sound pressure level was adjusted to 65 dB A. The participant could ask for a slight level adjustment for comfort if necessary, but all participants were comfortable with the proposed original playback level.

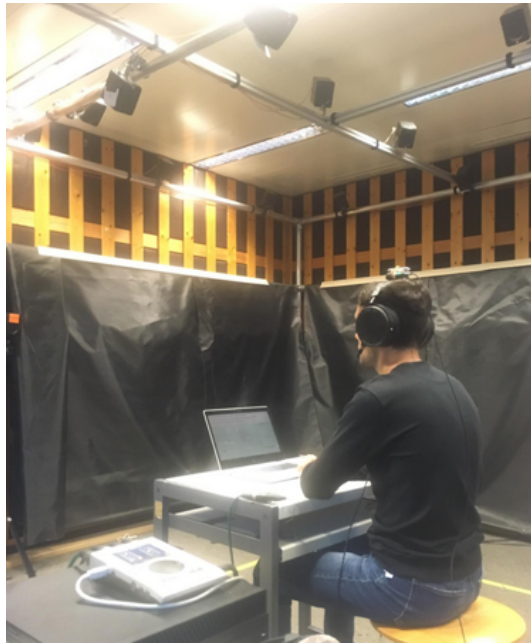


Figure 7.4: Participant equipped with the headphones and head-tracking device in the listening room.

7.4.2 Preparation / Training

Prior to the actual experiment, the listeners had a short introduction to sound externalization. They used an interface on which they could play and listen to two versions of a pre-recorded audio file. The first version was achieved using a pair of binaural microphones placed on the artificial ears of a KEMAR manikin, while the second version was a mono reduction of this recording. This recording was made in the same room than the one in which the experiment was taking place in order to ensure room congruence for this demo. The content of the recording consisted of 70 s of a male voice that was moving around the listener and briefly giving explanations about the experiment. In particular, the recording highlighted the subtle difference between auditory externalization and auditory distance. For this purpose, at some point of the audio file, the voice was moving from very close of the ear of the manikin to the corner of the room, the goal being to showcase that a variation of the distance could still be perceived in the mono recording, while the sound clearly remained perceived as externalized. The listener was asked to listen entirely to each recording first, and then they could listen as much as they want to both recordings and could switch between them. The goal of this short introduction was first to ensure that they understood the meaning of externalization, and check that they were able perceive it. The second goal was to give them references points for the evaluation, of what could be a very well externalized sound in comparison to a sound perceived as internalized. After this introduction the experimenter gave further explanations about externalization and on the distinction between auditory externalization and distance, which appeared clear for all participants after this demo.

7.4.3 Stimuli and conditions

The base stimulus consisted of a 8.5 s excerpt from an anechoic male speech recording in English. The same sentence was used for every run. The first samples of the BRIRs (corresponding to the direct sound) were set to zeros to avoid the superimposition with the direct sound spatialized using generic HRTFs. The BRIRs were truncated to 10 ms for every stimulus. This value was chosen as the constraints of wearable devices encourage to limit the memory usage, and thus investigate the possibility to provide externalization with short BRIRs. The value was determined during informal pre-tests in which the ER time was not found to be the most influential parameter. Moreover this value is mentioned as a threshold in [101], below which head-tracking might be more beneficial for externalization. The different head-tracking conditions ("no head-tracking", "head-tracking latency of 0.4 s" and "head-tracking amplitude mismatch") are denoted respectively: **No HT**, **Lat 0.4** and **Amp. Artefact**.

For this part of the experiment, the effect of the direct-to-reverberant ratio (DRR) as well as the effect of a low-pass filtering were also tested. The following conditions were applied:

Parameter	Setting
Head-tracking condition	No HT - Reference - Lat 0.4 - Amp. Artefact
DRR	2 dB - 5 dB - 30 dB
Low-pass filter cut-off	7 kHz - 10 kHz

Table 7.1: List of conditions used in the externalization experiment.

It should be noted that the DRR values of 2 dB, 5 dB and 30 dB correspond to the full length BRIRs, the true values with the 10-ms truncation are respectively 5.34 dB, 8.36 dB and 33.35 dB. The output was level-normalized in the real-time model with a gain at the output to obtain a consistent level across every DRR setting. The laterality of the sound source has a substantial effect on the perception of externalization [81]. All stimuli evaluated in this experiment were simulated in front of the listener at an azimuth of 0°.

A full factorial design was used for this experiment, each subject rated every combinations of the conditions described in this section. Each of these 24 combinations was presented with 4 repetitions for every participant, and the order was randomized inside each of the 4 repetition blocks. Hence, with the 18 additional training stimuli, each participant had to evaluate a total of 114 stimuli. The experiment lasted about 60 min (explanations and breaks included). A mandatory break after half of the stimuli had been evaluated was imposed to avoid listening fatigue and maintain focus of the participant due to the task being quite repetitive. The participant could have a break at any time, upon request.

7.4.4 Task

The subject was asked to perform the same motion for every stimulus. First they were looking in front of them at a mark placed on the wall at 0° . Then they pressed the "play" button on the interface. As soon as the speech was heard they were asked to turn their head, and look at a mark placed on the wall at $+80^\circ$ first, then turn the head to the other side to look at a mark placed at -80° , and finally point back at the initial position at 0° . They were asked to perform this motion in synchronisation with the words of the speech sample, in particular and most importantly, they were asked to be back at the center position for a specific word in the sentence. This was to ensure that they heard the last 2 seconds of the speech stimulus as a source in front of them while they remained still. This protocol was inspired by the one used in [74]. This protocol ensures that the potentially observed increase in externalization is not based on the lateralization of the sound sources while the listener is turning the head from it. Instead, it aims at assessing if the cues provided by the movement of the listener yield a persistent impression of externalization after the movement.

Frontal sources are known to be perceived as less externalized, and were shown to potentially benefit more from the addition of head-tracking in comparison to lateral sources [74]. After performing the instructed movement in this experiment, the amplitude mismatch artefact can lead to a final angle that is different from 0° . To prevent this from biasing the result, the artefact simulation was adapted for this part of the experiment. At a certain time $t = 5.5$ s, the tracking simulation was set to smoothly converge to fit the reference tracking. This was to ensure that, after completing the movement, the final angle was not larger for the **Amp. Artefact** condition compared to the other conditions. A short silence was included in the recording from $t = 5.5$ s, so that this compensation did not affect the spatial processing. An example of the resulting tracking is depicted in Fig. 7.5 derived from the original artefact simulation pictured in Fig. 7.3.

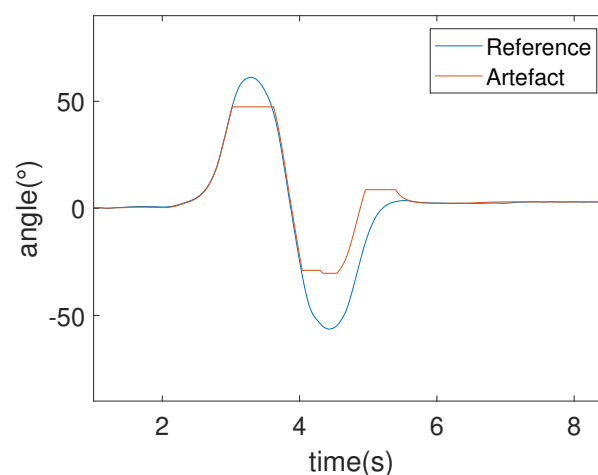


Figure 7.5: Example of the simulation of the amplitude mismatch artefact adapted to the externalization experiment.

For the participants, understanding how to perform this motion in synchronisation with the speech sample took between three and eight runs. The first 18 runs of the experiment served only as training runs and were not included in the results. This number was defined during informal pre-test sessions, which enabled investigation of how many runs it took for the participants to perform the task easily and to focus mainly on their perception of externalization. The experimenter continuously monitored that correct synchronization of the motion was achieved.

Listeners were asked to rate the degree of externalization perceived at the end the motion, for the last 2 seconds of the speech stimulus when they were back at the center and remained still. A diotic 2 s pink noise sample was played between every stimulus. This served as a perceptual memory "reset" to limit the effect of the order of presentation, so that the spatial attributes of the previous stimulus would not influence the perception of the next one.

To rate the externalization, the listeners had to use a continuous scale labeled at the extremities from "completely internalized" to "completely externalized", which corresponded to ratings of 0% and 100%. The listeners had to move the cursor from an initial 50% position to give their rating and then validate it with a dedicated button. A similar scale was used to rate externalization in [97]. This method was preferred over categorical scales such as the ones used in [74, 98, 103], or scales with visual references used in auditory distance estimation experiments [41] as no auditory or visual reference was available to the listener. Moreover, the design of the experiment did not allow to use MUSHRA-style interface or paired comparisons, as it would require too much memory effort for the listener to remember accurately the perceived externalization of the previous stimuli after completing the movement of the next ones. Moreover the duration might have been too long with paired comparisons with the number of tested condition. In [97], the listeners had to close their eyes. The purpose was to enable the listener to rate externalization without being constrained by auditory distance matching with visual references. For the present study, listeners kept their eyes open, as it was necessary to ensure that they were accurately at 0° after the movement, which would have been too complicated to achieve without sight. With available visual cues, it is possible that a stimulus for which the binaural cues for externalization are preserved, but the cues associated with perceived distance are distorted (e.g. the DRR) might never be given a 100% externalization rating, even if is neatly perceived outside the head. This effect might have been partially limited in the present study as no potential visual reference to match was proposed to the listener.

7.5 Externalization experiment results

7.5.1 Auditory externalization

The raw results from this experiment naturally exhibit significant standard deviation differences from one subject to another as each subject can use the scale in a different manner. For

example, a subject could rate every stimulus within a small range of the scale, while another one could be using the full scale. In order to normalize the results and focus on the relative effect of the parameters on the perception of externalization rather than on absolute ratings, these ratings were transformed into z-scores [93] for each subject. The z-score were calculated for each subject by subtracting the mean from each raw rating and then dividing the difference by the standard deviation of this subject's ratings. With this transform, no conclusion can be made on the absolute externalization ratings, but only on the relative ratings. To use this transformation, it is assumed that the externalization scale can be considered as interval-level. Indeed, the training aimed at underlining that the task consisted in rating externalization independently from auditory distance. This was to ensure that the scale would not be confused with one associated with auditory distance in rooms which is usually considered to be non-linear [23, 122]. The z-score ratings are displayed in Fig. 7.6.

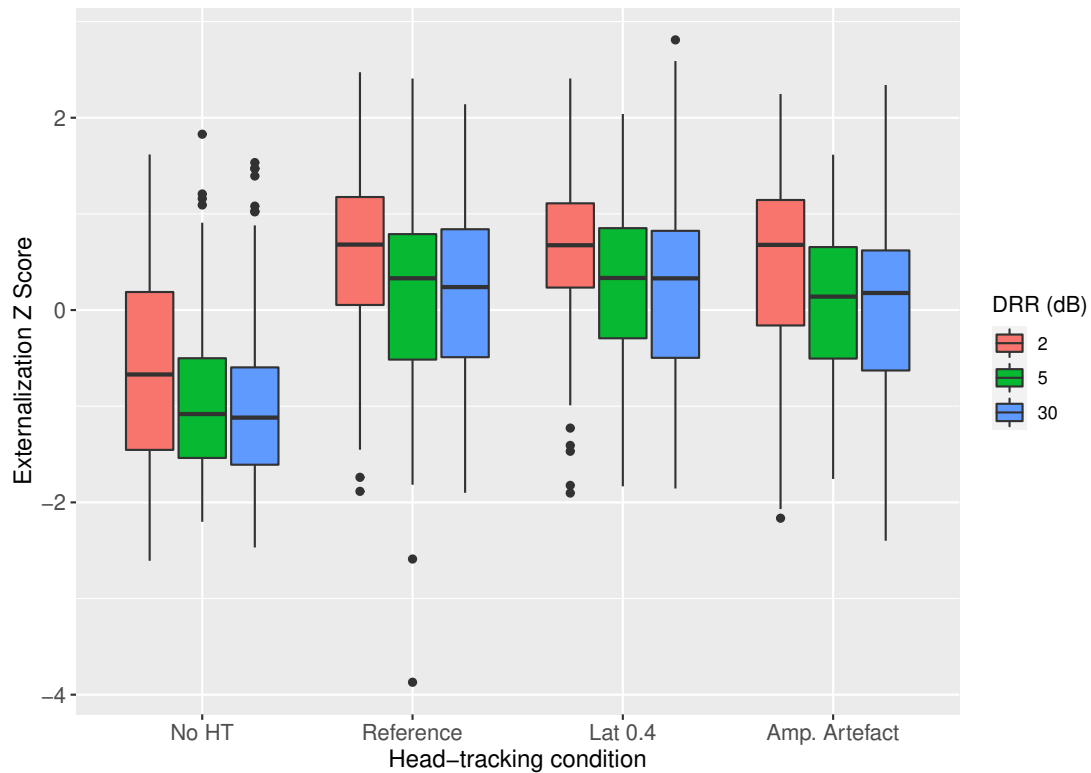


Figure 7.6: Externalization ratings per head-tracking condition and DRR setting (z-scores transformed).

The normal distribution was visually checked by plotting the QQ-plots of the residuals for each of the independent variables (see appendix Fig. A.4 and the corresponding histograms in Fig. A.3 for the head-tracking condition). Indeed, for large sample sizes, as in this experiment, slight deviations from the normal distribution can lead to a significant result in evaluations such as the Shapiro Wilk test. Thus visual checks of the histograms and QQ-plots can be performed alternatively [57]. The distribution of the **No HT** condition is slightly skewed, never-

theless, the literature suggests that as long as the group sizes are equal and the observations are independent, the ANOVA can be used and is robust against deviations of normality [65, 51]. A repeated-measures ANOVA was conducted with four within-subject factors: the head-tracking condition, the DRR, the low-pass filter value and the repetition number. The assumption for sphericity was checked using Mauchly's test, and the Greenhouse–Geisser correction was applied when necessary. Significant effects were found for the processing [$F(3, 168) = 112.17$; $p < 0.001$] and the DRR [$F(2, 112) = 73.55$; $p < 0.001$]. The complete ANOVA table is available in appendix in Tab. A.2. The effect size was evaluated using the eta-squared method. Following Cohen's rule of thumb, a large effect size was found for both the head-tracking condition ($\eta^2_{\text{partial}} = 0.58$) and the DRR ($\eta^2_{\text{partial}} = 0.17$). No significant effect was found for the low-pass filter value [$F(1, 56) = 0.929$; $p = 0.339$] and the repetition number [$F(3, 168) = 2.83$; $p = 0.066$], so the score for the two cutting frequency and all four repetitions were mixed for the rest of the analysis and on the plot in Fig. 7.6. No interaction was found between the head-tracking condition and the DRR value [$F(6, 336) = 0.206$; $p = 0.97$].

A *post hoc* Tukey's HSD analysis was conducted with a 95 percent confidence level for the two significant independent variables, i.e. the head-tracking condition and the DRR. It was found that the **No HT** condition was significantly less externalized than the **Reference** ($p < 0.001$), the **Lat 0.4** condition ($p < 0.001$) and the **Amp. Artefact** condition ($p < 0.001$). This observation remains true regardless of the DRR and even the setting with **No HT** and DRR = 2 dB is significantly rated as lower compared to any of the other conditions ($p < 0.001$ in all cases). No significant difference was found between the **Reference** and the **Lat 0.4** conditions ($p = 0.54$) and the **Reference** and the **Amp. Artefact** conditions ($p = 0.099$). Finally, with a small but significant difference, the **Lat 0.4** was perceived as more externalized compared to the **Amp. Artefact** ($p = 0.017$). Nevertheless, by comparing between the two latter conditions, a significant difference is only observed for the DRR = 5 dB stimuli, where the **Amp. Artefact** stimuli were perceived slightly less externalized compared to the **Lat 0.4** stimuli ($p = 0.027$).

Within each head-tracking condition group, the stimuli with DRR = 2 dB were perceived as more externalized compared to both the DRR = 5 dB ($p < 0.001$ for **No HT**, **Reference**, **Lat 0.4** and **Amp. Artefact**) and DRR = 30 dB ($p < 0.001$ for **No HT**, **Reference**, **Lat 0.4** and **Amp. Artefact**) stimuli, whereas no significant difference was found between the DRR = 5 dB and DRR = 30 dB stimuli for every condition.

7.5.2 Head movements

The head trajectories were recorded during the experiment for two purposes. The first is to verify that the participants could follow the instructions precisely enough, especially in terms of timing with the speech sample. The second is to assess if certain parameters of the head motion had an influence on the externalization perception. The following dependent variables were computed from the trajectories recordings: the maximum (right) and minimum (left) azimuth angle reached during the motion, the total amplitude of the motion, the final angle

reached at the end of the motion (averaged over the last second) and the time spent at azimuth 0° at the end of the motion. The medians and interquartile ranges on the global data for those variables are summarized in Tab. 7.2.

	Median	Interquartile Range	Spearman's ρ
Maximum angle (right)	61.92°	12.74°	0.017
Minimum angle (left)	-62.18°	15.77°	0.019
Amplitude of the motion	124.24°	26.14°	0.021
Final azimuth angle	0.92°	2.52°	0.061
Time left after movement	2.23 s	0.28 s	0.017

Table 7.2: Medians, interquartile ranges and Spearman's ρ in relation to the z-score of externalization for several dependent variables of the participant's movements.

These results suggest that the participants successfully managed to achieve the task as instructed by the experimenter. As can be seen in Tab. 7.2, no correlation was found between any of dependent variables linked with the amplitude of the movement and the perceived externalization (negligible Spearman's ρ), which suggest that the variability and range in terms of amplitude of the movement in azimuth was small enough to not affect the externalization ratings. No correlation was found either with the final azimuth angle, which remained rather small, and did not affect the externalization results. Finally the time left after movement, i.e. the time when the participant was static and had to evaluate externalization, did not affect externalization either. The participants very rarely arrived too late (after the silence gap before the last part of the sentence), so no bias was created by participants failing to synchronize with the motion. Even though the 0.5 s silence gap before the last part of the speech on which externalization was evaluated should be enough in most cases to compensate for the latency in the **Lat 0.4** condition, this makes this condition potentially more sensitive to the ability of the participant to synchronize with the speech. Nevertheless, no correlation was found for the subgroup of the **Lat 0.4** condition either (Spearman's $\rho = 0.038$), suggesting that the silence gap was probably enough to compensate for the latency and that it did not affect the perception of externalization for this condition.

The errors of tracking simulated in the **Amp. Artefact** condition can vary from run to run, depending on the movement of the listeners. A score of distance to the related reference tracking was computed for every run in the **Amp. Artefact** condition by summing differences between the two trajectories sample by sample. No correlation was found between the distance to the reference and the externalization ratings (Spearman's $\rho = 0.006$).

7.6 Externalization experiment discussion

7.6.1 Effect of the head-tracking condition

In this experiment, in which large movements were performed and only frontal azimuth sources were simulated, an increase in the perceived externalization was obtained for all conditions that included head-tracking. This is in agreement with [74] and [98], in which it has been shown that head-tracking combined with head movements provided an advantage for the perception of externalization if sources were frontal and movements were large.

In every condition where head-tracking was available, even with latency of amplitude mismatch, the additional cues provided to the listener by the head-tracking always helped the listeners to better externalize compared to the situation where no head-tracking was available. Additionally, in particular, in the **Amp. Artefact** condition, the amount of error did not cause a reduction in the perception of externalization compared to the **Reference** tracking. This could be explained by the fact the listener was still exposed to similar cues to the ones available in the **Reference** condition, especially when the listener reaches the most lateral azimuths. As no correlation was found between externalization ratings and the amplitude and maximum values in azimuth reached, it is likely that the underestimation of the tracking usually happening in the **Amp. Artefact** condition did not prevent the listener from externalizing either. Concerning latency in particular, this confirms what was found in [164], in which latency up to 500 ms did not have an influence on externalization.

It was clearly stated to the participant that the simulated speech source was in front of them at an azimuth of 0°, which was materialized by a visual mark. In addition, they were instructed that the sound source was not supposed to move. Because of the mismatch of estimation and the resulting dynamic binaural synthesis, a "slewing" of the perceived sound direction is likely to be experienced by the listener. It can be hypothesized that this was not enough to affect plausibility, which can be degraded in the case of non-matching audio-visual presentations. Indeed, this "slewing" might have been perceived by the auditory system as if the source was moving, regardless of the instruction that the sound source should remain stable. Moreover, Li et al. [101] suggested that source movements can increase the perception of externalization. In this case, there is no reason to think that the considered artefacts should have affected externalization significantly. Indeed, one could imagine that if the sound source would be made more realistic with a visual reference, plausibility could be more severely affected by tracking artefacts. Consequently, externalization could be affected too. This could be done e.g. with a real person or a video of a speaker displayed on a screen placed in the direction the sound is supposed to be coming from. The same remark can be made for the **Lat 0.4** stimuli. This should be tested in further studies.

7.6.2 Effect of the early reflections and binaural cues

Early reflections play an essential role in the perception of auditory externalization and the length of the impulse response affects externalization [45], with a larger influence up to approximately 30-40 ms [149]. For long BRIRs, a recent study suggests that head-tracking and head movements might not provide a very substantial increase in externalization compared to shorter BRIRs [101]. The present study investigated the possibility to increase externalization with 10-ms truncated non-individualized BRIRs measured in a listening room superimposed to a generic direct sound. In [74], the spatialization was non-individualized, and the authors intentionally chose a room with "not too much reverberation" to record their stimuli. The full-length reverberation was used in the latter, for a room having a $RT_{60} = 0.24s$, i.e. slightly larger but comparable to the listening room of the present study ($RT_{60} = 0.17s$). Conversely no reverberation was available in [98], but generic HRTFs were also used.

Every value mentioned for the DRR in this section were computed from the full length BRIRs (360 ms). Hence with the 10-ms truncation applied to the BRIRs, the DRR values used in the experiment (2 dB, 5 dB and 30 dB) correspond respectively to the DRR equivalent values of 5.34 dB, 8.36 dB and 33.35 dB. A clear effect of the DRR on the perception of externalization was observed, with the DRR = 2 dB stimuli being perceived as significantly more externalized compared to both the DRR = 5 dB and DRR = 30 dB stimuli. Little to no difference was observed between the DRR = 5 dB and DRR = 30 dB conditions which could be explained by the BRIR time length which is rather short and in this case the ERs are probably not loud enough to provide a significant difference of externalization.

The results suggest that the cues derived during a large movement were sufficient to provide significantly more externalization, even though the spatialization was made using generic HRTFs for the direct sound and generic 10-ms truncated BRIRs for the ERs. In [20], results are in agreement and additionally suggest that with non-individualized HRTFs, the benefit brought by head-tracking can be larger than with individualized HRTFs. Nevertheless, individualized HRTFs were not tested in the experiment reported in this chapter.

Finally, the BRIRs used in this experiment were measured in the same listening room as the one in which the experiment took place. Room congruence has an important influence on auditory externalization [170], and it could be interesting to check the potential improvement in externalization provided by head movements, depending on the divergence between the playback room and the room in which BRIRs were measured.

7.6.3 Considerations for applications to binaural communication devices

The present study suggests that, in the context of wearable binaural communication devices, a significant improvement in perceived externalization can be brought when the listener is performing movements and that head-tracking is available. Indeed, in this study, a significant enhancement in externalization was provided using a simplified binaural rendering algorithm.

The spatialization used and described in this study is intentionally minimal, as it is designed to be potentially implementable on wearable devices. All HRTFs and BRIRs are generic, and the length of the BRIRs was always truncated to 10 ms.

In Chapter 6 an algorithm aiming at providing head-tracking with the constraints of wearable binaural communication devices was designed. Such algorithm has limitations that were described in the previous chapter, as it is relying solely on two 3-axis accelerometers. However, even if the head-tracking had a significant amplitude mismatch artefact, it did not affect the perception of externalization in the listening test of this study. A large latency (0.4 s), did not affect auditory externalization either. Such latency is much larger than any latency potentially occurring in the context of wearable binaural communication devices with RM. The design of the present test did not include any visual reference corresponding to the virtual sound source. Hence, further studies should try to investigate if the presence of such visual reference would make externalization collapse, as plausibility could be severely affected.

In application, when using partitioned convolution to generate ERs, the BRIRs would be fixed and stored in the wearable binaural communication devices. This means that the BRIRs might not always be congruent with the room in which the user is. This raises the question about the dependence between the degree of divergence and the potential improvement that head-tracking combined with head motions could provide to the listener. Nevertheless, some of the methods described in Chapter 3 aim at extracting ERs from the signal retrieved from the microphones available on the devices themselves, which mean that they would include the room information in every situation.

Externalization is known to be substantially influenced by visual cues and phenomena such as the ventriloquist effect [163]. As mentioned above, the experiment did not include a realistic visual reference for the sound source. Doing so could have decreased plausibility when the head-tracking artefacts yielded a perceptually moving auditory source while the visual reference was static. In a practical wearable binaural communication device application, the presence of a real sound source that is physically present and thus visible to the listener (such as a person speaking in a real life situation), might change how the tracking artefacts affect externalization, at least during movements. When the listener is not moving and the speaker is in the visual field, it is likely that ventriloquist effect could provide a certain amount of externalization. Hence, visual cues might increase externalization in every head-tracking condition and/or reduce the improvements in externalization provided by head-tracking. Future works concerning the effects of head-tracking artefacts on externalization should include more realistic visual references to answer those questions, and confirm that the artefacts of this study are indeed not problematic for the perception of externalization.

7.7 Localization experiment: protocol

7.7.1 Setup

The experiment took place in the same room and with the same setup as the externalization experiment. A part of the stimuli (real sources) were played through 8 loudspeakers (ELAC 301.2) placed around the listener and amplified using an 8-channel amplifier (Allen & Heath GR8A). The loudspeakers were placed on a rectangular frame present in the listening room, and hidden behind a black curtain. The transfer function of the curtains was measured in an anechoic room, and all sound played through the loudspeakers were compensated for the small attenuation in the high frequency due to the curtains. In addition, every stimulus played through the loudspeakers was compensated with a gain corresponding to the position of each loudspeaker, so that the level was constant and equal to 65 dB A at the listener's position. This served at minimizing the effect of the level cue, that plays a major role in distance perception, and thus might have biased the azimuth localization. The rest of the stimuli (virtual sources) were played through the same pair of open headphones and amplifier as in the first experiment. Moreover, the ERs of the virtual sources presented over headphones in the rest of the experiment are simulated with BRIRs with a constant distance from the listener. The same complete IMU/AHRS device as in the first experiment was mounted on the top of the headphones. In addition, a class 1 laser pointer was placed on top of the head-tracker in coincidence with its x-axis and was activated during this experiment. The laser beam was used by the listeners who could aim at the perceived location of the sound sources by pointing the head toward it. A schematic representation of the setup is depicted in Fig. 7.7.

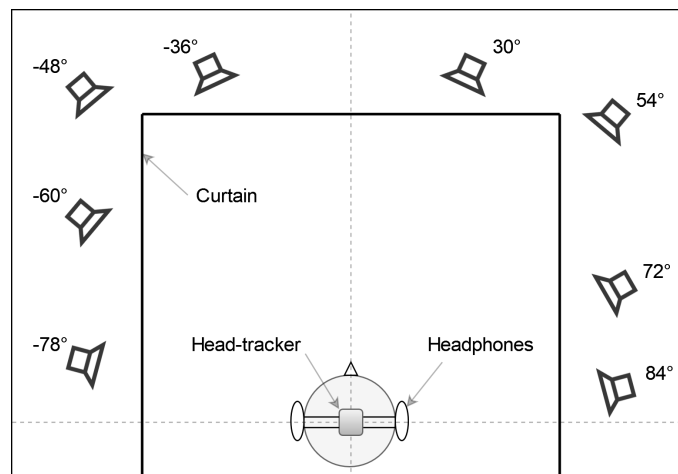


Figure 7.7: Schematic representation of the setup used of the localization experiment, with 8 loudspeakers hidden behind black acoustically transparent curtains, and the listener equipped with headphones and the IMU/AHRS head-tracking device.

7.7.2 Stimuli

The recording used to generate every stimulus in this part of the experiment consisted of a 25 s excerpt from an anechoic male speech recording in English. In this part of the experiment, the BRIR are still truncated to 10 ms. The low-pass filter value is fixed to 10 kHz and the DRR is fixed to 5 dB (8.36 equivalent with the 10-ms truncation). The independent variables are reported in Tab. 7.3.

Parameter	Setting
Head-tracking condition	Real sources - Reference - Lat 0.4 - Amp. Artefact
Target azimuth	-78°, -60°, -48°, -36°, 30°, 54°, 72°, 84°

Table 7.3: List of conditions used for the localization experiment.

Each combination of head-tracking condition and target azimuth was presented with 4 repetitions. 4 initial stimuli were used as training runs at the start of each of both the real sources and virtual sources sub-part of the experiment. Hence the subject had to locate a total of 36 sound sources for the real sources part and 100 sound sources in the virtual sources part. The total duration of the experiment was between 45 and 60 min depending on the participant's pace (explanations and breaks included). The subject could take a break at any time if necessary.

7.7.3 Procedure

First, the listener was asked to point toward a mark at azimuth 0°, in front of them. The laser helped them to achieve this precisely and easily. This served to initialize the head-tracking between each stimulus. Then, they could press the space bar on the computer keyboard to play the stimulus. As soon as the stimulus could be heard, they were instructed to locate the sound source, and point the head toward the perceived azimuth using the laser placed on their head. It was suggested in [8] that this method might be the most reliable one. They could then validate their answer by pressing the space bar again. The instruction by the experimenter stated that accuracy was the priority in this task, but mentioned that the time was measured too, and that they should validate right away when they were sure of the answer. After validation, a message appeared on the screen to remind the listener to point the laser at the mark in front of them to initialize the head-tracker before playing the next stimulus. For the virtual sources, a diotic 2 s pink noise sample was played after each validation. The aim of this was to limit the effect of the order of presentation, so that the previous stimulus minimally affect the next one. In addition, the experimenter informed the participant that the sound sources could only come from azimuths between -90° and +90°, i.e. the sound source could not come from rear positions.

7.8 Localization experiment: results

7.8.1 Localization error

The general results are reported in Fig. 7.8, with the boxplot of the absolute error for each processing.

The real sources were presented before the virtual sources in the experiment procedure, so they were not randomized in the same presentation block. Thus, results of the localization of real sources are only presented as an indicative value of the performance of the listeners and can be analyzed alone, but were not mixed together in the statistical analysis about the effect of the head-tracking condition and target azimuth. The medians and interquartile ranges are reported for each head-tracking condition in Tab. 7.4.

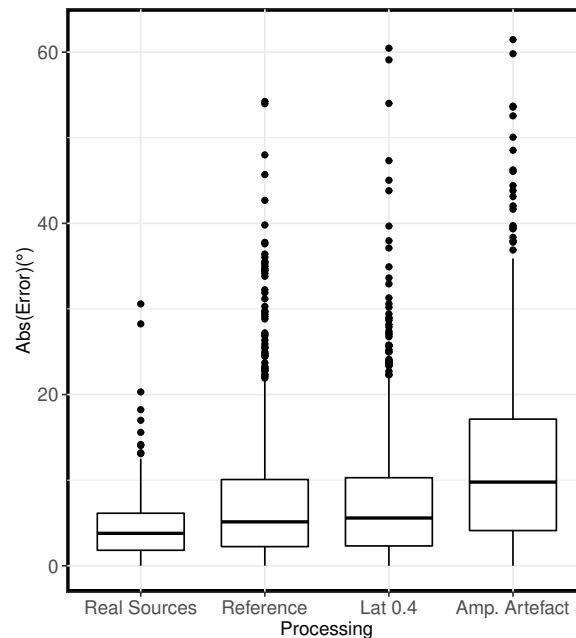


Figure 7.8: Absolute error of localization for each head-tracking condition.

	Median	Interquartile Range
Real Sources	3.80°	4.33°
Reference	5.15°	7.86°
Lat 0.4	5.59°	7.99°
Amp. Artefact	9.78°	13.02°

Table 7.4: Medians and interquartile ranges of the absolute error in azimuth localization for each head-tracking condition.

The absolute error data was transformed using a cubic root transformation in order to obtain

a normal distribution of the data, which was visually checked using QQ-plots of the residuals for each condition (see appendix Fig. A.6 and the corresponding histograms in Fig. A.5 for the head-tracking condition). All the statistical analysis was conducted on the transformed data, the values reported in the descriptive statistics in Tab. 7.4 are reported with the re-transformed data.

Real sources

The results indicate a median absolute error of 3.80° with an interquartile range of 4.33° . The results were not included in the ANOVA analysis as this condition was tested in a single block before the virtual sources, but it can still be noted that this condition yielded the best performance, and gives an idea of the performance of the listeners.

The listener was sitting on a non-rotating chair which position was fixed. The azimuth positioning of the loudspeakers was measured from the point where the listener's head was intended to be during the experiment. It is estimated that the position of the head of the listener compared to its expected position may vary only in a range of about ± 10 cm maximum. This results in a potential uncertainty of $\pm 2^\circ$ in the answers collected in the case of the real sources.

The choice was made to compensate the gain for each loudspeaker so that the level at the listener's position was constant and equal to 65 dBA. This was to limit the effect of the level cue and compensate for the positioning of the loudspeakers on a rectangular frame rather than a circular one. However, even if the loudspeakers were hidden, the shape of the frame could possibly influence the listener.

Virtual sources

The assumption for sphericity was checked using Mauchly's test, which showed that sphericity had not been violated for any of the independent variables. A repeated-measures ANOVA was conducted on the transformed data for the virtual sources, and a significant effect was found for the head-tracking condition [$F(2, 58) = 27.25$; $p < 0.001$]. Following Cohen's rule of thumb, a medium effect size was found for the head-tracking condition ($\eta^2_{\text{partial}} = 0.10$). No effect was found for the repetition number [$F(3, 87) = 0.10$; $p = 0.96$]. The complete ANOVA table is available in appendix in Tab. A.3.

A *post hoc* Tukey's HSD analysis with a 95 percent confidence level was conducted on the head-tracking condition. It was found that the **Amp. Artefact** condition was located significantly less accurately compared to both the **Reference** ($p < 0.001$) condition and the **Lat 0.4** condition ($p < 0.001$). On the contrary, the results suggest that the listeners did not perform differently between the **Reference** and the **Lat 0.4** conditions ($p = 0.88$).

The number of times the listener switched the sign of their head trajectories while they were

trying to locate the sound source was computed out of the recorded trajectories. It could be expected that listeners might perform more movements when they have more difficulty to locate the sound. However, no correlation was found with the localization performance for the virtual sources (Spearman's $\rho = 0.033$) nor for the **Amp. Artefact** condition group (Spearman's $\rho = 0.011$). Additionally, an error score compared to the reference tracking was computed from the trajectories for the **Amp. Artefact** condition group, but the correlation with the localization performance was negligible (Spearman's $\rho = 0.14$)

Effect of the target azimuth

When looking at the mixed data for the virtual sources, a significant effect was found for the target azimuth [$F(7, 203) = 3.65$; $p < 0.001$], with a small-medium effect size ($\eta^2_{\text{partial}} = 0.050$). However, as pictured in Fig. 7.9, results suggest it is more interesting to look at the effect of the target azimuth for each head-tracking condition separately, as confirmed by the interaction between these variables [$F(14, 406) = 13.06$; $p < 0.001$]. The medians and interquartile ranges are summarized in Tab 7.5.

	-78°	-60°	-48°	-36°	30°	54°	72°	84°
Real Sources	5.72(4.73)°	5.21(4.60)°	2.97(4.5)°	2.22(3.32)°	1.65(2.65)°	3.43(2.67)°	5.51(4.01)°	4.92(3.68)°
Reference	5.85(9.70)°	4.72(7.19)°	5.36(7.48)°	4.52(7.98)°	5.36(7.46)°	5.66(7.27)°	5.60(7.44)°	4.77(8.07)°
Lat 0.4	5.62(5.86)°	5.72(7.78)°	5.94(7.74)°	4.88(8.39)°	6.13(8.88)°	5.51(8.28)°	5.55(6.60)°	5.21(9.35)°
Amp. Artefact	5.63(6.27)°	14.2(11.8)°	12.7(11.9)°	11.9(12.8)°	10.2(12.0)°	15.6(17.4)°	9.87(7.59)°	2.67(3.60)°

Table 7.5: Medians and interquartile ranges of the absolute error in azimuth localization for each head-tracking condition and target azimuth, formatted as median(interquartile range).

In the case of the **Real Sources**, a significant effect of the target azimuth on the localization performance was found [$F(7, 203) = 14.41$; $p < 0.001$]. The p -values of the *post hoc* Tukey's HSD analysis leads to too many combinations to be included here, but the complete results are reported in appendix Fig. A.9a. A clear tendency can be observed as larger errors occur for more lateral sources compared to the more frontal sources.

In the case of the **Lat 0.4** condition, no significant effect of the target azimuth was found [$F(7, 203) = 0.172$; $p = 0.991$]. No significant effect of the target azimuth was found either with the **Reference** head-tracking condition [$F(7, 203) = 0.829$; $p = 0.564$].

In the case of the **Amp. Artefact** condition a significant effect of the target azimuth was found [$F(7, 203) = 19.11$; $p < 0.001$]. The *post hoc* Tukey's HSD analysis (see appendix Fig. A.9b) indicates that the performance for the very lateral angles, -78° and +84° was significantly better compared to the other more frontal azimuths. It can be hypothesized that this effect might be due to listeners being instructed that sources should only be coming from frontal positions (i.e. from azimuths between $\pm 90^\circ$), which might yield this reduction of the error for the two more lateral angles which are close to the edge (boundary effect).

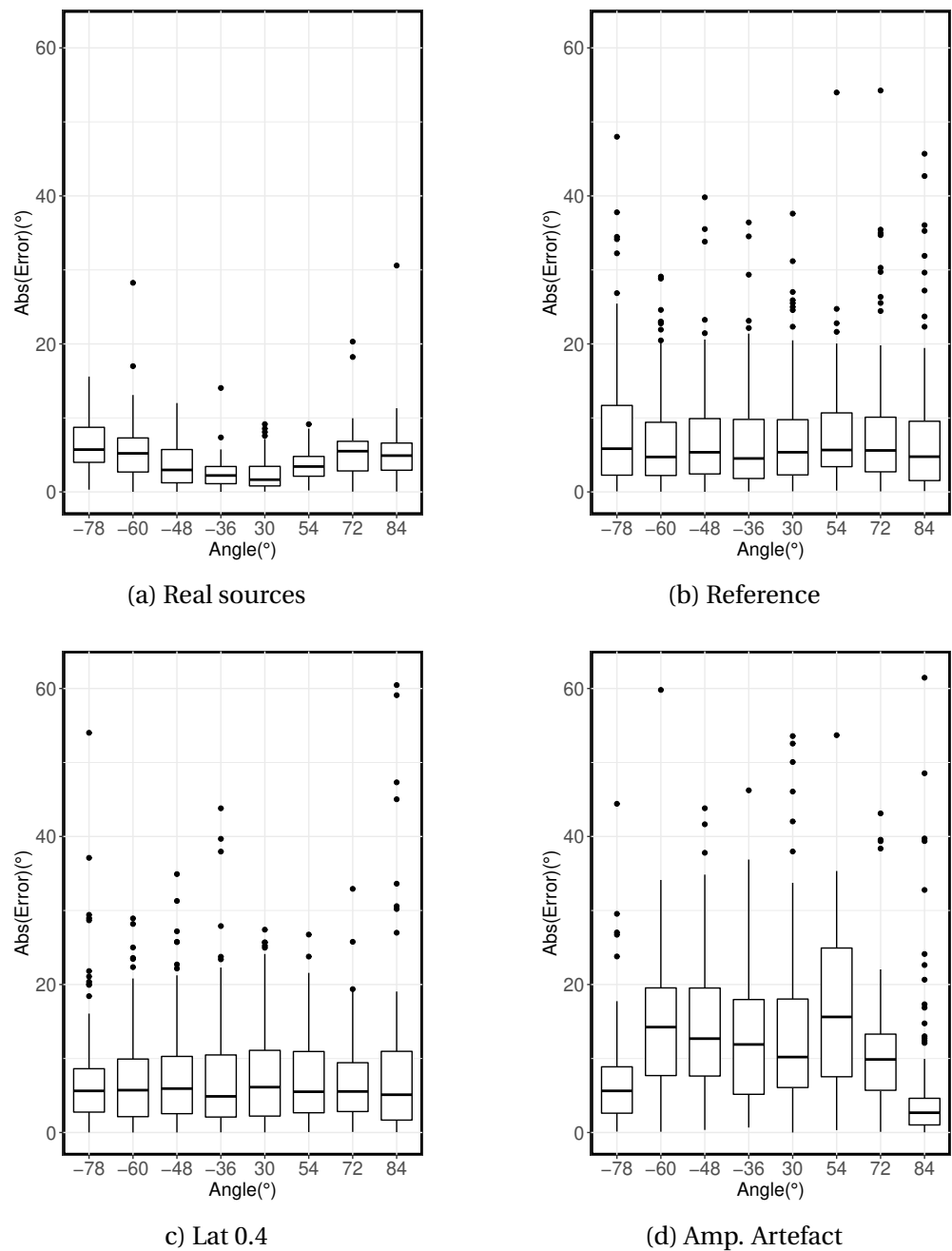


Figure 7.9: Absolute error in localization, effect of the target azimuth for each head-tracking condition.

7.8.2 Localization time

The raw localization time data distribution is neatly skewed. The data were transformed using a logarithmic transformation in order to obtain a normal distribution of the data, which was checked visually with QQ-plots of the residuals for each head tracking condition and target azimuth (see appendix Fig. A.8 and the corresponding histograms in Fig. A.7 for the head-tracking condition). The assumption for sphericity was checked using Mauchly's test, which showed that sphericity had not been violated for any of the independent variables.

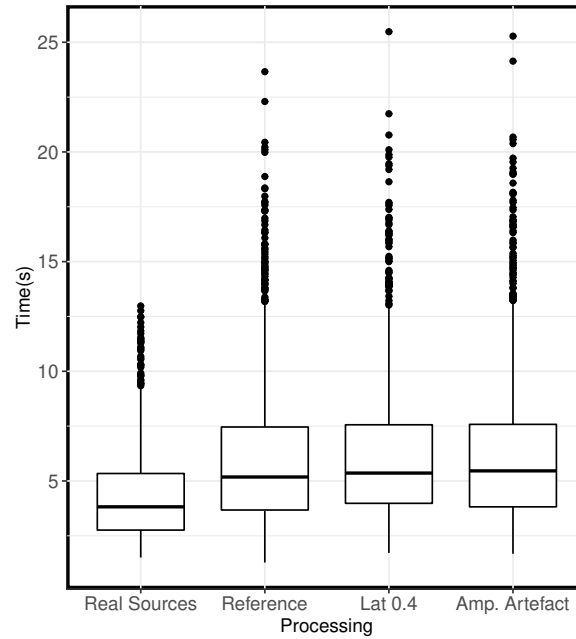


Figure 7.10: Scores for localization time per head-tracking condition.

The medians and interquartile ranges are reported for each head-tracking condition in Tab. 7.6.

	Median	Interquartile Range
Real Sources	3.82 s	2.58 s
Reference	5.18 s	3.78 s
Lat 0.4	5.36 s	3.58 s
Amp. Artefact	5.46 s	3.76 s

Table 7.6: Medians and interquartile ranges of the time of localization for each head-tracking condition.

A repeated-measures ANOVA was conducted on the log-transformed time of answer, in the virtual sources sub-group. A significant effect was found for the head-tracking condition [$F(2, 58) = 6.01$; $p = 0.0043$], with a large effect size ($\eta^2_{\text{partial}} = 0.53$). The **Reference** condition was located faster than both the **Lat 0.4** ($p = 0.0053$) and **Amp. Artefact** ($p = 0.027$) conditions.

No difference was found between the **Lat 0.4** and **Amp. Artefact** conditions ($p = 0.82$). Nevertheless, the general difference regarding the median is very modest when looking at Tab. 7.6. This suggests that neither the latency nor the amplitude mismatch did affect extensively the time it took for the listeners to confidently locate the sound source. In the case of the **Lat 0.4**, the slightly longer time could be simply explained by the 400 ms delay on the sound source.

No significant effect was found either regarding the target azimuth [$F(7, 203) = 2.28$; $p = 0.1967$]. This suggests that the difficulty was not increased for more lateral sources compared to frontal sources.

Finally, a significant difference was found for the repetition number [$F(3, 87) = 7.70$; $p < 0.001$]. This suggests that the listeners performed the task slightly more quickly over time during the session. Nevertheless a small effect size was associated with this observation ($\eta^2_{\text{partial}} = 0.011$). A *post hoc* Tukey's HSD analysis with a 95 percent confidence level was conducted and a significant effect was found between the first repetition and the three consecutive ones ($p = 0.033$ compared to repetition number 2, $p = 0.001$ compared to repetition number 3 and $p < 0.001$ for the 4th one). Those results are pictured in Fig. 7.11. The complete ANOVA table is available in appendix in Tab. A.4.

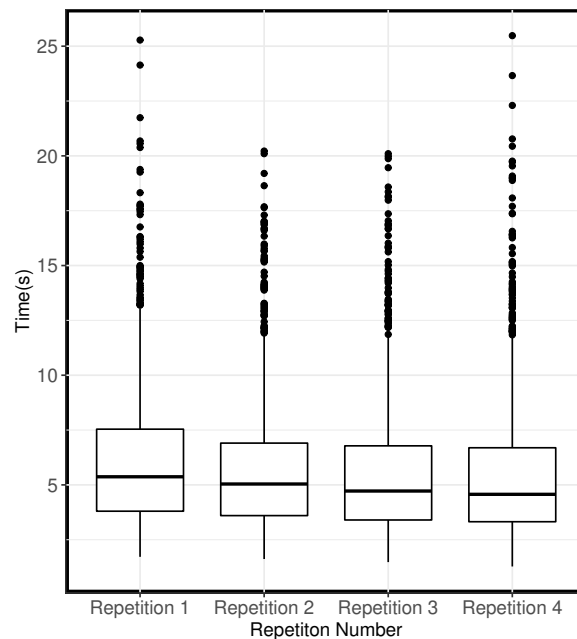


Figure 7.11: Localization time as a function of the repetition number.

7.9 Localization experiment: discussion

7.9.1 Localization error

The head-tracking latency did not decrease the performance in localization compared to a reference head-tracking in this experiment. It can be hypothesized that the subjects understood the nature of this type of artefact and were able to adapt to it. In [164], Wenzel investigated the effect of latency on various aspects of localization. The author used stimuli of 8 s to provide listeners with enough time for exploratory head movements. The study showed that localization was still accurate, even with a latency as large as 500 ms. In the study conducted in [165], which used individualized HRTFs, accuracy was generally comparable for the shortest and longest latencies. The authors suggest that listeners might have been able to ignore latency while they were actively trying to locate the source, despite the neat spatial “slewing” of the source for large latencies. In the same study, it was found that the duration of the sound stimulus affected moderately the localization performance as a function of latency. For the 8 s stimuli, front-back confusions were quite rare and increased moderately with larger latencies, whereas for the 3 s stimuli, a more pronounced increase in reversals was observed. The present study only used one setting of a large latency (400 ms) and a long speech stimulus. The rare occurrence of reversals with non-individualized HRTFs, as mentioned by the post experiment participant feed-backs, tends to confirm this observation. It is suggested in the literature that lower latency values might decrease the localization in the case of short stimuli. For example, this was the case for 1.5 to 2.5 s long stimuli and 96 ms latency in [143].

Several studies have shown that the performance in localization is poorer for lateral sources compared to frontal sources. For example results reported in [15] from two large scale studies (Preibisch-Effenberger 1966 and Haustein and Schirmer 1970), show that for a fixed head condition and a 100 ms white noise pulse, the localization uncertainty was $\pm 3.6^\circ$ for frontal sources (azimuth 0°), $\pm 5.5^\circ$ for rear sources (azimuth 180°), and around $\pm 10^\circ$ for lateral sources (azimuth 90° and 270°). The results in the present study suggest that the resort to a reference head-tracking or even a delayed head-tracking combined with head-movement yield a balanced performance for all azimuths. Indeed, by pointing the head toward the sound source, the source becomes frontal for the listener, which should thus compensate the larger error usually reported for more lateral azimuths. Nevertheless, for the real sources, an increase of error for lateral sources can still be observed. However, errors are slightly smaller for lateral sources in this study compared to the results reported in [15], suggesting that the listeners still benefited from pointing toward the source in this case.

In this experiment, the median absolute errors in localization were kept rather low for the reference and latency head-tracking conditions. In those cases, the accuracy of the listeners was comparable to the performance of untrained listeners with non-individualized HRTFs found in [118] in the case of rather frontal sources. Indeed, in this experiment head-tracking was available, and the listener had to point the head toward the sound source. Hence, any location may become frontal for the listener. This is nevertheless a different task than locating

a source presented at 0° for a listener standing still. This may explain that the errors in this test are slightly larger than the errors for the exact 0° azimuth in [118]. The absolute error was smaller for those two conditions compared to the average error reported with non-individual HRTFs in [13]. This might be explained by the fact the reported average error was mixed between azimuths in the latter. As expected, the listeners were less accurate in the case of the amplitude mismatch artefact. This is explained by the perceived angular shift of the sound source due to the tracking estimation mismatches.

It is possible that even a degraded head-tracking with amplitude mismatch or latency still provides differential integration of the binaural cues which are enough to resolve most of the front-back confusions. In addition, the experimenter also gave the information to the listener that the source would not be coming from rear positions, which might have helped the subjects to not experience reversals. Oral feedback from the listeners indicated that the subjects very rarely perceived the sounds coming from rear positions.

In [114], three listeners (among which two are co-authors) performed virtual and free-field localization for different sound-source locations. Front-back confusion rate and average localization error were equivalent between individualized virtual localization and free-field localization for each participant. Nevertheless, in the present study, the binaural synthesis used a combination of non-individualized HRTFs for the direct sound and non-individualized 10-ms truncated BRIRs for the ERs (in which the direct sound was set to zeros). In [168], the results suggest that most listeners can locate sound sources accurately without individualized HRTFs, particularly in the azimuth dimension. Begault [13] found that head-tracking can reduce localization errors in azimuth, for non-individual HRTFs.

7.9.2 Localization time

The results of the localization time suggest that the head-tracking artefacts did not affect the time it took for the listeners to locate the sound source, as there was very little difference between the head-tracking conditions with artefacts compared to the reference condition. The small increase in localization time for the 400 ms latency artefact is in agreement with the results in [26], which found that for continuous sound stimuli, the response time was larger for latencies above about 90 ms.

7.9.3 Considerations for applications to binaural communication devices

The results of this experiment suggest that the impact of a large latency on both auditory externalization and localization performance is small. Nevertheless, latencies might create some other disturbance on the long term. In the context of virtual reality, the so-called motion sickness, mostly thought to be visually induced, was shown to be potentially triggered by auditory cues as well [86]. Vection, which is the illusion of self-motion in the absence of real physical movement is another well-known artefact in this context. However the influence of

the visual cues might be larger than auditory cues for this type of issue [85].

The errors obtained with the **Amp. Artefact** condition, even if larger than for the **Lat 0.4** and **Reference** conditions, were not excessively large when looking at average values in this experiment. Nevertheless, one should bear in mind that the head-tracking time was rather short, and that errors due to the integration with this type of artefact might accumulate with longer time of measure. This means that such algorithm should necessarily include a method to re-initialize the position with a reference cue in order to limit the accumulation and maintain errors in more acceptable ranges. This can be done with magnetometers and gyroscopes in devices for which technical constraints allow it. In the context of HAs where the goal would be to use only accelerometers with algorithms as described in Chapter 6, one could imagine to resort to direction of arrival (DOA) algorithms, using a sporadically active sound source which position would be known in the room.

In this study, the listeners were not given time to train with the generic HRTFs. At best they listened with those HRTFs in the first session and a minimum one week gap was respected between the two session for every participant. No visual feedback was given neither for them to learn. It is likely that a training could help improve the localization of sound source in application. Indeed, with a wearable binaural communication device, the listener would constantly be stimulated by both audio and the corresponding visual feedback. This may help them learn quickly, and thus to perform better in localization [118].

7.10 Summary and conclusion

In this section, two subjective listening tests were performed to evaluate the effect of various head-tracking conditions on the perception of auditory externalization and the performance in localization. The binaural synthesis was achieved with 10 ms non-individualized BRIRs superimposed to a direct sound filtered with non-individualized HRTFs.

For each DRR setting, the condition with no head-tracking always resulted in poorer externalization ratings compared to all the conditions with head-tracking. The results suggest that amplitude mismatch of the head-tracking as well as a large latency (400 ms), might not affect auditory externalization. The experiment did not include a realistic visual cue, hence the potential perceived "slewing" of the sound source might not have disturbed the participants in their perception of externalization. This suggest that they could still benefit from the additional cue provided by the head motion to externalize the sound source, even when the head-tracking was substantially degraded. Further studies should investigate if this observation holds with a realistic visual reference, which might affect how the "slewing" is interpreted by the auditory system.

Large head-tracking latency did not affect the performance in localization compared to the reference tracking, suggesting that the listeners might have understood spontaneously the nature of the artefact in this case and adapted their strategy. It is also likely that even a

degraded head-tracking, with either latency or amplitude mismatch, is enough to provide differential integration of the binaural cues helping to resolve most front-back confusions. The estimation mismatch, which simulated the artefact of the algorithm described in Chapter 6, naturally led to larger errors in localization. This is explained by the head-tracking errors which result in a perceptual shift of the azimuth of the sound source.

8 Conclusion

This thesis addressed the improvement of the spatial rendering performed in binaural communication devices, such as HAs and hearables, used with a RM system. In particular, the perception of auditory distance and the perception of externalization were targeted for this purpose. Indeed, previous works in the field developed a localization and spatialization algorithm allowing localization in azimuth to HI listeners while preserving speech intelligibility. However, a lack of realism, due in particular to in-head localization, was experienced by listeners, and thus reported as a limitation of the developed algorithm. This motivated the investigation of technical solutions aiming at enhancing externalization and the perception of auditory distance of a RM signal. The design of these features was tailored to be compatible with the technical constraints of wearable devices. First, the investigation of DSP methods that could enable the introduction of ERs in the spatial rendering of such wearable devices was addressed. Additionally, an accelerometer-based head-tracking algorithm compatible with binaural devices was proposed. These investigations were motivated by the potential improvement in externalization and auditory distance perception brought by ERs and head-tracking in the perception of spatial hearing. Several subjective studies aimed at evaluating if the proposed methods did actually help listeners to perceive auditory distance and externalize sound sources. The main contributions and results are summarized below. The associated perspectives^I are also reported in each section.

Contributions and perspectives

Adding ERs in the RM signal

Two different types of strategies were explored with the purpose of superimposing ERs to a direct sound spatialized with generic HRTFs. The first consists in removing the noise and late reflections from the signal picked up by the binaural devices' microphones to extract the direct sound and ERs of the speech. In particular, the developed methods take advantage of the "clean" speech signal picked up by the RM to process the binaural devices' microphones signal.

^IThe perspectives are indicated with a ★ symbol.

- The MWF, a state-of-the-art method was first implemented and investigated. Although showing efficient performance for the intended purpose, MWF is too heavy computationally and might not be compatible with the processing power of most wearable devices.
 - An alternative method based on the coherence between the signals picked up by the RM and the wearable devices' microphone was investigated. Despite yielding lower performance compared to the MWF, the method still achieves the intended purpose, but at a lower computational cost.
 - Another original method based on the use of the frequency-dependent envelopes of the RM to directly filter the wearable devices' microphone signals was proposed and implemented. While showing promising performance in informal listening sessions, the method currently requires further investigations and comparison with the other methods.
 - The second type of strategy that was considered synthesized artificial ERs from the RM signal. Partitioned convolution was implemented and investigated for this purpose. The method offered several advantages such as an accurate definition of the ER time and the possibility to provide clean ERs regardless of the amount of surrounding noise.
- ★ The work in this thesis mainly focused on the effect of those strategies on spatial perception, but did not assess other perceptual aspects. Subjective listening studies could be conducted in further works to assess speech quality and an overall preference of the denoised signals with those methods.

Effect of ERs on auditory distance perception in NH and HI listeners

Two studies were conducted to assess if the methods proposed to provide ERs in the spatialized RM actually improve the perception of auditory distance.

First study

The first study included NH and HI aided listeners with moderate-to-highly-profound HL. The listeners had to evaluate the perceived auditory distance of various binaural rendering strategies while the visual cues were available.

- The results showed that the addition of ERs with the coherence-based method did improve the perception of auditory distance in both NH and HI listeners compared to the baseline algorithm using non-individualized HRTFs. In particular, it helped increasing substantially the rate of externalization.
- It was also shown that the non-linear amplification (WDRC) performed in most HAs did not prevent HI listeners from perceiving differences in auditory distance between the stimuli.
- The auditory distance evaluation results additionally suggest that the performance of HI listeners is more similar to that of NH listeners when the WDRC is performed after the spatialization, i.e. when the audibility is optimized rather than the accuracy of the binaural

cues. This could be explained by the long term accustomization of the HI listeners to their own compression setting.

- This study also showed that severe-to-profound HI listeners have a contracted perception of auditory distance compared to NH listeners. This confirms previously published results in the literature, which observed the same phenomenon in mild-to-moderate HI listeners.

- ★ Due to technical constraints, the HI listeners did not use their HAs and the sound was rendered through headphones with a simulation of their amplification settings instead. Whether the difference in acoustic coupling between those two scenarios affects the perception of auditory distance could be addressed in future studies. Further work could also assess to what extent the non-availability of visual cues might affect the results found in this study. Various azimuths of the sound source could be additionally evaluated.

Second study

A follow-up study with a similar protocol and purpose was conducted with NH listeners. The experiment additionally included the partitioned convolution-based algorithm. The baseline algorithm was tested in this experiment with an additional signal picked up from the hearables' microphones as commonly done in current RM use cases.

- The partitioned convolution method was shown to yield similar auditory distance perception to the coherence-based one.

- The baseline algorithm method enables only including ERs with a lower amplitude, and was therefore perceived closer than the two other methods.

- This study also aimed to verify that the two methods that provide additional ERs do not affect the ability of the listener to detect a surrounding non-visible acoustic event in a challenging auditory situation (low SNR). With the proposed test, no significant difference was found between the three considered strategies.

- The combined results of the two experiments suggest that the two proposed algorithms providing additional ERs might improve auditory distance perception without affecting the ability of listeners to detect other surrounding events.

- ★ This study was restricted to NH listeners. Future works could include HI listeners, and focus on the investigation of the influence of the two proposed algorithms on speech intelligibility and spatial awareness. Indeed, those two tasks are more challenging for HI listeners compared to NH listeners, thus it could be expected that the results might differ. This is of particular interest as speech intelligibility remains the most important goal of HAs, and spatial awareness is crucial in the RM application to ensure safety. Future works could also investigate the tuning of the algorithm to determine the settings yielding the optimal trade-off between auditory distance perception, spatial awareness and speech intelligibility, as a function of the SNR and the degree of impairment of the listener.

Development of a head-tracking algorithm using two accelerometers

The next contribution consisted in the proposition of a head-tracking algorithm compatible with the technical constraints of wearable devices.

- The developed algorithm was designed for the purpose to estimating the azimuth orientation of the head, i.e. the yaw, using solely two accelerometers (one per device). The yaw is particularly challenging to estimate with accelerometers, as accelerations around the gravity axis cannot be detected by such sensors.
 - The proposed method includes several original updates and substantial improvements to a method of the literature adapted to simpler cases. It achieves yaw estimation of a human head movements without knowing the initial orientation of the sensors placed over the ears.
 - The head-tracking strategy was evaluated with realistic measurements of human motions, in a task combining sound localization and writing. The results show that a good performance can be achieved on a majority of the measurements.
 - A factorial analysis was performed, and showed that with the current set of tunable parameters, it is not possible to find a unique combination yielding a good performance on all the measurements. One major limitation of the current version of the algorithm is that it is not robust against sensor displacement after calibration. A second limitation, is that the algorithm is not capable of estimating slow movements, as the inherent sensor noise does not allow to use accelerations which are too low in amplitude.
- ★ As no unique set of parameters could be found to achieve a good performance on all the recorded measurements, it can be suggested that some of these parameters might have to be defined dynamically, based on individual-dependent characteristics of the retrieved accelerations. This could be addressed in future works. Additionally, the algorithm relies on two integration stages. Thus, small errors might accumulate on long term estimations and lead to larger shifts. Hence, further works could investigate the possibility to include an absolute reference to compensate for the occurrence of such errors. For example, in the RM use-case in a classroom, it is possible to imagine resorting to a DOA estimation algorithm with a sound source whose position is fixed and known in the room.

Effects of head-tracking artefacts on auditory spatial perception

Finally, a subjective listening test was conducted to assess the effect of two types of head-tracking artefacts on auditory spatial perception. The first artefact was a simulated estimation mismatch, mimicking a limiting behaviour of the algorithm developed in this thesis. The second artefact consisted of a large latency of head-tracking (400 ms). The spatialization of a speech sample was performed with non-individual HRTFs to which were superimposed 10 ms of ERs with partitioned convolution and generic BRIRs of the listening room where the test took place.

- The first experiment assessed the effect of those artefacts on auditory externalization of a frontal source, while the listener was performing a large head movement. The results showed that head-tracking coupled with head movements yielded a higher degree of externalization compared head movements with no head-tracking, confirming the results of previous studies.
 - In addition, this externalization improvement remained valid when the head-tracking included the tested artefacts, suggesting that the listener could still take advantage of spatial cues provided by the head movement when the tracking was degraded.
 - The second experiment consisted of a localization task in azimuth with the same head-tracking artefacts. The results showed that a large latency did not affect the ability of the listeners to locate virtual sound sources compared to a reference head-tracking situation. It can be hypothesized that the listener understood the nature of the artefact in this case, and managed to compensate for it.
 - As expected, the estimation mismatch artefact decreased the localization performance in azimuth. This is explained by the resulting perceived angular shift of the sound source.
- ★ In the externalization experiment, the direction of the virtual sound sources was only materialized by a visual mark on the wall. It is possible that even if the listeners were instructed that the sound should be coming from this direction and be static, the auditory system might have interpreted the perceived "slewing" as a movement of the sound source. Further studies could include a more realistic visual reference, e.g. a video of a speaker on a screen, to assess whether the mismatch between the visual cue and the degraded dynamic auditory information would make externalization collapse.

Final word

The research undertaken in this thesis has lead to promising contributions in the field of spatial hearing. In particular, it can be expected from the achieved developments and results, that future RM systems could provide a more realistic and natural auditory experience to most listeners. This is of high interest for aided HI people, for whom the RM system allows participating in many daily life activities such as attending a class or working and communicating with others.

A Additional figures and data

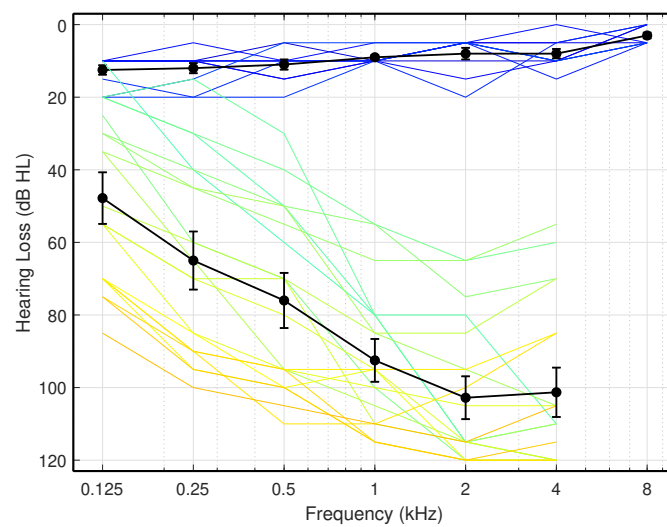


Figure A.1: Audiograms, measured at the best ear, of the NH listeners (blue) and HI listeners (green, yellow, orange). The upper and lower thicker black lines corresponds to the average of the NH listeners and HI listeners respectively. (Chapter 4).

Degree of hearing loss	Hearing loss range (dB HL)
Normal	25 dB
Slight	26–40 dB
Moderate	41–60 dB
Severe	61–80 dB
Profound	81 dB or greater

Table A.1: WHO's grades of hearing impairment [132].

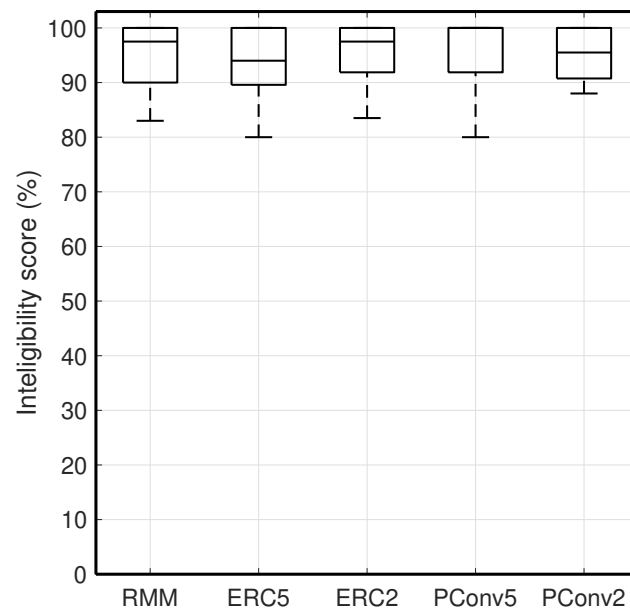


Figure A.2: Intelligibility scores in the spatial awareness experiment (Chapter 5).

Effect	<i>df</i>	<i>df_{residual}</i>	<i>F</i>	<i>p</i>
HT Condition	2	112.17	137.26	< 0.001
DRR	1.31	73.55	26.756	< 0.001
Num. of Repetitions	2.82	157.66	2.835	0.066
Filter	1	56	0.929	0.339
Filter x HT Condition	2	112.17	0.850	0.430
Filter x DRR	1.31	73.55	0.443	0.5610
Filter x Num. of Repetitions	2.82	157.66	1.993	0.121
HT Condition x DRR	5.13	287.06	0.206	0.962
HT Condition x Num. of Repetitions	7.29	408.09	0.978	0.449
DRR x Num. of Repetitions	5.36	300.01	0.211	0.525

Table A.2: ANOVA Table with Greenhouse-Geisser correction : Z-score of externalization (Chapter 7).

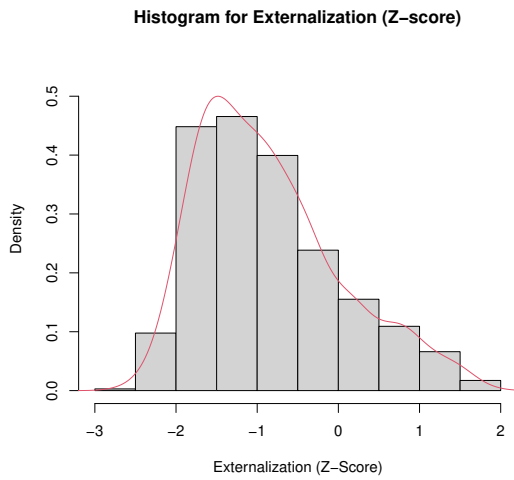
Additional figures and data

Effect	df	$df_{residual}$	F	p
HT Condition	2	58	27.25	< 0.001
Angle	7	203	3.65	< 0.001
Num. of Repetitions	3	87	0.104	0.957
HT Condition x Angle	14	406	13.058	< 0.001
HT Condition x Num. of Repetitions	6	174	0.253	0.958
Angle x Num. of Repetitions	21	609	1.335	0.145

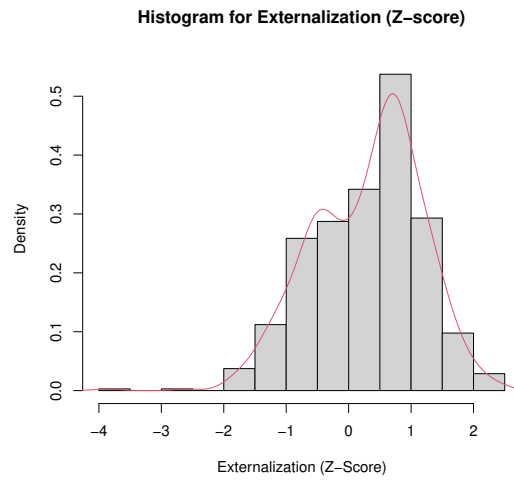
Table A.3: ANOVA Table : Error of localization (Chapter 7).

Effect	df	$df_{residual}$	F	p
HT Condition	2	58	6.008	0.004
Angle	7	203	2.275	0.030
Num. of Repetitions	3	87	7.703	0.001
HT Condition x Angle	14	406	1.167	0.298
HT Condition x Num. of Repetitions	6	174	1.770	0.108
Angle x Num. of Repetitions	21	609	1.532	0.061

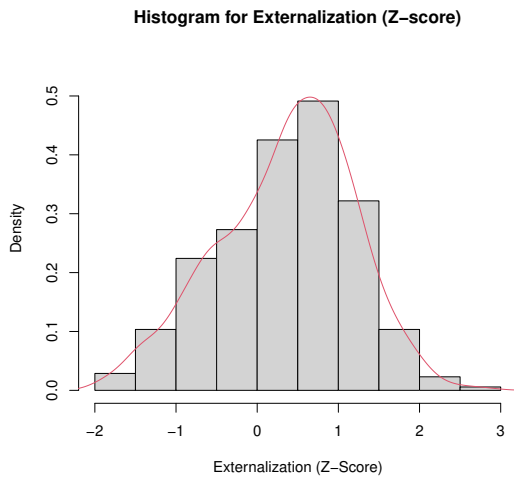
Table A.4: ANOVA Table : Time of localization (Chapter 7).



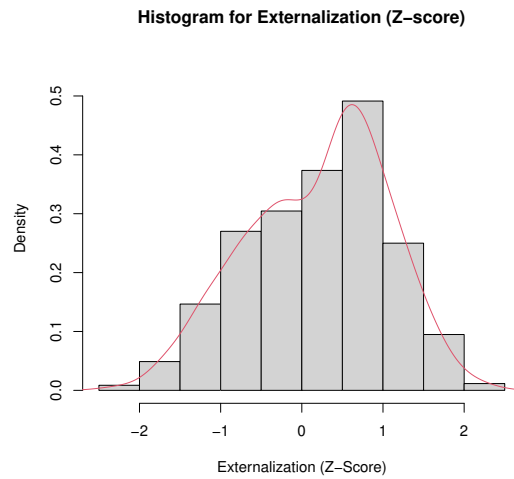
(a) No HT



(b) Reference



(c) Lat 0.4



(d) Amp. Artefact

Figure A.3: Histogram of the z-score externalization rating per head-tracking condition (Chapter 7).

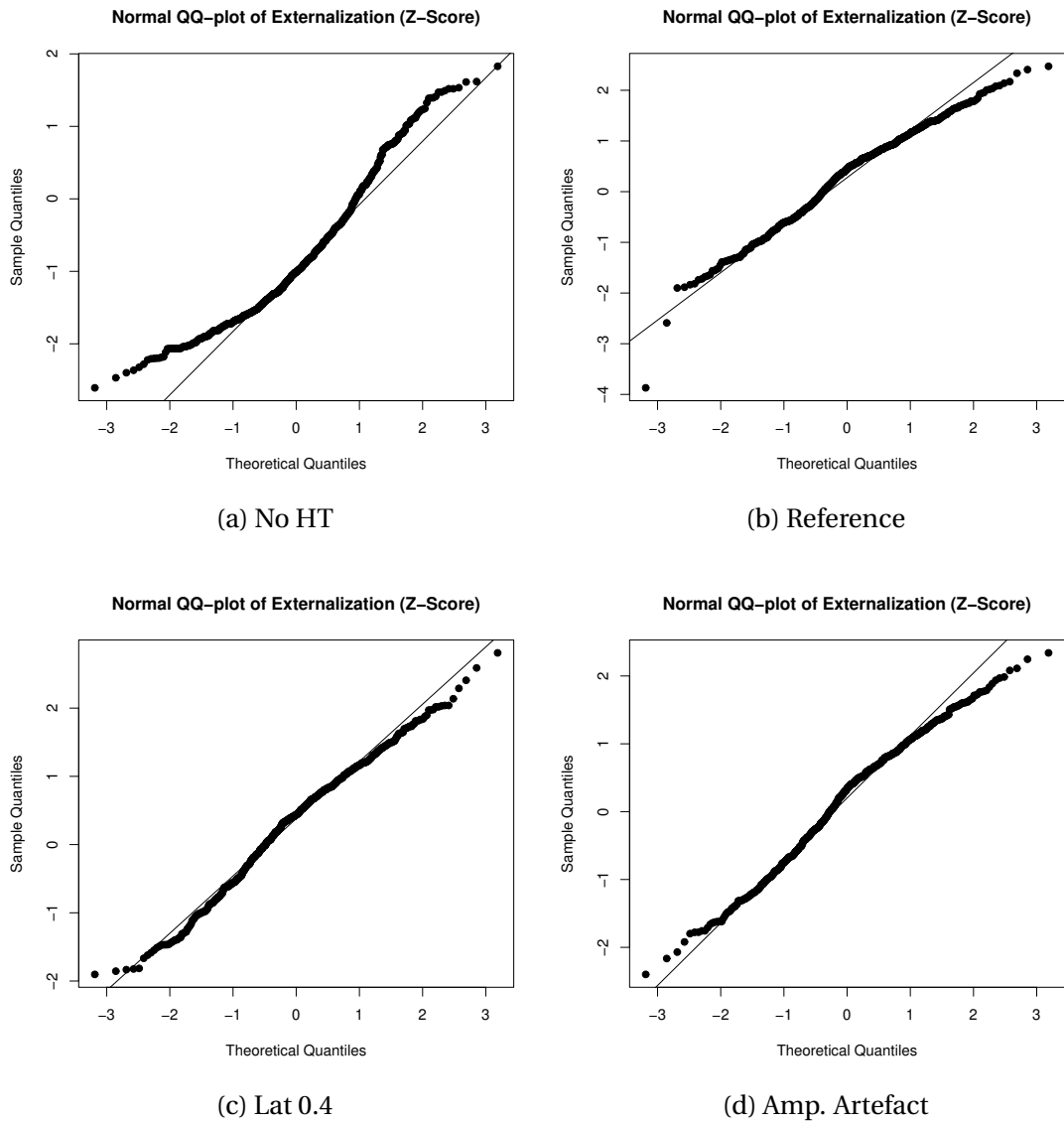
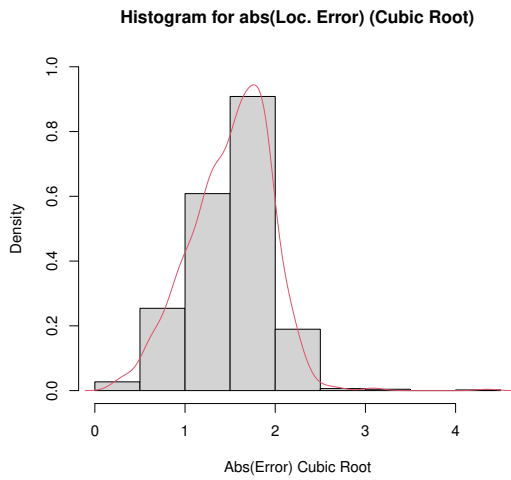
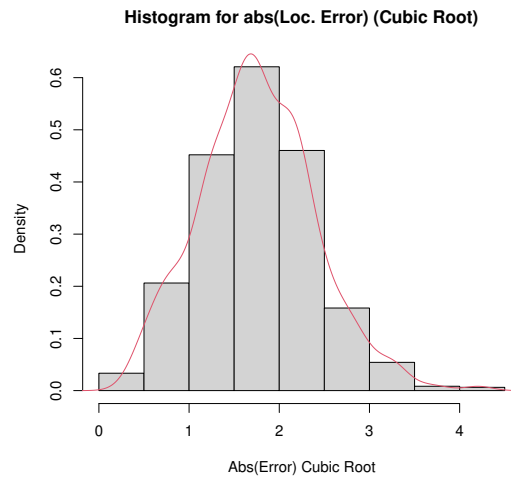


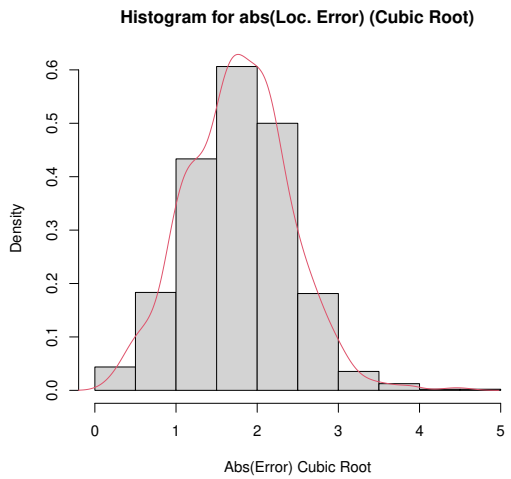
Figure A.4: QQ-Plots of the residuals per head-tracking condition for the externalization ratings (z-scores) (Chapter 7).



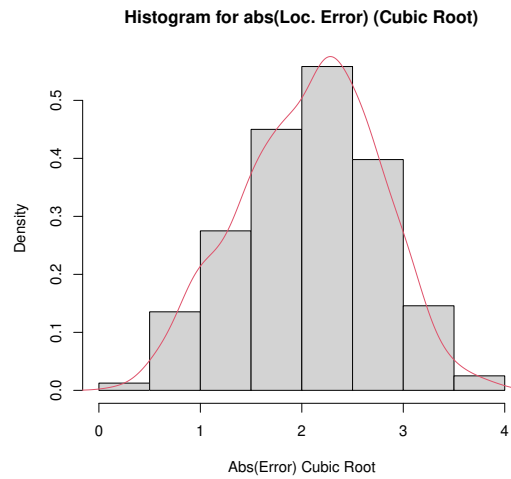
(a) Real sources



(b) Reference



(c) Lat 0.4



(d) Amp. Artefact

Figure A.5: Histogram of the absolute error of localization (cubic root transformed) per head-tracking condition (Chapter 7).

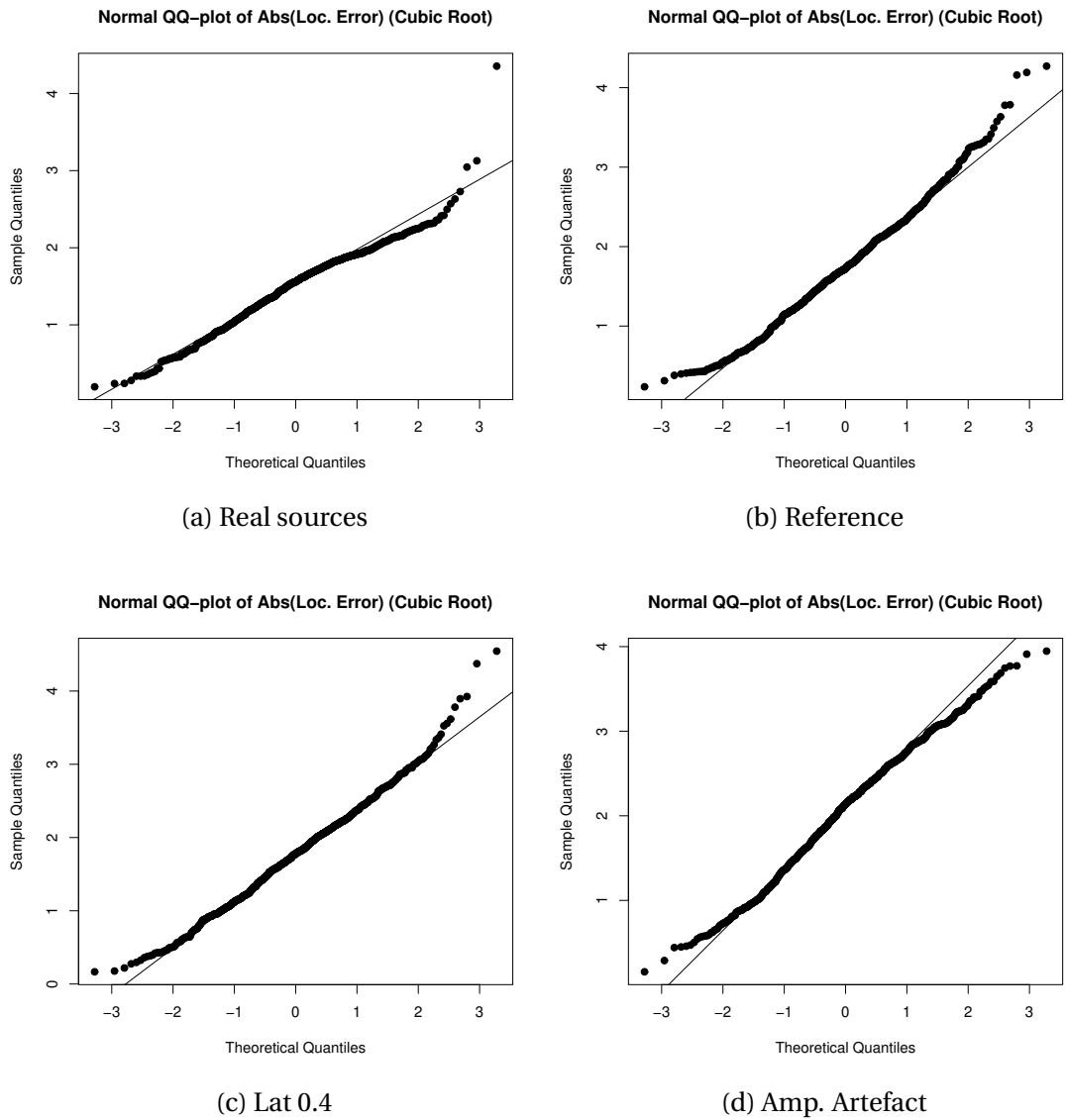


Figure A.6: QQ-Plots of the residuals per head-tracking condition for the absolute error of localization (cubic root transformed) (Chapter 7).

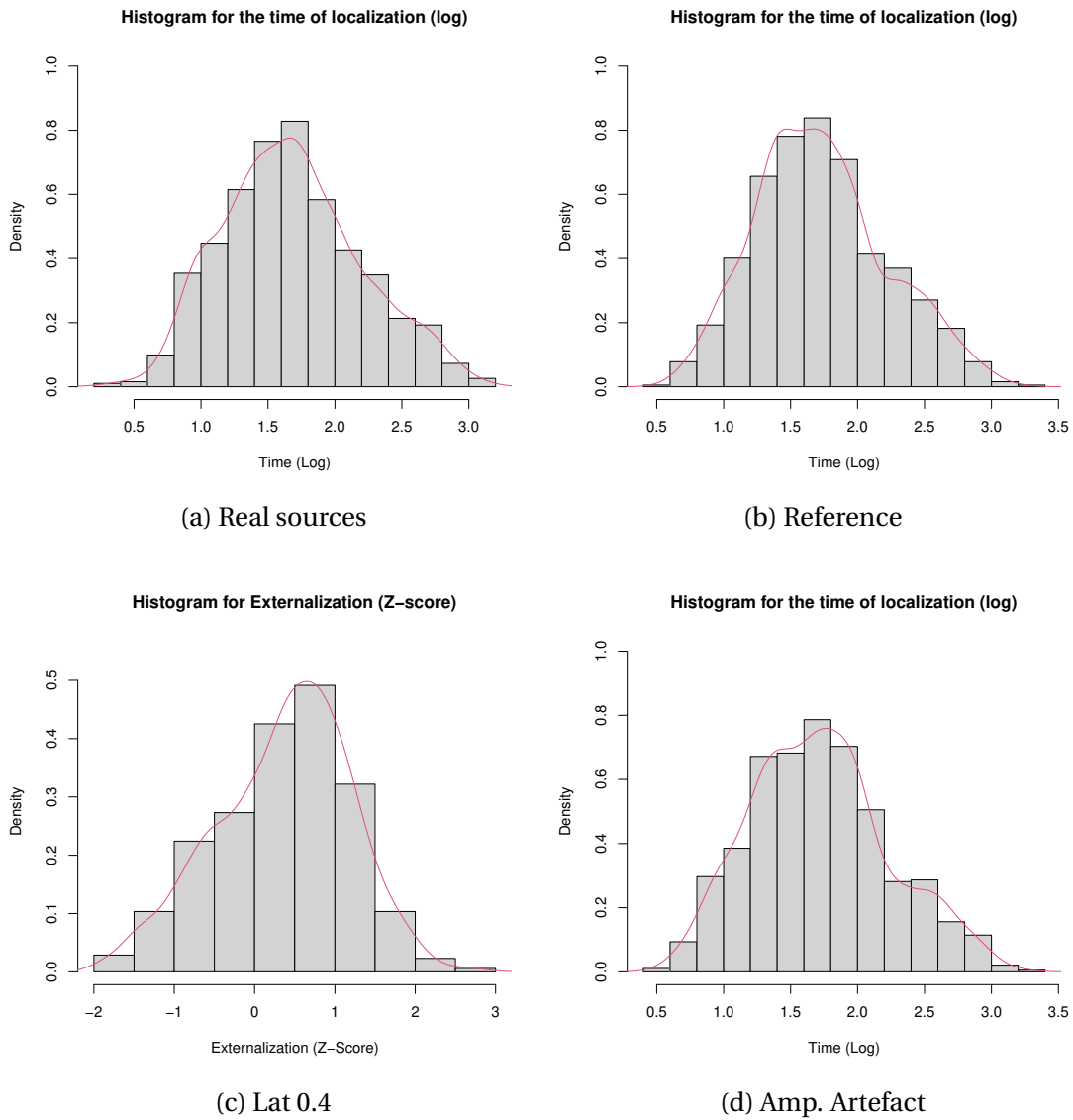


Figure A.7: Histogram of the absolute error of localization (cubic root transformed) per head-tracking condition (Chapter 7).

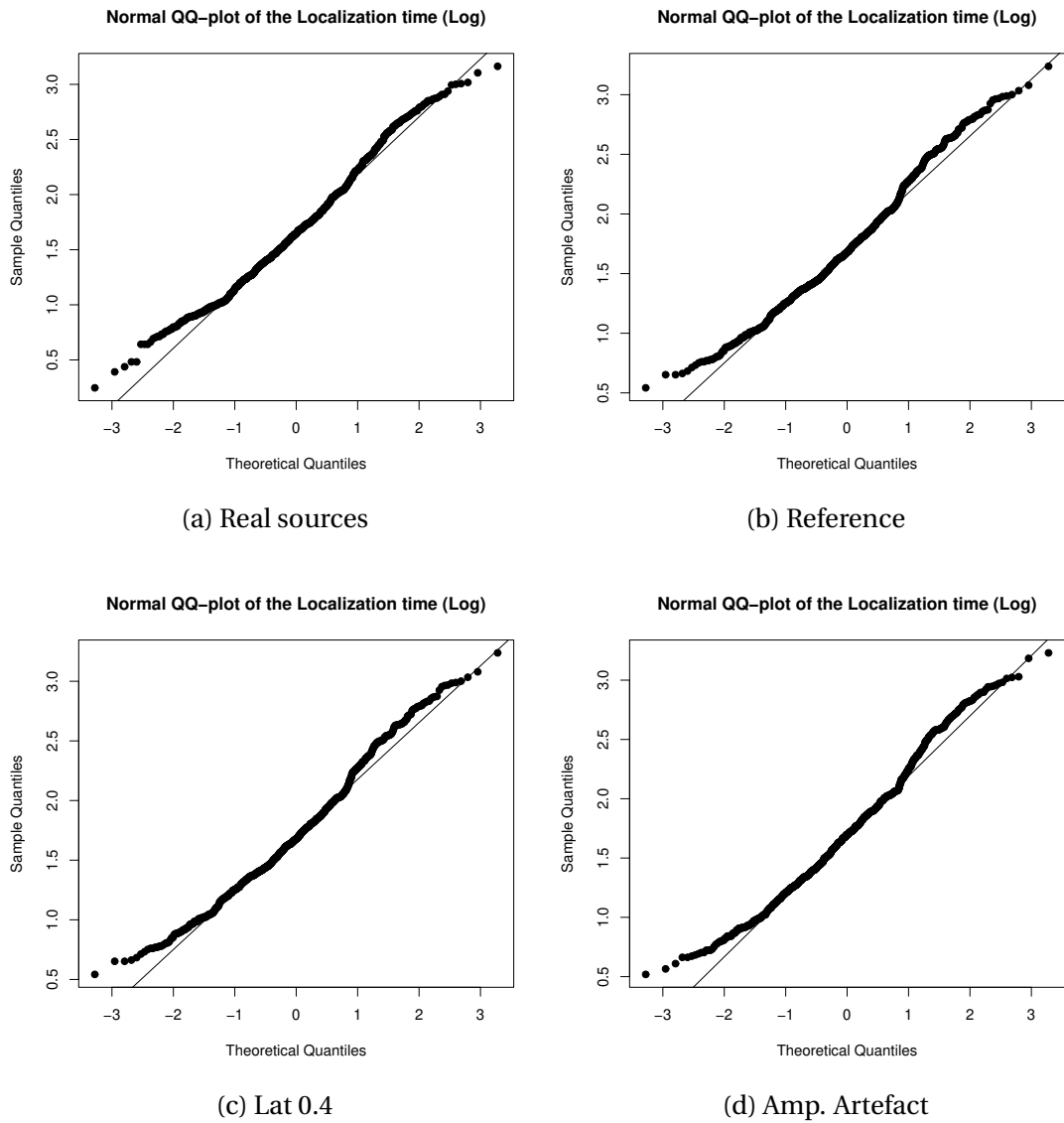


Figure A.8: QQ-Plots of the residuals per head-tracking condition for the absolute error of localization (cubic root transformed) (Chapter 7).

contrast	estimate	SE	df	t.ratio	p.value
(-78) - (-60)	0.14368	0.0776	203	1.851	0.5859
(-78) - (-48)	0.36722	0.0776	203	4.732	0.0001
(-78) - (-36)	0.52553	0.0776	203	6.771	<.0001
(-78) - 30	0.56537	0.0776	203	7.285	<.0001
(-78) - 54	0.34063	0.0776	203	4.389	0.0005
(-78) - 72	0.11640	0.0776	203	1.500	0.8069
(-78) - 84	0.11357	0.0776	203	1.463	0.8258
(-60) - (-48)	0.22354	0.0776	203	2.880	0.0820
(-60) - (-36)	0.38185	0.0776	203	4.920	<.0001
(-60) - 30	0.42170	0.0776	203	5.434	<.0001
(-60) - 54	0.19695	0.0776	203	2.538	0.1858
(-60) - 72	-0.02728	0.0776	203	-0.352	1.0000
(-60) - 84	-0.03011	0.0776	203	-0.388	0.9999
(-48) - (-36)	0.15831	0.0776	203	2.040	0.4579
(-48) - 30	0.19815	0.0776	203	2.553	0.1797
(-48) - 54	-0.02659	0.0776	203	-0.343	1.0000
(-48) - 72	-0.25082	0.0776	203	-3.232	0.0305
(-48) - 84	-0.25365	0.0776	203	-3.268	0.0273
(-36) - 30	0.03985	0.0776	203	0.513	0.9996
(-36) - 54	-0.18490	0.0776	203	-2.382	0.2557
(-36) - 72	-0.40913	0.0776	203	-5.272	<.0001
(-36) - 84	-0.41195	0.0776	203	-5.308	<.0001
30 - 54	-0.22474	0.0776	203	-2.896	0.0788
30 - 72	-0.44898	0.0776	203	-5.785	<.0001
30 - 84	-0.45180	0.0776	203	-5.821	<.0001
54 - 72	-0.22423	0.0776	203	-2.889	0.0802
54 - 84	-0.22706	0.0776	203	-2.926	0.0728
72 - 84	-0.00282	0.0776	203	-0.036	1.0000

(a) Real sources

contrast	estimate	SE	df	t.ratio	p.value
(-78) - (-60)	-0.5880	0.106	203	-5.552	<.0001
(-78) - (-48)	-0.5636	0.106	203	-5.322	<.0001
(-78) - (-36)	-0.4720	0.106	203	-4.457	0.0004
(-78) - 30	-0.4464	0.106	203	-4.215	0.0010
(-78) - 54	-0.6667	0.106	203	-6.295	<.0001
(-78) - 72	-0.3234	0.106	203	-3.053	0.0514
(-78) - 84	0.2826	0.106	203	2.668	0.1385
(-60) - (-48)	0.0244	0.106	203	0.231	1.0000
(-60) - (-36)	0.1160	0.106	203	1.095	0.9572
(-60) - 30	0.1417	0.106	203	1.338	0.8834
(-60) - 54	-0.0787	0.106	203	-0.743	0.9955
(-60) - 72	0.2647	0.106	203	2.499	0.2018
(-60) - 84	0.8706	0.106	203	8.221	<.0001
(-48) - (-36)	0.0916	0.106	203	0.865	0.9888
(-48) - 30	0.1172	0.106	203	1.107	0.9547
(-48) - 54	-0.1031	0.106	203	-0.974	0.9777
(-48) - 72	0.2402	0.106	203	2.268	0.3163
(-48) - 84	0.8462	0.106	203	7.990	<.0001
(-36) - 30	0.0257	0.106	203	0.242	1.0000
(-36) - 54	-0.1947	0.106	203	-1.838	0.5949
(-36) - 72	0.1487	0.106	203	1.404	0.8547
(-36) - 84	0.7546	0.106	203	7.125	<.0001
30 - 54	-0.2203	0.106	203	-2.080	0.4311
30 - 72	0.1230	0.106	203	1.161	0.9417
30 - 84	0.7290	0.106	203	6.883	<.0001
54 - 72	0.3434	0.106	203	3.242	0.0296
54 - 84	0.9493	0.106	203	8.963	<.0001
72 - 84	0.6060	0.106	203	5.722	<.0001

(b) Amp. Artefact

Figure A.9: Complete *Post hoc* Tukey HSD for the effect of the target azimuth on the localization absolute error in the **Real Sources** (a) and **Amp. Artefact** (b) conditions. The values in the "contrast" column correspond to the target azimuths in degrees (Chapter 7).

Bibliography

- [1] I. Adnanul. *Detecting head movement using gyroscope data collected via in-ear wearables*. University of Oulu. Master: Faculty of information technology and electrical engineering, 2021.
- [2] M.A. Akeroyd, S. Gatehouse, and J. Blaschke. “The detection of differences in the cues to distance by elderly hearing-impaired listeners”. In: *J Acoust Soc Am* 121.2 (2007), pp. 1077–89.
- [3] MA. Akeroyd. “The effect of hearing-aid compression on judgments of relative distance”. In: *J Acoust Soc Am* 127.1 (2010), pp. 9–12.
- [4] P. Anderson and P. Zahorik. “Auditory/visual distance estimation: Accuracy and variability”. In: *Frontiers in psychology* 5 (Oct. 2014), p. 1097.
- [5] E. Armelloni, C. Giottoli, and A. Farina. “Implementation of real-time partitioned convolution on a DSP board”. In: Nov. 2003, pp. 71–74. ISBN: 0-7803-7850-4.
- [6] I. Arweiler and J.M. Buchholz. “The influence of spectral characteristics of early reflections on speech intelligibility”. In: *The Journal of the Acoustical Society of America* 130.2 (2011), pp. 996–1005.
- [7] D.H. Ashmead, D.F. LeRoy, and R.D. Odom. “Perception of the relative distances if nearby sound sources”. In: *Perception & Psychophysics* 47.4 (1990), pp. 326–331.
- [8] H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel. “Comparison of different egocentric pointing methods for 3D sound localization experiments”. In: *Acta acustica united with Acustica* 102.1 (2016), pp. 107–118.
- [9] S. Banerjee and Starkey Laboratories. *The Compression Handbook: An Overview of the Characteristics and Applications of Compression Amplification*. Starkey Laboratories, 2005.
- [10] D.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.
- [11] E. Battenberg and R. Aviûienis. “Implementing Real-Time Partitioned Convolution Algorithms on Conventional Operating Systems”. In: In 14th International Conference on Digital Audio Effects (DAFx-11), Paris. 2011.

-
- [12] D.R. Begault. "Perceptual effects of synthetic reverberation on three-dimensional audio systems". In: *AES: Journal of the Audio Engineering Society* 40 (Dec. 1992).
- [13] D.R. Begault, E. M. Wenzel, and M. R. Anderson. "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: *J Audio Eng Soc.* 49.10 (2001), pp. 904–16.
- [14] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopco. "Sound Externalization: A Review of Recent Research". In: *Trends in hearing* 24 (Jan. 2020), p. 2331216520948390.
- [15] J. Blauert. *Spatial hearing – The psychophysics of human sound localization*. The MIT Press, 1999.
- [16] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Monnen. "Binaural cue preservation for hearing aids using an interaural transfer function multi-channel Wiener Filter". In: *Proc. ICASSP*. 2007, 2007.
- [17] A. Boothroyd. "Hearing aid accessories for adults: the remote FM microphone". In: *Ear Hear* 25.1 (Feb. 2004), pp. 22–33.
- [18] A.W. Boyd, M. Whitmer, J.J. Soraghan, and M.A. Akeroyd. "Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers". In: *J Acoust Soc Am* 131 (Mar. 2012), p. 3.
- [19] E. Branda and T. Wurzbacher. "Motion Sensors in Automatic Steering of Hearing Aids". In: *Semin Hear* 42 (2021), p. 03.
- [20] W.O. Brimijoin, A.W. Boyd, and M.A. Akeroyd. "The Contribution of Head Movement to the Externalization and Internalization of Sounds". In: *PLoS One* 8 (2013), p. 12.
- [21] O.B. Britta. *What is Apple Spatial Audio, how does it work and how to get it?* from, 2021.
- [22] A.W. Bronkhorst. "Effect of stimulus properties on auditory distance perception in rooms". In: *Physiological and Psychological Bases of Auditory Function* (2001), pp. 184–191.
- [23] A.W. Bronkhorst and T. Houtgast. "Auditory distance perception in rooms". In: *Nature*. 11.397 (Feb. 1999), p. 6719.
- [24] C.P. Brown and R.O. Duda. "A structural model for binaural sound synthesis". In: *IEEE T. Audio* 6 (Sept. 1998).
- [25] D.S. Brungart, N.I. Durlach, and W.M. Rabinowitz. "Auditory localization of nearby sources. II. Localization of a broadband source". In: *J Acoust Soc Am* 106.4 (Oct. 1999), pp. 1956–68.
- [26] D.S. Brungart, A.J. Kordik, and B.D. Simpson. "Effects of headtracker latency in virtual audio displays". In: *journal of the audio engineering society* 54.1/2 (Jan. 2006), pp. 32–44.
- [27] D.S. Brungart and W.M. Rabinowitz. "Auditory localization of nearby sources". In: *Head-related transfer functions* 106.3 (Sept. 1999), pp. 1465–79.

BIBLIOGRAPHY

- [28] J. R. Burwinkel, X. Buye, and J. Crukley. “Preliminary Examination of the Accuracy of a Fall Detection Device Embedded into Hearing Instruments”. In: *J Am Acad Audiol* 31 (2020), p. 06.
- [29] R.A. Butler and R.A. Humanski. “Localization of sound in the vertical plane with and without high-frequency spectral cues”. In: *Perception & psychophysics* 51.2 (1992), pp. 182–186.
- [30] R.A. Butler, E.T. Levy, and W.D. Neff. “Apparent distance of sounds recorded in echoic and anechoic chambers”. In: *J Exp Psychol Hum Percept Perform* 6.4 (Nov. 1980), pp. 745–50.
- [31] E.R. Calcagno, E.L. Abregú, M.C. Eguía, and R. Vargara. “The role of vision in auditory distance perception”. In: *Perception* 41 (2012), pp. 175–192.
- [32] R. Carhart. “Monaural and Binaural Discrimination against Competing Sentences”. In: *The Journal of the Acoustical Society of America* 37.6 (1965), pp. 1205–1205.
- [33] J. Catic, S. Santurette, J.M. Buchholz, F. Gran, and T. Dau. “The effect of interaural-level-difference fluctuations on the externalization of sound”. In: *J Acoust Soc Am* 134.2 (Aug. 2013), pp. 1232–41.
- [34] J. Catic, S. Santurette, and T. Dau. “The role of reverberation-related binaural cues in the externalization of speech”. In: *J Acoust Soc Am* 138 (2015), p. 1154.
- [35] W. Chee. “Yaw Rate Estimation Using Two 1-Axis Accelerometers”. In: 2005 (June 2005), pp. 8–10.
- [36] B. Cornelis, M. Moonen, and J. Wouters. “Reduced-bandwidth Multi-channel Wiener Filter based binaural noise reduction and localization cue preservation in binaural hearing aids”. In: *Signal Process* 99 (2014), pp. 1–16.
- [37] G. Courtois. “Early reflections extraction and cleaning”. In: *Restricted. Internal report, EPFL* (2017).
- [38] G. Courtois. “Spatial hearing rendering in wireless microphone systems for binaural hearing aids”. PhD thesis. Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2016.
- [39] G. Courtois and E. Georganti. *Methods and systems for hearing device signal enhancement using a remote microphone*. Patent No. US 2020/0178004 A1, June 2020.
- [40] G. Courtois, V. Grimaldi, I. Kodrasi, H. Lissek, and E. Georganti. *Experimental evaluation of speech enhancement methods in remote microphone systems for hearing aids*. Heraklion, Crete - Greece: Euronoise, 2018.
- [41] G. Courtois, V. Grimaldi, H. Lissek, P. Estoppey, and E. Georganti. “Perception of Auditory Distance in Normal-Hearing and Moderate-to-Profound Hearing-Impaired Listeners”. In: *Trends in Hearing* 23 (2019). PMID: 31774032, pp. 1–18.

-
- [42] G. Courtois, V. Grimaldi, H. Lissek, I. Kodrasi, and E. Georganti. "Experimental evaluation of speech enhancement methods in remote microphone systems for hearing aids". In: *Proceedings of Euronoise 2018, European Acoustics Association* (2018), pp. 8. 1–8.
- [43] G. Courtois, H. Lissek, P. Estoppey, Y. Oesch, and X. Gigandet. "Effects of Binaural Spatialization in Wireless Microphone Systems for Hearing Aids on Normal-Hearing and Hearing-Impaired Listeners". In: *Trends in Hearing* 22 (2018), pp. 1–17.
- [44] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. "Binaural hearing aids with wireless microphone systems including speaker localization and spatialization," in: Warsaw, Poland: 138th Audio. Eng. Soc. Convention, May 2015.
- [45] R. Crawford-Emery and H. Lee. "The subjective effect of BRIR length perceived headphone sound externalisation and tonal colouration. 136th Audio Eng". In: *Soc. Convention* 26 (Apr. 2014).
- [46] J. Cubick, J.M. Buchholz, V. Best, M. Lavandier, and T. Dau. "Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners". In: *The Journal of the Acoustical Society of America* 144.5 (2018), pp. 2896–2905.
- [47] J. Cubick, S. Santurette, S. Laugesen, and T. Dau. "The influence of visual cues on auditory distance perception". In: *Forschritte der Akustik DAGA'15* (2015), pp. 1220–1223.
- [48] D.D. Dirks and R.H. Wilson. "The Effect of Spatially Separated Sound Sources on Speech Intelligibility". In: *Journal of Speech and Hearing Research* 12.1 (1969), pp. 5–38.
- [49] S. Doclo, S. Gannot, M. Moonen, and A. Spriet. "Acoustic beamforming for hearing aid applications". In: *Handbook on Array Processing and Sensor Networks, Chapter 10* (2008).
- [50] S. Doclo, A. Spriet, J. Wouters, S. Gannot, and M. Moonen. "Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction". In: *Speech Enhancement* (2005), pp. 199–228.
- [51] T.S. Donaldson. "Robustness of the F-Test to Errors of Both Kinds and the Correlation Between the Numerator and Denominator of the F-Ratio". In: *Journal of the American Statistical Association* 63.322 (1968), pp. 660–676.
- [52] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. "On the Externalization of Auditory Images". In: *Presence: Teleoperators and Virtual Environments* 1.2 (May 1992), pp. 251–257.
- [53] N. I. Durlach, C. L. Thompson, and H. S. Colburn. "Binaural Interaction in Impaired Listeners: A Review of Past Research". In: *Audiology* 20.3 (1981), pp. 181–211.
- [54] C. Faller, F. Menzer, and C. Tournery. *Binaural Audio with Relative and Pseudo Head Tracking*. 138th Audio Eng. Warsaw, Poland: Soc. Convention, 2015.

BIBLIOGRAPHY

- [55] M. Farmani, M. Syskind Pedersen, and Jesper J. “Sound Source Localization for Hearing Aid Applications Using Wireless Microphones”. In: *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. 2018, pp. 455–459.
- [56] A. Ferlini, A. Montanari, C. Mascolo, and R. Harle. “Head Motion Tracking Through in-Ear Wearables”. In: *Proceedings of the 1st International Workshop on Earable Computing. , United Kingdom, Association for Computing Machinery*. 2019, pp. 8–13.
- [57] A. Field, J. Miles, and Z. Field. *Discovering statistics using R*. Sage London, 2012.
- [58] EP. Fowler. “The recruitment of loudness phenomenon”. In: *Laryngoscope* 60 (July 1950), p. 680.
- [59] R.L. Freyman, K.S. Helfer, D.D. McCall, and R.K. Clifton. “The role of perceived spatial separation in the unmasking of speech”. In: *The Journal of the Acoustical Society of America* 106.6 (Dec. 1999), pp. 3578–3588. ISSN: 0001-4966.
- [60] M. Froehlich, E. Branda, and K. Freels. “New dimensions in automatic steering for hearing aids: clinical and real-world findings”. In: *Hearing Review* 26.11 (2019), pp. 32–36.
- [61] M.B. Gardner. “Proximity Image Effect in Sound Localization”. In: *The Journal of the Acoustical Society of America* 43.1 (1968), pp. 163–163.
- [62] W. Gardner. “Efficient convolution without input-output delay”. In: *journal of the audio engineering society* 43.3 (Mar. 1995), pp. 127–136.
- [63] J.C. Gil-Carvajal, J. Cubick, S. Santurette, and T. Dau. “Spatial Hearing with Incongruent Visual or Auditory Room Cues, Scientific Reports, vol. 6, Art. numb.: 37342”. In: 2016 () .
- [64] P. Giller, F. Wendt, and R. Höldrich. “The influence of different BRIR modification techniques on externalization and sound quality”. In: Sept. 2019.
- [65] G.V. Glass, P.D. Peckham, and J.R. Sanders. “Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance”. In: *Review of Educational Research* 42.3 (1972), pp. 237–288.
- [66] V. Goverdovsky, W. von Rosenberg, T. Nakamura, D. Looney, D. Sharp, C. Papavassiliou, M. Morrell, and D. Mandic. “Hearables: Multimodal physiological in-ear sensing”. In: *Scientific Reports* 7 (Sept. 2016).
- [67] V. Grimaldi, G. Courtois, L.S.R. Simon, and H. Lissek. “Externalization of virtual sounds using low computational cost spatialization algorithms for hearables”. In: *Forum Acusticum, Lyon, France* (Dec. 2020), pp. 917–921.
- [68] V. Grimaldi, H. Lissek, G. Courtois, E. Georganti, and P. Estoppey. “Auditory externalization in hearing impaired listeners with remote microphone systems for hearing aids”. In: *Proc. of the 26th ICSV Conv., Montreal, Canada* (2019).
- [69] V. Grimaldi, L.S.R. Simon, M. Sans, G. Courtois, and H. Lissek. “Human Head Yaw Estimation based on Two 3-axis Accelerometers”. In: *IEEE Sensors* Under review, submitted (2022).

-
- [70] J.H.L. Hansen and B.L. Pellom. “An effective quality evaluation protocol for speech enhancement algorithms”. In: *ICSLP*. 1998.
- [71] W.M. Hartmann and A. Wittenberg. “On the externalization of sound images”. In: *J Acoust Soc Am* 99.3678 (), p. 1996.
- [72] H.G. Hassager, F. Gran, and T. Dau. “The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment”. In: *J Acoust Soc Am* 139 (May 2016), p. 5.
- [73] H.G. Hassager, A. Wiinberg, and T. Dau. “Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment”. In: *The Journal of the Acoustical Society of America* 141.4 (2017), pp. 2556–2568.
- [74] E. Hendrickx, P. Stitt, J.C. Messonnier, J.M. Lyzwa, B. FG. Katz, and de C. Boishéraud. “Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis”. In: *The Journal of the Acoustical Society of America* 141.3 (2017), pp. 2011–2023.
- [75] N. Hunn. *Hearables – the new Wearables: Wearable Technologies*. Aug. 2015.
- [76] Y. Iwaya, Y. Suzuki, and D. Kimura. “Effects of head movement on front-back error in sound localization”. In: *Acoustical Science and Technology* 24 (Sept. 2003), pp. 322–324.
- [77] L.A. Jeffress and R.W. Taylor. “Lateralization vs Localization”. In: *The Journal of the Acoustical Society of America* 33.4 (1961), pp. 482–483.
- [78] T. G. Stockham Jr. “High-Speed Convolution and Correlation”. In: *Managing Requirements Knowledge, International Workshop on*. Vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 1966, p. 229.
- [79] J.M. Kates and K.H. Arehart. “Improving auditory externalization for hearing-aid remote microphones”. In: *2017 51st Asilomar Conference on Signals, Systems, and Computers*. 2017, pp. 1895–1899.
- [80] J.M. Kates, K.H. Arehart, and L.O. Harvey. “Integrating a remote microphone with hearing-aid processing”. In: *The Journal of the Acoustical Society of America* 145.6 (2019), pp. 3551–3566.
- [81] J.M. Kates, K.H. Arehart, R.K. Muralimanohar, and K. Sommerfeldt. “Externalization of remote microphone signals using a structural binaural model of the head and pinna”. In: *J. Acoust. Soc. Am* 143 (May 2018), p. 5.
- [82] J.M. Kates and K.H. Areheart. “The Hearing-Aid Speech Quality Index (HASQI) Version 2”. In: *J. Audio Eng* 62 (2014), p. 3.
- [83] J. Kawaura, Y. Suzuki, F. Asano, and T. Sone. “Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear”. In: *Journal of the Acoustical Society of Japan (E)* 12.5 (1991), pp. 203–216.

BIBLIOGRAPHY

- [84] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery. "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers". In: *International Journal of Audiology* 45.10 (2006), pp. 563–579.
- [85] B. Keshavarz, L. Hettinger, D. Vena, and J. Campos. "Combined effects of auditory and visual cues on the perception ofvection". In: *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale* 232 (Dec. 2013).
- [86] B. Keshavarz, L.J. Hettinger, R.S. Kennedy, and J.L. Campos. "Demonstrating the Potential for Dynamic Auditory Stimulation to Contribute to Motion Sickness". In: *PLOS ONE* 9 (July 2014), pp. 1–9.
- [87] S.M. Kim and W. 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.
- [88] Y.G. Kim, C.J. Chun, H.K. Kim, Y.J. Lee, D.Y. Jang, and K. Kang. "An integrated approach of 3D sound rendering techniques for sound externalization". In: *Advances in Multimedia Information Processing—PCM*. K.M. Lam, H. Kiya, X.-Y. Xue, C.-C. J. Kuo, and M. S. Lew: G. Qiu, 2010, pp. 682–693.
- [89] Kionix. "Using Two Tri-Axis Accelerometers for Rotational Measurements". Document Number: AN 019. 2015.
- [90] M. Kok, J.D. Hol, and T.B. Schön. "Using Inertial Sensors for Position and Orientation Estimation, :" in: *Foundations and Trends in Signal Processing* 11.1-2 (2018), pp. 1–153.
- [91] A.J. Kolarik, C.J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss". In: *Atten Percept Psychophys* 78.2 (2016), pp. 373–395.
- [92] N. Kopco and B.G. Shinn-Cunningham. "Effect of stimulus spectrum on distance perception for nearby sources". In: *J Acoust Soc Am* 130 (2011), p. 1530.
- [93] E. Kreyszig. *Advanced Engineering Mathematics (Fourth ed.)* Wiley, 1979, p. 880.
- [94] A. Kuklasinski, S. Doclo, S.H. Jensen, and J. Jensen. "Multi-channel Wiener Filter for Speech Dereverberation in Hearing Aids - Sensitivity to DoA Errors". In: *AES* 60 (2016).
- [95] A. Kulkarni and H.S. Colburn. "Role of spectral detail in sound-source localization". In: *Nature* 396 (Dec. 1998), pp. 747–9.
- [96] B.D. Kulp. "Digital Equalization Using Fourier Transform Techniques". In: *Audio Engineering Society Convention* 85. Nov. 1988.
- [97] T. Leclère, M. Lavandier, and F. Perrin. "On the externalization of sound sources with headphones without reference to a real source". In: *The Journal of the Acoustical Society of America* 146.4 (2019), pp. 2309–2320.
- [98] S. Li, Jiaxiang E, R. Schlieper, and J. Peissig. "The impact of trajectories of head and source movements on perceived externalization of a frontal sound source". In: *Journal of the Audio Engineering Society* (May 2018).

-
- [99] S. Li, E. Jiaxiang, R. Schlieper, and J. Peissig. "The Impact of Trajectories of Head and Source Movements on Perceived Externalization of a Frontal Sound Source". In: *AES Convention*: 144 (May 2018).
 - [100] S. Li, R. Schlieper, and J. Peissig. "The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room". In: *The Journal of the Acoustical Society of America* 144.2 (2018), pp. 966–980.
 - [101] S. Li, R. Schlieper, and J. Peissig. "The impact of head movement on perceived externalization of a virtual sound source with different BRIR lengths". In: *Journal of the Audio Engineering Society* (Mar. 2019).
 - [102] S. Li, R. Schlieper, and J. Peissig. "The Role of Reverberation and Magnitude Spectra of Direct Parts in Contralateral and Ipsilateral Ear Signals on Perceived Externalization". In: *Applied Sciences* 9.3 (2019). ISSN: 2076-3417.
 - [103] S. Li, R. Schlieper, A. Tobbala, and J. Peissig. "The Influence of Binaural Room Impulse Responses on Externalization in Virtual Reality Scenarios". In: *Applied Sciences* 11 (Oct. 2021), p. 10198.
 - [104] A. Lindau. "The perception of system latency in dynamic binaural synthesis". In: *Proceedings of NAG/DAGA 2009 - Rotterdam*. 2009.
 - [105] R.Y. Litovsky, H.S. Colburn, W.A. Yost, and S.J. Guzman. "The precedence effect". In: *The Journal of the Acoustical Society of America* 106.4 (1999), pp. 1633–1654.
 - [106] A.D. Little, D.H. Mershon, and P.H. Cox. "Spectral content as a cue to perceived auditory distance". In: *Perception*. 21.3 (1992), pp. 405–16.
 - [107] J.P.A. Lochner and J.F. 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.
 - [108] J.M. Loomis, C. Hebert, and J.G. Cicinelli. "Active localization of virtual sounds". In: *The Journal of the Acoustical Society of America* 88.4 (1990), pp. 1757–1764.
 - [109] J.M. Loomis, R.L. Klatzky, J.W. Philbeck, and R.G. Golledge. "Assessing auditory distance perception using perceptually directed action". In: *Percept Psychophys*. 60.6 (Aug. 1998), pp. 966–80.
 - [110] O.M. Lord Rayleigh and R.S. Pres. "XII. On our perception of sound direction". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74 (1907), pp. 214–232.
 - [111] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, J. Nyström Å. and Pettersen, and R. Bergman. "Experimental design and optimization". In: *Chemometrics and intelligent laboratory systems* 42.1-2 (1998), pp. 3–40.
 - [112] M. Luzardo, M. Karppa, J. Laaksonen, and T. Jantunen. "Head Pose Estimation for Sign Language Video". In: *Image Analysis*. Ed. by Joni-Kristian Kämäräinen and Markus Koskela. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 349–360.

BIBLIOGRAPHY

- [113] R.G. Lyons. *Understanding Digital Signal Processing*. Prentice Hall, 2011. ISBN: 9780137-027415.
- [114] R.L. Martin, K.I. Mcanally, and M.A. Senova. “free-field equivalent localization of virtual audio”. In: *journal of the audio engineering society* 49.1/2 (Jan. 2001), pp. 14–22.
- [115] K.I. McAnally and R.L. Martin. “Sound localization with head movement: implications for 3-d audio displays”. In: *Frontiers in Neuroscience* 8 (2014), p. 210. ISSN: 1662-453X.
- [116] C. Mendonça. “A review on auditory space adaptations to altered head-related cues”. In: *Frontiers in Neuroscience* 8 (2014), p. 219. ISSN: 1662-453X.
- [117] C. Mendonça, G. Campos, P. Dias, and J.A. Santos. “Learning Auditory Space: Generalization and Long-Term Effects”. In: *PLOS ONE* 8.10 (Oct. 2013), null.
- [118] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. p. Ferreira, and J.A. Santos. “On the improvement of localization accuracy with non-individualized hrtf-based sounds”. In: *journal of the audio engineering society* 60.10 (Oct. 2012), pp. 821–830.
- [119] C. Mendonça and S. Delikaris-Manias. “Statistical tests with MUSHRA data”. In: *journal of the audio engineering society*. May 2018.
- [120] D.H. Mershon, W.L. Ballenger, A.D. Little, P.L. McMurtry, and J.L. Buchanan. “Effects of Room Reflectance and Background Noise on Perceived Auditory Distance”. In: *Perception*. 18 (1989), p. 3.
- [121] D.H. Mershon, D.H. Desaulniers, T. L. Amerson, and S. A. Kiefer. “Visual capture in auditory distance perception: Proximity image effect reconsidered”. In: *Journal of Auditory Research* 20.2 (1980), pp. 129–136.
- [122] D.H. Mershon and L.E. King. “Intensity and reverberation as factors in the auditory perception of egocentric distance”. In: *Perception & Psychophysics* 18 (1975), p. 409.
- [123] A. W. Mills. “On the minimum audible angle”. In: *The Journal of the Acoustical Society of America* 30.4 (1958), pp. 237–246.
- [124] P. Minnaar, S.K. Olesen, F. Christensen, and H. Møller. “Localization with binaural recordings from artificial and human heads”. In: *journal of the audio engineering society* 49.5 (May 2001), pp. 323–336.
- [125] P. Minnaar, S.K. Olesen, F. Christensen, and H. Møller. “The importance of head movements for binaural room synthesis”. English. In: *Proceedings of ICAD 2001, July 29-August 1, Espoo, Finland*. 1996.
- [126] S. Musa-Shufani, M. Walger, H. von Wedel, and H. Meister. “Influence of dynamic compression on directional hearing in the horizontal plane”. In: *Ear Hear* 27.3 (June 2006), pp. 279–85.
- [127] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen. *Incorporating the Conditional Speech Presence Probability in Multi-Channel Wiener Filter Based Noise Reduction in Hearing Aids*. EURASIP J. Adv. Sig. Pr, 2009.

-
- [128] W. Noble. "Auditory localization in the vertical plane: Accuracy and constraint on bodily movement". In: *The Journal of the Acoustical Society of America* 82.5 (1987), pp. 1631–1636.
 - [129] J. Odelius and "O. Johansson. "Self-assessment of classroom assistive listening devices". In: *Int. J. Audiol.* 49.7 (2010), pp. 508–517.
 - [130] B. Ohl, S. Laugesen, J. Buchholz, and L. Dau. "Externalization versus Internalization of Sound in Normal-hearing and Hearing-impaired Listeners". English. In: *In Fortschritte der Akustik*. Vol. MI. 14:25. Jahrestagung der Deutschen Gesellschaft für Akustik, DAGA 2010 ; Conference date: 15-03-2010 Through 18-03-2010. Deutsche Gesellschaft für Akustik, 2010.
 - [131] World Health Organization. In: *Deafness and hearing loss* 15 (Mar. 2018).
 - [132] World Health Organization. "Prevention of blindness and deafness". In: (2005).
 - [133] K. Papafotis and P. Sotiriadis. "Multiple Accelerometers and Magnetometers Joint Calibration and Alignment". In: *IEEE Sensors* 4 (2020).
 - [134] M. Pedley. "Tilt Sensing Using Linear Accelerometers". In: *(NXP) Freescale Semiconductor, Document Number: AN 3461* (2013).
 - [135] S. Perrett and W. Noble. "The contribution of head motion cues to localization of low-pass noise". In: *Attention, Perception & Psychophysics* 59 (Jan. 1997), pp. 1018–1026.
 - [136] J. Plazak and M. Kersten-Oertel. "A Survey on the Affordances of "Hearables"". In: *Inventions* 3.3 (2018). ISSN: 2411-5134.
 - [137] G. Plenge. "On the differences between localization and lateralization". In: *The Journal of the Acoustical Society of America* 56.3 (1974), pp. 944–951.
 - [138] L. Prud'homme and M. 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.
 - [139] M. Rahme, P. Folkeard, and S. Scollie. "Evaluating the Accuracy of Step Tracking and Fall Detection in the Starkey Livio Artificial Intelligence Hearing Aids: A Pilot Study". In: *American Journal of Audiology* 30.1 (2021), pp. 182–189.
 - [140] M. Risoud, J.-N. Hanson, F. Gauvrit, C. Renard, P.-E. Lemesre, N.-X. Bonne, and C. Vincent. "Sound source localization". In: *European Annals of Otorhinolaryngology, Head and Neck Diseases* 135.4 (2018), pp. 259–264. ISSN: 1879-7296.
 - [141] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *IEEE* 2011 (Nov. 2001), pp. 749–752.
 - [142] U. Röijezon, M. Djupsjöbacka, M. Björklund, H. Häger-Ross C. Grip, and D.G. Liebermann. "Kinematics of fast cervical rotations in persons with chronic neck pain: a cross-sectional and reliability study". In: *BMC Musculoskelet Disord* 11 (2010), p. 222.

BIBLIOGRAPHY

- [143] J. Sandvad. “Dynamic aspects of auditory virtual environments”. In: *journal of the audio engineering society* (May 1996).
- [144] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo. “Acoustic transparency in hearables—perceptual sound quality evaluations”. In: *Journal of the audio engineering society* 68.7/8 (July 2020), pp. 495–507.
- [145] J.G. Selby, A. Weisser, and E.N. MacDonald. “Influence of a remote microphone on localization with hearing aids”. In: 6 (Dec. 2017), pp. 405–411.
- [146] B.G. Shinn-Cunningham, N. Kopco, and T.J. Martin. “Localizing nearby sound sources in a classroom: binaural room impulse responses”. In: *J Acoust Soc Am* 117 (2005), pp. 3001–3115.
- [147] J.A. Da Silva. “Scales for perceived egocentric distance in a large open field: Comparison of three psychological methods”. In: *Am. J. Psychol.* 98 (1985), pp. 119–144.
- [148] H.J. Simon and H. Levitt. “Effect of Dual Sensory Loss on Auditory Localization: Implications for Intervention”. In: *Trends Amplif.* 11.4 (Dec. 2007), pp. 259–272.
- [149] J. Sinker and B. Shirley. *The Effect of Early Impulse Response Length and Visual Environment on Externalization of Binaural Virtual Sources*. 140th Audio Eng. Paris, France: Soc. Convention, 2016.
- [150] G.A. Soulodre, N. Popplewell, and John S. Bradley. “Combined effects of early reflections and background noise on speech intelligibility”. In: *Journal of Sound and Vibration* 135.1 (1989), pp. 123–133.
- [151] T. Sporer, S. Werner, and F. Klein. “Adjustment of the direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System”. In: *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Sept. 2016.
- [152] J.C. Steinberg and M.B. Gardner. “The dependence of hearing impairment on sound intensity”. In: *The Journal of the Acoustical Society of America* 9.1 (1937), pp. 11–23.
- [153] P. Stitt, E. Hendrickx, J-C. Messonnier, and B. Katz. “the influence of head tracking latency on binaural rendering in simple and complex sound scenes”. In: *journal of the audio engineering society* (May 2016).
- [154] J. Szurley, A. Bertrand, B. Van Dijk, and M. Moonen. “Binaural Noise Cue Preservation in a Binaural Noise Reduction System with a Remote Microphone Signal”. In: *IEEE T. Audio* 24 (2016).
- [155] W.R. Thurlow and P.S. Runge. “Effect of Induced Head Movements on Localization of Direction of Sounds”. In: *The Journal of the Acoustical Society of America* 42.2 (1967), pp. 480–488.
- [156] E.E. Toole. “In-Head Localization of Acoustic Images”. In: *J Acoust Soc Am* (1970), pp. 48–943.

-
- [157] A. Torger and A. Farina. "Real-time partitioned convolution for Ambiophonics surround sound". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. 2001, pp. 195–198.
 - [158] J. Van Opstal. *The auditory system and human sound-localization behavior*. Academic Press, 2016.
 - [159] F. Völk. "Externalization in data-based binaural synthesis: effects of impulse response length". In: *Proc. of Intern. Conf. on Acoustics NAG/DAGA*, 2009, pp. 1075–1078.
 - [160] S. C. Voss, K. Pichora-Fuller, A. Pereira, J. Seiter, N. Guindi, and J. Qian. *Evaluating the accuracy of motion detection using a behind-the-ear sensor*. Arizona, USA: The Annual Scientific and Technology Conference of the American Auditory Society. Scottsdale, 2020.
 - [161] S.C. Voss, K. Pichora-Fuller, I. Ishida, A. Pereira, J. Seiter, N. El Guindi, V. Kuehnel, and J. Qian. "Evaluating the benefit of hearing aids with motion-based beamformer adaptation in a real-world setup". In: *International Journal of Audiology* (2021), pp. 1–13.
 - [162] H. Wallach. "The role of head movements and vestibular and visual cues in sound localization." In: *Journal of Experimental Psychology* 27 (1940), pp. 339–368.
 - [163] D.H. Warren, R.B. Welch, and T.J. McCarthy. "The role of visual-auditory compellingness in the ventriloquism effect: Implications for transitivity among the spatial senses Percept Psychophys". In: *Vol. 30.6* (1981), pp. 557–564.
 - [164] E.M. Wenzel. "Effect of Increasing System Latency on Localization of Virtual Sounds". In: *AES 16th International conference on Spatial Sound Reproduction*. 1999.
 - [165] E.M. Wenzel. "Effect of increasing system latency on localization of virtual sounds with short and long duration". In: *Proceedings of the 2001 International Conference on Auditory Display*. 2001, pp. 185–290.
 - [166] E.M. Wenzel. "The relative contribution of interaural time and magnitude cues to dynamic sound localization". In: *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. 1995, pp. 80–83.
 - [167] E.M. Wenzel, M. Arruda, Doris J. Kistler, and FL. Wightman. "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1 (1993), pp. 111–123.
 - [168] E.M. Wenzel, FL. Wightman, and D.J Kistler. "Localization with non-individualized virtual acoustic display cues". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1991, pp. 351–359.
 - [169] EM. Wenzel, M. Arruda, DJ. Kistler, and FL. Wightman. "Localization using non-individualized head-related transfer functions". In: *J Acoust Soc Am* (1993), pp. 111–23.
 - [170] S. Werner, F. Klein, and K. Mayenfels T.and Brandenburg. "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events". In: June 2016.

BIBLIOGRAPHY

- [171] F. L. Wightman and Doris J. Kistler. "Resolution of front-back ambiguity in spatial hearing by listener and source movement". In: *The Journal of the Acoustical Society of America* 105.5 (1999), pp. 2841–2853.
- [172] J. Wolfe, E. Schafer, N. Martella, M. Morais, and M. Mann. "Evaluation of Extended-Wear Hearing Technology for Children with Hearing Loss". In: *J Am Acad Audiol*. 26.7 (2015), pp. 615–31.
- [173] R. Woodworth, B. Barber, and H. Schlosberg. *Experimental psychology*. Oxford and IBH Publishing, 1954.
- [174] P. Zahorik. "Assessing auditory distance perception using virtual acoustics". In: *J Acoust Soc Am* 111 (2002), p. 1832.
- [175] P. Zahorik. "Auditory display of sound source distance". In: *Proc. Int. Conf. on Auditory Display*. 2002, pp. 326–332.
- [176] P. Zahorik. "Distance localization using non-individualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 108.5 (2000), pp. 2597–2597.
- [177] P. Zahorik. "Estimating sound source distance with and without vision". In: *Optometry and vision science* 78.5 (2001), pp. 270–275.
- [178] P. Zahorik, D.S. Brungart, and A.W. Bronkhorst. "Auditory distance perception in humans: A summary of past and present research". In: *Acta Acustica* 91 (2005), pp. 409–420.
- [179] P. Zahorik, D.J. Kistler, and F. L. Wightman. "Sound localization in varying virtual acoustic environments". In: Georgia Institute of Technology. 1994.
- [180] J. Zhao. "A Review of Wearable IMU (Inertial-Measurement-Unit)-based Pose Estimation and Drift Reduction Technologies". In: *J. Phys* 1087 (2018).
- [181] X.L. Zhong and B.S. Xie. *Head-Related Transfer Functions and Virtual Auditory Display, Soundscape Semiotics - Localization and Categorization*. InTech, 2014.
- [182] E. Zwicker. "Subdivision of the audible frequency range into critical bands". In: *J. Acoust. Soc. Am* 33 (1961), pp. 248–248.

Curriculum Vitae

Vincent Grimaldi was born in Paris and grew up in Angers, France. He received the Dipl. Ing. degree (BSc/MSc combined) from École Nationale des Art et Métiers ParisTech (Lille and Paris, France) with a third year major in Bioengineering in 2015. In 2016, he received a second MSc in Sciences and Technologies "Parcours ATIAM" (Signal Processing, Acoustics and Computer Science applied to Music) from Université Pierre-et-Marie-Curie (UPMC, Paris, France), in collaboration with the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) and Telecom ParisTech. He joined the École Polytechnique Fédérale de Lausanne (EPFL) as a doctoral assistant in the Acoustic Group of the Signal Processing Laboratory 2 (LTS2) in 2017. As a teaching assistant, he supervised multiple bachelor and master semester projects as well as internships. He has also been involved in theoretical and practical exercises classes in Audio Engineering.

List of publications

Journal papers

V. Grimaldi, L.S.R. Simon, M. Sans, G. Courtois, H. Lissek. "Human head yaw estimation based on two 3-axis accelerometers". *IEE Sensors, Submitted in Jan. 2022.*

G. Courtois, V. Grimaldi, H. Lissek, P. Estoppey, E. Georganti. "Perception of auditory distance in normal-hearing and moderate-to-profound hearing-impaired listeners". *Trends in Hearing, Volume 23: 1-18, 2019.*

Conference papers

V. Grimaldi, G. Courtois, L.S.R. Simon, H. Lissek. "Externalization of virtual sounds using low computational cost algorithms for hearables". *Proc. Forum Acusticum Lyon, December 2020.*

V. Grimaldi, G. Courtois, P. Estoppey, E. Georganti, H. Lissek. "Auditory externalization in hearing-impaired listeners with remote microphone systems for hearing aids". *Proc. ICSV, Montréal, Canada, July 2019.*

V. Grimaldi, G. Courtois, E. Georganti, H. Lissek. "Objective evaluation of static beamforming on the quality of speech in noise". *Proc. Euronoise, Heraklion, Crete, Greece, May 2018.*

G. Courtois, V. Grimaldi, E. Georganti, H. Lissek. "Experimental evaluation of speech enhancement methods in remote microphone systems for hearing aids". *Proc. Euronoise, Heraklion, Crete, Greece, 2018.*

V. Grimaldi, C. Böhm, S. Weinzierl and H. von Coler. "Parametric synthesis of crowd noises in virtual acoustic environments". *Journal of the Audio Engineering Society, 2017.*