Evaluation of audio source separation in the context of 3D audio

THÈSE N^{O} 6723 (2015)

PRÉSENTÉE LE 28 AOÛT 2015

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE D'ÉLECTROMAGNÉTISME ET ACOUSTIQUE
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Lukas ROHR

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. J. R. Mosig, Dr H. Lissek, directeurs de thèse
Dr M. Rosenthal, rapporteur
Dr D. Schröder, rapporteur
Prof. T. Ebrahimi, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

To my beautiful wife…

I like to listen.
I have learned a great deal from listening carefully.
Most people never listen.
— Ernest Hemingway —

Acknowledgements

The completion of this work was only made possible through the contributions of numerous people and organisations and I would like to thank them in the preamble to this thesis.

First, in a more formal part, the financial and thematic framework of this thesis was provided by the i3Dmusic project, supported by the EUREKA Eurostars Programme and co-funded by the Swiss Federal Office for Professional Education and Technology (grant agreement E!5582) and the European Union (grant agreement INT.2010.0023). The past few years would probably look different for me, if this project had not seen the light of day.

I would like to thank the two co-directors of my thesis: Prof. Juan Mosig, director of the Laboratory of Electromagnetics and Acoustics and Dr. Hervé Lissek, head of the Acoustics Group, now part of the Signal Processing Laboratory 2 at EPFL. Thank you Juan for accepting to direct this thesis and welcoming me in the lab, providing the support for Hervé's group to be able to function. You always spread a good mood and you make sure to keep the lab's informal traditions that contribute to keeping it together. I wish you a long and joyous soon-to-be retirement. In that context, thanks also to Prof. Pierre Vandergheynst for welcoming the Acoustics Group in his lab since 2015.
Thank you, Hervé, for having me in the Acoustics Group. Your feedback proved to be valuable and your support was very welcome, always making sure that I had all I needed to perform my research. A solution could always be found to any problem that I presented you with. Thank you for your flexibility with my somewhat uncommon working hours and your confidence in my work.

I also want to express my gratitude to the members of the jury, Prof. Jean-Philippe Thiran, Prof. Touradj Ebrahimi, Dr. Matthias Rosenthal and Dr. Dirk Schröder for their interest in my work and their valuable feedback.

My thanks go to the whole i3Dmusic team who largely contributed to this thesis, allowing me to start out in the topic and providing valuable insights, feedback, criticism and encouragement. Special thanks to: Dr. Etienne Corteel and Dr. Khoa-Van NGuyen for the continuous collaboration, especially at the beginning of the project, and the support with the 3D WFS system; Dr. Emmanuel Vincent and Dr. Laurent Simon for all the discussions about the psychoacoustic experiments conducted in this thesis and for actually running the experiment presented in chapter 6; Dr. Pierre Leveau for your feedback about all source separation matters,

but also for heading the project, organising all the meetings, etc.
Although not involved in the project, Prof. Anthony C. Davison and Dr. Jean-Marie Fuerbringer allowed me to put their extensive statistical knowledge to good use, for which I'm very grateful.

I would also like to thank the present and former members of the Laboratory of Electromagnetics and Acoustics who I had the privilege to know and work with. You made it a memorable journey in a good working atmosphere. My special thanks to: Dr. Xavier Falourd, for introducing me to the world of acoustics through my Master thesis and for helping out at the start of my work on this thesis; Dr. Patrick Marmaroli, for the scientific discussions that were always fruitful and cheerful at the same time; Dr. Sami Karkar, Dr. Romain Boulandet and Etienne Rivet, with whom I shared an office for a while and who could always answer my more technical questions; Gilles Courtois, for sharing the office with me during the last part of this work and always being open to psychoacoustic (amongst other) discussions; Baptiste Crettaz and Cédric Monchâtre for the technical support at the lab (many cables have been made and many experiments conducted); David Desscan for the tireless IT support; Eulalia Durussel and Mercedes Quintas for their administrative support, facilitating the formalities and always available for any organisational request.

Finally, thanks to my friends and family who supported me throughout the period of this thesis. Very special thanks to my parents Christian and Susanne, for allowing me to choose the long path of university studies and supporting me all along. My deepest gratitude goes to my wife, Tabea, who is always there for me. Thank you for your patience, your love, your support and all the other unquantifiable contributions to this work and to my life.

Lausanne, 29th May 2015 Lukas

Abstract

The emergence and broader availability of 3D audio systems allows for new possibilities in mixing, post-production and playback of audio content. Used in movie post-production for cinemas, as special effect by disk jockeys for example and even for live concerts, 3D rendering immerses the listener more than ever before. When existing audio material is to be employed, Audio Source Separation (ASS) techniques enable the extraction of single sources from a mixture. Modern mixing approaches for 3D audio do not assign individual gains and delays for each source in every channel. A sound scene is rather designed, with individual sources treated as objects to be placed within a scene. The hardware layer is mostly irrelevant for mixing in such a setting. ASS is therefore a valuable tool to "disassemble" a more traditional monophonic, stereophonic, or multichannel mix. However, due to the complexity of the ASS problem, extracted sources are subject to degradations. While state-of-the-art objective measures for ASS quality build on monaural auditory models, they don't take into account binaural listening and the psychoacoustic phenomena that are involved, such as binaural unmasking.

In this thesis, an extension to Perceptive Evaluation Methods for Audio Source Separation (PEASS) [42] is proposed with spatial rendering in mind. Additionally a new binaural model for ASS evaluation in the context of 3D audio is presented. The performance of the basic and extended versions of PEASS, as well as the proposed binaural model is evaluated in two subjective studies. The first study is conducted with binaural spatialisation presented over headphones, while the second experiment uses a 3D Wave Field Synthesis (WFS) system. A set of artificial ASS degradation algorithms is proposed and used for the stimuli of the subjective studies. Results of the studies indicate monotonic decrease of the perceived quality as a function of the amounts of degradations introduced. The most important degradation is found to be target distortion, followed by onset misallocation and musical noise-type artifacts. Additionally, spatialising the extracted target source away from the residue or having it louder than the residue negatively affects the results, indicating a perceived quality degradation. In 3D WFS conditions, results show evidence for monaural and binaural unmasking.

The performance of the proposed binaural model is consistently superior to that of the basic or extended PEASS versions. In the binaural spatialisation experiment, a correlation coefficient of 0.60 between subjective and objective results is achieved, versus 0.57 and 0.53 with the extended and basic PEASS version respectively. For the 3D WFS study, the binaural model achieves 0.67 prediction accuracy whereas both PEASS versions get 0.57.

The perceptual validity of the WFS formulation is also verified in a localisation experiment. Vertical localisation is found to be nearly as good as physical source localisation for an extended listening area with localisation precision of 6° - 9°. The response time is also used as an indicator of localisation performance.

Key words: audio source separation evaluation, subjective audio quality evaluation, objective audio quality evaluation, spatial audio, 3D wave field synthesis, audio source separation degradations, binaural listening, binaural model

Résumé

L'émergence et la disponibilité commerciale de systèmes audio 3D ouvrent de nouvelles possibilités pour le mixage, la post-production et le rendu de contenu audio. Utilisé dans la post-production de films pour le cinéma, en tant qu'effet spécial par des DJs par exemple, ou même pour des concerts en direct, le rendu 3D immerge l'auditeur plus que jamais auparavant. Quand du matériel audio existant est utilisé, les techniques de séparation de sources audio (SSA) permettent l'extraction de sources isolées d'un mélange. Les approches modernes de mixage audio 3D n'assignent pas un gain et un délai individuel à chaque source dans chaque canal. Une scène sonore est plutôt conçue, en considérant les différentes sources comme des objets à placer dans celle-ci. Le niveau matériel est donc essentiellement irrélévant pour le mixage dans un tel environnement. La SSA est donc un outil précieux pour "désassembler" un mélange monophonique, stéréophonique ou multicanal plus traditionnel. La complexité du problème de la SSA a pour conséquence que les sources extraites contiennent généralement des dégradations. Tandis que les mesures objectives de qualité de la SSA actuelles se basent sur des modèles auditifs monoraux, elles ne prennent pas en compte l'écoute binaurale et les phénomènes psychoacoustiques qui y sont associés, tels que le démasquage spatial.

Cette thèse propose une extension à l'état de l'art donné par PEASS [42] en prenant en compte un rendu spatialisé. Un nouveau modèle binaural pour l'évaluation de la SSA dans le contexte de l'audio 3D est également présenté. La performance des versions de base et étendue de PEASS, ainsi que du modèle binaural proposé est évaluée au travers de deux études subjectives. La première étude est conduite en conditions de spatialisation binaurale présentée au casque, tandis que la seconde expérience utilise un système WFS 3D. Des algorithmes de dégradation artificielle de SSA sont présentés et utilisés pour les stimuli des études subjectives. Les résultats indiquent une décroissance monotonique de la qualité perçue en fonction de la quantité de dégradations introduites. La distortion de la source cible semble être la dégradation la plus importante, suivie par la mauvaise affectation des débuts de sons et les artéfacts de type bruit musical. Le déplacement de la source cible par rapport au résidu ou un niveau plus fort de la cible relativement au résidu influencent également négativement les résultats, indiquant une perte de qualité perçue. En conditions WFS 3D, les résultats prouvent la présence de démasquage monaural et binaural.

La performance du modèle binaural proposé est systématiquement supérieure à celle de la version basique ou étendue de PEASS. Lors de l'expérience à spatialisation binaurale, un coefficient de corrélation de 0.60 entre les résultats subjectifs et objectifs est atteint, contre 0.57 et 0.53 pour la version étendue et basique de PEASS respectivement. Pour l'étude WFS 3D, le modèle binaural atteint une précision de prédiction de 0.67, alors que les deux versions de PEASS arrivent à 0.57.

La validitié perceptuelle de la formulation WFS utilisée est également vérifiée à l'aide d'une expérience de localisation. La localisation verticale se révèle être quasiment aussi précise que pour des sources physiques pour une aire d'écoute étendue, avec une précision de localisation de 6° à 9°. Le temps de réponse est également utilisé en tant qu'indicateur de performance de localisation.

Mots clefs : séparation de sources audio, évaluation subjective de qualité audio, évaluation objective de qualité audio, spatialisation audio, synthèse de champs d'onde 3D, dégradations de séparation de source audio, écoute binaurale, modèle binaural

Zusammenfassung

Das Aufkommen und die kommerzielle Verfügbarkeit von 3D Tonausgabesystemen eröffnen neue Möglichkeiten bei der Abmischung, der Postproduktion und der Wiedergabe von Audiomaterial. Die 3D Audiowiedergabe kann etwa bei nachbearbeiteten Kinofilmen benutzt werden, von DJs als Spezialeffekt eingesetzt werden, oder sogar an Konzerten zum Einsatz kommen und sie gibt dem Hörer ein Erlebnis noch nie dagewesener Immersion. Wenn vorhandenes Audiomaterial benutzt werden soll, ermöglicht Audio-Quellentrennung (AQT) die Extrahierung einzelner Quellen aus einer Summe. Moderne Herangehensweisen für die 3D Abmischung weisen nicht jeder Quelle in jedem Kanal eine Amplitude und eine Verzögerung zu. Stattdessen wird eine Klangszene erstellt, in der einzelne Quellen räumlich platziert werden. Die Hardware-Ebene ist in solch einem Vorgang weitgehend irrelevant. Die AQT ist also ein wertvolles Werkzeug um ein traditionelleres Mono-, Stereo- oder Multikanal-Summensignal zu zerlegen. Die Komplexität des AQT Problems hat jedoch zur Folge, dass extrahierte Quellen Qualitätsverluste hinnehmen müssen. Aktuelle Qualitätsmaße für AQT bauen auf monaurale Hörmodelle auf, in welchen binaurale Hörvorgänge und die damit verbundenen psychoakustischen Phänomene, wie räumliches Demaskieren, nicht berücksichtigt werden.

In dieser Dissertation wird eine Erweiterung zum Stand der Technik nach PEASS [42] im Kontext von räumlichem Hören vorgeschlagen. Zusätzlich wird ein binaurales Modell zur Qualitätsbeurteilung von AQT für 3D Audio vorgestellt. Die Leistung der aktuellen und erweiterten Versionen von PEASS, sowie die des vorgestellten binauralen Modells werden in zwei subjektiven Studien beurteilt. In der ersten Studie wird eine Annäherung des räumlichen Schallfeldes über Kopfhörer wiedergegeben, während in der zweiten Studie ein 3D WFS System zum Einsatz kommt. Eine Serie von künstlichen AQT Degradierungsalgorithmen wird vorgeschlagen und für die Erzeugung der Reizklänge für die subjektiven Studien verwendet. Die Ergebnisse der Studien zeigen eine monotone Verminderung der wahrgenommenen Qualität in Abhängigkeit der Menge zugeführter Degradierungen an. Die Verzerrung der Zielquelle scheint der wichtigste Degradierungstyp zu sein, gefolgt von der Fehlallokation von Einsätzen und Störgeräuschen. Zusätzlich wird festgestellt dass die räumliche Trennung von Zielquelle und Hintergrund, sowie ein unausgewogenes Lautstärkeverhältnis zwischen diesen beiden Signalen als Qualitätsverlust wahrgenommen werden. In 3D WFS Bedingungen zeigen die Resultate Hinweise für monaurale und binaurale Reduzierung des Verdeckungseffekts.

Die Leistung des vorgestellten binauralen Modells ist durchgehend besser als jene der aktuellen oder erweiterten Versionen von PEASS. Im Experiment mit Kopfhörern erreicht das neue Modell einen Korrelationskoeffizient von 0.60 zwischen den subjektiven und den objektiven Resultaten, im Gegensatz zu 0.57 mit der erweiterten und 0.53 mit der aktuellen Versionen von PEASS. In der 3D WFS Studie erreicht das binaurale Modell eine Vorhersagegenauigkeit von 0.67, wogegen beide PEASS Versionen auf 0.57 kommen.

Die subjektive Gültigkeit der verwendeten WFS Formulierung wird ebenfalls in einem Lokalisierungs-Experiment geprüft. Den Resultaten zufolge können virtuelle Schallquellen in einem erweiterten Zuhörerbereich fast so gut wie physische Schallquellen mit einer Genauigkeit von 6° bis 9° geortet werden. Die Reaktionszeit wird ebenfalls als Leistungsindikator analysiert.

Stichwörter: Qualitätsbeurteilung Audio-Quellentrennung, subjektive Audio-Qualitätsbeurteilung, objektive Audio-Qualitätsbeurteilung, 3D Tonausgabe, 3D Wellenfeldsynthese, Audio-Quellentrennungs-Degradierung, binaurales Hören, binaurales Modell

Contents

List of Figures

List of Tables

List of acronyms

AIC Akaike information criterion

ANN Artificial neural network

APS Artifact-related perceptual score

ASS Audio source separation

BIC Bayesian information criterion

BMLD Binaural masking level difference

BRIR Binaural room impulse response

CASP Computational auditory signal-processing and perception model

DCT Discrete cosine transform

DIX Distortion index

ERB Equivalent rectangular bandwidth

FFT Fast fourier transform

FIR Finite impulse response

FSNR Frequential signal-to-noise ratio

HOA Higher order ambisonics

HpTF Headphone transfer function

HRIR Head-related impulse response

HRTF Head related transfer function

IACC Interaural cross-correlation

ICC Intraclass correlation coefficient

IID Interaural intensity difference

IIR Infinite impulse response

ILD Interaural level difference

IPD Interaural phase difference

IPS Interference-related perceptual score

ISI Intersymbol interference

ISO International organisation for standardisation

ISR Source image to spatial distortion ratio

ITD Interaural time difference

ITU International telecommunication union

LMM Linear mixed model

MAA Minimum audible angle

ML Maximum likelihood

MLT Modulated lapped transform

MOA Mixed-order ambisonics

MOS Mean objective score

MUSHRA Multiple stimulus with hidden reference and anchor

NMR Noise-to-mask ratio

OASE Objective audio signal evaluation

OLA Overlap-and-add

OPS Overall perceptual score

PAQM Perceptual audio quality measure

PEAQ Perceptual evaluation of audio quality

PEASS Perceptive evaluation methods for audio source separation

PEMO-Q Perceptual model of audio quality

PERCEVAL Perceptual evaluation of the quality of audio signals

POM Perceptual objective measure

PSM Perceptual similarity measure

QESTRAL Quality evaluation of spatial transmission and reproduction using an artificial listener

Q-Q Quantile-quantile

RMS Root mean square

Roex Rounded exponential

SAR Signal to artifacts ratio

SDR Signal to distortion ratio

SIR Signal to interference ratio

SNR Signal to noise ratio

SOR Signal to onset ratio

SPL Sound pressure level

TDOA Time-difference of arrival

TPS Target-related perceptual score

TRR Target to residue ratio

VBAP Vector base amplitude panning

WFS Wave field synthesis

Introduction

Context and objectives of this thesis

Spatial audio is used in more and more applications. Professional and consumer-level solutions make use of the more broadly available spatialisation technologies and of the advantages coming with them: increased sense of immersion, better stability of the created spatial impression, etc. However, the use of existing audio material in such applications is not straightforward, since it is often intended and mixed for stereo or surround playback. A simple translation to spatialised audio by simulating the number of loudspeakers adapted to the mixing technique - a technique known as upmixing - is one possibility, but it lacks at least some of the advantages of spatialisation (notably image stability). An alternative to this is given by Audio Source Separation (ASS). ASS algorithms aim at extracting one or several components out of a mixture. This allows the remixing of the content for other systems and spatialisation techniques.

This is the general context for the i3Dmusic project [1]. The goal of this project is to enable the playback of existing musical content on 3D audio systems through the use of ASS techniques and sound spatialisation algorithms. Supported by the EUREKA Eurostars Programme and co-funded by the Swiss Federal Office for Professional Education and Technology (grant agreement E!5582) and the European Union (grant agreement INT.2010.0023), the i3Dmusic project is an international collaboration between academic and industrial partners. The 3D audio knowledge is brought in by the Swiss start-up Sonic Emotion who specializes in the production of consumer and professional 3D audio systems. The ASS side is covered by the French company Audionamix with their experience in post-production and real-life application of ASS algorithms and by INRIA for the more theoretical part. The laboratory of Electromagnetics and Acoustics (LEMA) of EPFL is in charge of the subjective evaluation part of the project in coordination with INRIA and Sonic Emotion, encompassing subjective evaluation of the ASS algorithms for 3D audio applications and usability studies of the developed products.

The objective of this thesis is to provide insights about subjective ASS evaluation in the context of 3D audio and the differences that might be encountered when comparing to the more classical stereophonic approach. Additionally, since subjective evaluation campaigns are

[1] http://i3dmusic.audionamix.com

costly in financial terms and in time, investigations should be undertaken to model the subjective quality of ASS algorithms in the context of spatial audio and based on the state of the art in ASS. While the i3Dmusic project provided the framework of this thesis, its results might therefore be used in the broader context of binaural auditory modelling and more general audio quality evaluation, especially because the use of 3D audio is on the rise and evaluation tools for audio quality in this context are therefore still to be developed.

Motivation

ASS, and especially blind ASS, where no *a priori* information about the sources in a mix is known, is an active area of research, with complex mathematical, statistical and modelling aspects. This is due to the wide variety of possible source characteristics (frequency content, temporal evolution, etc.) and the associated unpredictability. State-of-the-art ASS algorithms are therefore far from perfect, yet regularly improving. This results in errors in the extracted estimated source signals, leading to potentially audible degradations.

Those degradations may become even more audible on 3D audio systems than in the stereo case for which most current ASS algorithms are designed. The use of signals issued from ASS in the context of spatial audio poses a series of challenges that need to be addressed in order to enable the improvement and usage of ASS algorithms for 3D audio.

While there is a complete lack of data for subjective evaluation of ASS signals in the 3D audio context, there are some approaches for the objective modelling of the perceived quality of ASS signals that might be adapted to the spatial audio context. This thesis intends to bridge the gap between these state-of-the-art approaches and modern 3D audio rendering and therefore proposes some improvements to the ASS evaluation approaches and collects data against which the objective evaluation approaches can be tested.

Outlines and original contributions of the thesis

Chapter 1 of this thesis gives an overview of the state of the art in objective ASS quality evaluation. Different approaches are presented, leading to the largely used subjectively motivated least-squares degradation decomposition. Some improvements in the perspective of spatial audio are finally proposed, resulting in an extended degradation decomposition with one additional feature.

Chapter 2 then provides an overview of the main psychophysical processes that are taken into account by the state-of-the-art auditory models. Such models have been used for a long time in audio quality evaluation and are now being used in ASS quality evaluation. The integration of psychoacoustic knowledge enables to predict subjective data more accurately than other objective techniques. A few models along the history of objective quality evaluation are highlighted, leading to models that are used in today's ASS quality evaluation.

Chapter 3 bridges the gap between chapters 1 and 2 by presenting the state-of-the-art ASS evaluation model. It uses an auditory model presented in chapter 2 to assess the perceptual salience of the different decomposition components presented in chapter 1. The new extended decomposition proposed in chapter 1 is then integrated into this model and results are presented using existing subjective data. The performance of the state-of-the-art model and of its extended version are compared and perspectives drawn concerning its application in 3D audio.

Chapter 4 introduces spatial hearing principles, highlighting some phenomena that are not yet integrated in today's ASS evaluation models. An introduction to binaural spatialisation techniques is also given. A practical formulation for 3D Wave Field Synthesis (WFS) used in some commercially-available systems is then presented as well as a subjective validation of the rendering of source elevation that was conducted as part of this thesis. Results of an experiment of localisation in elevation are analysed to confirm the validity of the height rendering.

In chapter 5, a binaural pre-study on ASS evaluation in a 3D audio context that was conducted for this thesis is presented together with a new algorithm aiming at synthesising degradations linked to ASS. The subjective results of the pre-study are analysed to validate the use of synthetic degradations for further studies. A simplified model for objective ASS evaluation in 3D audio is then proposed based on an existing auditory model that was previously used as part of the model from chapter 3. First sensitivity studies are conducted and the performance of the proposed model is tested against the subjective data gathered during the pre-study to support its further use in follow-up experiments.

Chapter 6 presents a study on subjective ASS evaluation with binaural rendering that was conducted for this thesis. A statistical model is fit to the subjective data that was gathered in order to assess the relative importance of the different degradation components. The performance of the extended state-of-the-art model presented in chapter 3 using the data from this study is then assessed as well as that of the simplified model presented in chapter 5.

The same methodology is used in a 3D WFS environment and the resulting experiment is presented in chapter 7. The differences with the binaural study are highlighted through the use of statistical analysis and the performance of the two ASS evaluation models under test is commented.

In chapter 8 finally, conclusions are drawn about the results of the two presented models in the context of 3D audio. Recommendations are also made for future subjective ASS evaluation studies, based on the insights gained in the two presented studies. Possible continuations of this work are then outlined as a conclusion to this thesis.

1 Objective audio source separation evaluation

1.1 Introduction

Audio Source Separation (ASS) aims at extracting one, several or all "source" signals from a mixture. The associated mathematics, techniques and applications are beyond the scope of this work, but the interested reader may find reviews in the literature [21, 25, 83, 138, 139]. For better understanding of the main topic of this thesis however, an introduction to mixture is given at the beginning of this chapter. Existing objective ASS quality measures are then presented, including the state of the art.
The framework of this thesis is the use of ASS in the context of spatial audio. The last part of this chapter therefore proposes modifications to the state of the art with the context of 3D audio in mind.

1.2 Mixture types

A mixture can be seen as a linear time-invariant system [133]: It is assumed that J source signals s_j are mixed into I mixture signals x_i. Introducing mixing filters $\alpha_{ij}(\tau)$, the mixture signals can be expressed as

$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{+\infty} \alpha_{ij}(\tau) s_j(t-\tau) + n_i(t), \qquad i = 1, ..., I \tag{1.1}$$

where $n_i(t)$ is additive noise in the ith mixture channel (such as for example microphone noise). The term $\sum_{\tau=0}^{+\infty} \alpha_{ij}(\tau) s_j(t-\tau)$ contains all information about spatial position in a stereo mix if applicable, mixing gains and other operations on the source signals and is therefore interpreted as "spatial image" s_{ij}^{img} [132]. This mixture can be expressed in matrix notation, such that:

$$\mathbf{x} = \mathbf{A} * \mathbf{s} + \mathbf{n} \tag{1.2}$$

where $*$ is the convolution operator and bold letters indicate vectors or matrices.
The goal of ASS is then to recover $\mathbf{s}$ or part of it from $\mathbf{x}$ by estimating a demixing system $\mathbf{W}$. The structure of $\mathbf{W}$ depends on the complexity of the approach. In the most simple case, with a noiseless determined system (*i.e.* $\mathbf{n} = \mathbf{0}$ and $I = J$), $\mathbf{W} = \mathbf{A}^{-1}$ and the sources can be recovered as $\mathbf{s} = \mathbf{W}\mathbf{x}$. Most mixtures are however the result of a more complex procedure. Commonly encountered mixture cases can be classified into several categories:

- *Instantaneous mixture.* As stated by Ikram [59], it is assumed, that in that case, the mixed signals x_i at time instant t depend only on the values of the source signals at the same instant t:

  $$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t), \qquad i = 1, ..., I \tag{1.3}$$

  where a_{ij} are mixing coefficients.

- *Convolutive mixture.* In that case, the mixed signals depend on present and past values of the source signals:

  $$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{P-1} h_{ij}(\tau) s_j(t-\tau), \qquad i = 1, ..., I \tag{1.4}$$

  where h_{ij} models the P-point impulse response from source j to receiver i [59].
  In some cases, this is also written with the convolution operator [3]:

  $$x_i(t) = (h_{ij} * s_j)(t), \qquad i = 1, ..., I \tag{1.5}$$

- *Professional mixture.* Mixtures containing not only instantaneous components or convolutive components, but a mix of those or other special effects (*e.g.* chorus, distortion, vocoder, delay, equalisation, compression) are referred to as “professional mixtures”. Most of modern-day published music falls into this category.

The task of ASS can therefore be rather complex and the involved mathematical and statistical models achieve various degrees of performance. This results in degradations in the estimated source signals as compared to the corresponding clean versions before mixing. The next section will provide a state of the art of objective ASS evaluation measures.

1.3 State of the art in objective ASS evaluation

ASS performance measurements are a way to quantify the accuracy of the separation by comparing the extracted version of the source(s) $\hat{\mathbf{s}}$ to the original source(s) $\mathbf{s}$. The first measures were developed for ASS algorithms extracting monophonic signals from potentially multichannel mixtures (as in the situation where one source is recorded by several microphones).

Vincent *et al.* [133] give an overview of such measures, which will be reported here for clarity.

1.3.1 Row Intersymbol Interference (ISI)

If **W** is assumed to be a time-invariant linear demixing system, the estimated sources can be modelled as $\hat{\mathbf{s}} = \mathbf{B} * \mathbf{s}$ where $\mathbf{B} = \mathbf{W} * \mathbf{A}$ is an expression of the inaccuracy of the ASS algorithm. The quality measure is then given by the row ISI as explained by Lambert [78] and cited by Vincent *et al.* [133]:

$$\mathrm{ISI}_j := \frac{\sum_{l,\tau} |B_{jl}(\tau)|^2 - \max_{l,\tau} |B_{jl}(\tau)|^2}{max_{l,\tau}|B_{jl}(\tau)|^2}. \tag{1.6}$$

where the index l $(1 \leq l \leq J)$ relates to the true sources s_l whereas the index j refers to the estimated source $\hat{s}_j$. The advantage of this measure is that it is positive and equal to zero only when the extracted source matches the original source up to a gain and a delay. However, when the demixing system **W** cannot be supposed to be time-invariant linear (which is often the case), other measures are necessary. Additionally, the demixing system has to be known to be able to compute this measure.

1.3.2 L_2-normalised relative square distance

Another approach is to compare the estimated signal $\hat{s}_j$ and the original signal s_l directly, as with the L_2-normalised relative square distance [133, 144] :

$$D := \min_{\epsilon=\pm 1} \left\| \frac{\hat{s}_j}{\|\hat{s}_j\|} - \epsilon \frac{s_j}{\|s_j\|} \right\|^2. \tag{1.7}$$

This has the same properties as the former measure, but it does not allow for a delay. However, the maximum value of D is 2, even if the extracted signal is orthogonal to the original signal. D also does not represent perceptive differences consistently [133].

1.3.3 Orthogonal projection

Both of the previous performance measures are overall quality measures, *i.e.* not taking into account the acceptability of certain types of degradations depending on the signals ASS is applied to. As stated by Vincent *et al.* [133], speech may still be intelligible as a low-pass filtered version and that may not degrade the perceived quality too much. In music however, filtering or musical noise may be perceived as being more annoying than in speech for example. More differentiated criteria have therefore been developed. Vincent *et al.* [133] propose an orthogonal decomposition of the separated signals into a scaled version of the true source signal and different error terms. They state that this concept can be applied to any usual ASS situation given that the decomposition can be adapted to the application.

For music, a decomposition into target distortion s^{target} (a scaled version of the true source), interference of other sources e^{interf}, additional noise e^{noise} (such as sensor noise *e.g.*) and artifacts e^{artif}, which contains all degradations that could not be assigned to the other categories, is proposed. The extracted source is therefore expressed as

$$\hat{s}_j = s^{\text{target}} + e^{\text{interf}} + e^{\text{noise}} + e^{\text{artif}}. \tag{1.8}$$

This can be achieved by orthogonal projection [133]. If $\Pi\{y_1,...,y_k\}$ denotes the orthogonal projector onto the subspace spanned by the vectors $y_1,...,y_k$, 3 projectors can be formulated for the ASS problem:

$$P_{s_j} := \Pi\left\{s_j\right\} \tag{1.9}$$
$$P_{\mathbf{s}} := \Pi\left\{(s_l)_{1\leq l\leq J}\right\} \tag{1.10}$$
$$P_{\mathbf{s},\mathbf{n}} := \Pi\left\{(s_l)_{1\leq l\leq J}, (n_i)_{1\leq i\leq I}\right\}, \tag{1.11}$$

and the components of equation (1.8) can then be expressed as

$$s^{\text{target}} := P_{s_j}\hat{s}_j \tag{1.12}$$
$$e^{\text{interf}} := P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j \tag{1.13}$$
$$e^{\text{noise}} := P_{\mathbf{s},\mathbf{n}}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j \tag{1.14}$$
$$e^{\text{artif}} := \hat{s}_j - P_{\mathbf{s},\mathbf{n}}\hat{s}_j. \tag{1.15}$$

The projection of $\hat{s}_j$ onto s_j can be expressed as a simple inner product $s^{\text{target}} = \langle\hat{s}_j, s_j\rangle s_j/||s_j||^2$. The other projectors might not be as simple to compute. If the sources are mutually orthogonal, $P_{\mathbf{s}}$ can also be expressed as sum of inner products of the estimated source with all the other original sources. If they aren't mutually orthogonal, the calculation of the projection is a bit more complex. A vector $\mathbf{c}$ is then needed such that $P_{\mathbf{s}}\hat{s}_j = \sum_{l=1}^{J}\bar{c}_l s_l = \mathbf{c}^H\mathbf{s}$ where $(\cdot)^H$ is the Hermitian transposition [133]. $\mathbf{c}$ can be computed as $\mathbf{c} = \mathbf{R_{SS}}^{-1}\left[\langle\hat{s}_j, s_1\rangle,...,\langle\hat{s}_j, s_J\rangle\right]^H$ where $\mathbf{R_{SS}}$ is the Gram matrix of the sources and is defined as $(\mathbf{R_{SS}})_{jl} = \langle s_j, s_l\rangle$. The same computations can be made for $P_{\mathbf{s},\mathbf{n}}$, but noise signals can be assumed mutually orthogonal and orthogonal to all sources in most cases, such that $P_{\mathbf{s},\mathbf{n}}\hat{s}_j \approx P_{\mathbf{s}}\hat{s}_j + \sum_{i=1}^{I}\langle\hat{s}_j, n_i\rangle n_i/||n_i||^2$.

Based on this decomposition, Vincent *et al.* [133] define performance criteria by computing energy ratios between the different components: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) Signal to Noise Ratio (SNR) and Signal to Artifacts Ratio (SAR) which are defined as

$$\mathrm{SDR} := 10\log_{10}\frac{||s^{\mathrm{target}}||^2}{||e^{\mathrm{interf}}+e^{\mathrm{noise}}+e^{\mathrm{artif}}||^2} \tag{1.16}$$

$$\mathrm{SIR} := 10\log_{10}\frac{||s^{\mathrm{target}}||^2}{||e^{\mathrm{interf}}||^2} \tag{1.17}$$

$$\mathrm{SNR} := 10\log_{10}\frac{||s^{\mathrm{target}}+e^{\mathrm{interf}}||^2}{||e^{\mathrm{noise}}||^2} \tag{1.18}$$

$$\mathrm{SAR} := 10\log_{10}\frac{||s^{\mathrm{target}}+e^{\mathrm{interf}}+e^{\mathrm{noise}}||^2}{||e^{\mathrm{artif}}||^2}. \tag{1.19}$$

These performance measures can also be defined locally (framewise) by windowing the computed error components before computing the energy ratios to account for performance evolution over time. The variation over time or other statistics may then be computed.
These measures have the advantage of not assuming a particular demixing system. When comparing with the row ISI D, it can be shown that SDR is identical with D up to a logarithmic one-to-one mapping, but SDR has no lower bound. Additionally, the splitting of the error components allows to distinguish between 4 types of estimation errors.

Time-invariant delay

The four measures presented above proposed by Vincent *et al.* [133] are valid under the assumption that the degradations can be expressed by time-invariant gains, which corresponds to the instantaneous mixture case. But as the authors stated, the measures can be adapted to other situations by redefining the orthogonal projectors defined in equations (1.9)-(1.11). When the degradations can be expressed in terms of time-invariant filters for example, s^{target} is not a scaled version of s_j anymore, but a delayed version of it, due to the delay such a filter may introduce [133]. This corresponds more to the convolutive mixture case. The orthogonal projector therefore needs to be reformulated. If τ denotes the delay and s_l^τ and n_i^τ the delayed versions of the source signals and sensor noises respectively, the projectors can be formulated as

$$P_{s_j} := \Pi\left\{(s_j^\tau)_{0\le\tau\le L-1}\right\} \tag{1.20}$$

$$P_{\mathbf{s}} := \Pi\left\{(s_l^\tau)_{1\le l\le J, 0\le\tau\le L-1}\right\} \tag{1.21}$$

$$P_{\mathbf{s},\mathbf{n}} := \Pi\left\{\left\{(s_l^\tau)_{1\le l\le J}, (n_i^\tau)_{1\le i\le I}\right\}_{0\le\tau\le L-1}\right\}. \tag{1.22}$$

Caution has to be taken at the boundaries of the signals, to avoid multiple definitions and the support of the signals has to be reduced to $[0, T+L-2]$ where $[0, T-1]$ is the original length of the signals and $L-1$ is the allowed delay. Vincent *et al.* [133] also state that the Gram matrix corresponding to P_{s_j} is the empirical autocorrelation matrix of s_j defined by $(\mathbf{R}_{s_j s_j})_{\tau\tau'} = \langle s_j^\tau, s_j^{\tau'}\rangle$.

Time-variant gain

If the allowed degradations include time-varying gains, s^{target} is equal to a scaled version of s_j, but the gain slowly varies over time. In terms of mixture, this means that the instantaneous mixture coefficients a_{ij} in equation (1.3) are replaced by coefficients that depend on time $a_{ij}(t)$. For ASS, this gain can be parameterised as $g(t) = \sum_{u=0}^{U-1} \alpha_u v(t-uT')$ where v is a window that defines frames of length L' and with step size T' indexed by u. This allows for piecewise constant gains when v is a rectangular window and $L' = T'$, but a smoother varying gain can be obtained by choosing a smoother window. Under these assumptions, $s^{\text{target}}(t) = g(t)s_j(t) = \sum_{u=0}^{U-1} \alpha_u \times v(t-uT')s_j(t)$ and s^{target} therefore belongs to the subspace spanned by versions of s_j windowed by v. The orthogonal projectors can then be defined as:

$$P_{s_j} := \Pi\left\{(s_j^u)_{0\le u\le U-1}\right\} \tag{1.23}$$

$$P_{\mathbf{s}} := \Pi\left\{(s_l^u)_{1\le l\le J, 0\le u\le U-1}\right\} \tag{1.24}$$

$$P_{\mathbf{s},\mathbf{n}} := \Pi\left\{\left\{(s_l^u)_{1\le l\le J}, (n_i^u)_{1\le i\le I}\right\}_{0\le u\le U-1}\right\}. \tag{1.25}$$

The only condition for the window is that the sum of all windows is constant for all t: $\sum_{u=0}^{U-1} v(t-uT') = C$. This guarantees that the SDR is infinite when $\hat{s}_j = s_j$.

Time-variant gain and delay

For "professional" mixtures, both a time-varying delay and gain will probably occur. If the two previous decompositions should allow for time-varying gains and delays in the degradations (*i.e.* time-varying filters), the projections need to be reformulated. s^{target} is therefore expressed by a scaled and delayed version of s_j (with variations depending on t): $s^{\text{target}} = \sum_{\tau=0}^{L-1} h(\tau,t)s_j(t-\tau) = \sum_{\tau=0}^{L-1}\sum_{u=0}^{U-1} \alpha_{\tau u} \times v(t-uT')s_j(t-\tau)$. s^{target} therefore belongs to the subspace spanned by delayed versions of s_j windowed by v. Windowed delayed source and sensor noise signals $s_j^{\tau u}$ and $n_j^{\tau u}$ can be computed by windowing the delayed signals with v. The projectors can then be reformulated as

$$P_{s_j} := \Pi\left\{(s_j^{\tau u})_{0\le \tau\le L-1, 0\le u\le U-1}\right\} \tag{1.26}$$

$$P_{\mathbf{s}} := \Pi\left\{(s_l^{\tau u})_{1\le l\le J, 0\le \tau\le L-1, 0\le u\le U-1}\right\} \tag{1.27}$$

$$P_{\mathbf{s},\mathbf{n}} := \Pi\left\{\left\{(s_l^{\tau u})_{1\le l\le J}, (n_i^{\tau u})_{1\le i\le I}\right\}_{0\le \tau\le L-1, 0\le u\le U-1}\right\}. \tag{1.28}$$

1.3.4 Multichannel mixtures

As stated in section 1.2, mixtures are not limited to a single channel such as in mono recordings. In most cases in today's applications, sources are in 2 or more channels resulting in stereophonic, or more generally multichannel content. ASS must cope with these cases too and measures were developed for them. The first stereo audio source evaluation campaign [135] used a set of measures derived from the orthogonal projection measures presented above.

The decomposition is reformulated as

$$\hat{s}_{ij}^{\text{img}}(t) = s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t) \tag{1.29}$$

where the index ij refers to the jth source in the ith channel of the mixture. s_{ij}^{img} is the spatial image of a source j in channel i and its estimate is given by $\hat{s}_{ij}^{\text{img}}$. A spatial error term e_{ij}^{spat} is introduced, accounting for the fact that a source may be present in more than one channel and errors may be introduced by ASS, including components of the correct source, but from the wrong channel. These two components correspond to the s^{target} term from the previous measures. The noise error term e^{noise}, present in the previous measures, is however absent here, since it assumes the knowledge of all sensor noise signals, which is impossible in practice. Noise therefore is implicitly included in the e^{artif} artifacts error term.
The error terms can be (re-)formulated as

$$\begin{aligned}
e_{ij}^{\text{spat}}(t) &:= P_{s_j}\hat{s}_{ij}^{\text{img}}(t) - s_{ij}^{\text{img}}(t) && (1.30)\\
e_{ij}^{\text{interf}}(t) &:= P_{\mathbf{s}}\hat{s}_{ij}^{\text{img}}(t) - P_{s_j}\hat{s}_{ij}^{\text{img}}(t) && (1.31)\\
e_{ij}^{\text{artif}}(t) &:= \hat{s}_{j}^{\text{img}} - P_{\mathbf{s}}\hat{s}_{ij}^{\text{img}}(t), && (1.32)
\end{aligned}$$

where the projectors are given by

$$\begin{aligned}
P_{s_j} &:= \Pi\left\{(s_{kj}^{\tau})_{1\le k\le I, 0\le\tau\le L-1}\right\} && (1.33)\\
P_{\mathbf{s}} &:= \Pi\left\{(s_{kl}^{\tau})_{1\le k\le I, 1\le l\le J, 0\le\tau\le L-1}\right\}. && (1.34)
\end{aligned}$$

P_{s_j} corresponds to the projection on the subspace spanned by the true source image of source s_j in all channels, whereas $P_{\mathbf{s}}$ corresponds to the projection on the subspace spanned by all true source images of all sources in all channels.

According to the new decomposition, the performance criteria are adapted as follows, with the introduction of Source Image to Spatial Distortion Ratio (ISR) to account for the spatial errors:

$$\text{ISR}_j := 10\log_{10}\frac{\sum_{i=1}^{I}||s_{ij}^{\text{img}}||^2}{\sum_{i=1}^{I}||e_{ij}^{\text{spat}}||^2} \tag{1.35}$$

$$\text{SIR}_j := 10\log_{10}\frac{\sum_{i=1}^{I}||s_{ij}^{\text{img}} + e_{ij}^{\text{spat}}||^2}{\sum_{i=1}^{I}||e_{ij}^{\text{interf}}||^2} \tag{1.36}$$

$$\text{SAR}_j := 10\log_{10}\frac{\sum_{i=1}^{I}||s_{ij}^{\text{img}} + e_{ij}^{\text{spat}} + e_{ij}^{\text{interf}}||^2}{\sum_{i=1}^{I}||e_{ij}^{\text{artif}}||^2} \tag{1.37}$$

$$\text{SDR}_j := 10\log_{10}\frac{\sum_{i=1}^{I}||s_{ij}^{\text{img}}||^2}{\sum_{i=1}^{I}||e_{ij}^{\text{spat}} + e_{ij}^{\text{interf}} + e_{ij}^{\text{artif}}||^2} \tag{1.38}$$

Even though these measures are largely used in the ASS community [4, 5, 23, 132, 135], the

authors explicitly mention some flaws. For example, the distortion components seem to not always correspond to those perceived by human listeners. The authors attribute it to the fact that the time-varying projectors are computationally expensive and originally could only be implemented in low time resolution [133]. Additionally, the Finite Impulse Response (FIR) filter that was used has a constant frequency resolution, that does not match that of the ear [42]. The time-varying measures were therefore discarded since they did not improve the results.
Moreover, Emiya *et al.* [42] state that with the implementation of the proposed decomposition algorithm, the target distortion component can be nonzero - even when the target is not distorted - if the sources are correlated.
To tackle these points and especially to give more perceptual relevance to the decomposition, a follow-up measure has been proposed under the form of Perceptive Evaluation Methods for Audio Source Separation (PEASS) [42]. PEASS includes two stages:

1. perceptually-motivated least-squares signal decomposition

2. component-based objective measures with an auditory model

While the first stage is in the scope of this chapter, the second stage will be discussed in the next chapter.

1.3.5 Perceptually-motivated least-squares projection

To eliminate the flaws of the presented decomposition algorithms, Emiya *et al.* presented a perceptually-motivated approach to the signal decomposition [42]. This first stage of a framework called PEASS includes three steps.

Step 1: Time-frequency analysis

The estimated source $\hat{s}_{ij}(t)$ and the true source signals s_{kl} are filtered by a fourth-order gammatone filter bank. The implementation uses the proposal given by Hohmann [54] and Herzke & Hohmann [53]. The center frequencies of the resulting signals $\hat{s}_{ijb}$ and s_{klb}, where b indexes the filter band, are linearly distributed on the Equivalent Rectangular Bandwidth (ERB) scale between 20 Hz and the Nyquist frequency. For an explanation about the ERB scale and other similar scales, refer to section 2.5.1. 3 filters per ERB are used and the signals are consequently decimated by a factor depending on the center frequency of the filter.

In each sub-band, delayed versions of the true source signals s_{klb} are then computed, according to the “time-variant delay” approach cited above. The estimated source signal $\hat{s}_{ijb}(t)$ and the delayed true source signals $s_{klb}(t-\tau)$, $1 \le k \le I$, $1 \le l \le J$, $1 \le b \le B$, $-L/2 \le \tau \le L/2$ are then partitioned into time frames indexed by u with an analysis window w_a with step size N:

$$\hat{s}_{ijbu}(t) = w_a(t)\hat{s}_{ijb}(t-uN) \tag{1.39}$$

$$s^{\tau}_{klbu} = w_a(t)s_{klb}(t-uN-\tau). \tag{1.40}$$

The authors of the decomposition suggest a sine windows with 75% overlap. They also note that because of the decimation process, this decomposition will result in variable time resolution, depending on the center frequency of filter band b.

Step 2: Joint Least-Squares Decomposition

Since the bandwidth of the employed gammatone filters is quite wide, the degradation components are estimated by applying an additional FIR time-invariant filter bank to every signal in each sub-band and each time frame. The residual is taken as e^{artif}:

$$e^{\text{target}}_{ijbu}(t) = \sum_{k=1}^{I}\sum_{\tau=-L/2}^{L/2}\alpha_{ijbu,kj}(\tau)s^{\tau}_{kjbu}(t) \tag{1.41}$$

$$e^{\text{interf}}_{ijbu}(t) = \sum_{k=1}^{I}\sum_{l\neq j}\sum_{\tau=-L/2}^{L/2}\alpha_{ijbu,kl}(\tau)s^{\tau}_{klbu}(t) \tag{1.42}$$

$$e^{\text{artif}}_{ijbu}(t) = \hat{s}_{ijbu}(t) - s^{0}_{ijbu}(t) - e^{\text{target}}_{ijbu}(t) - e^{\text{interf}}_{ijbu}(t). \tag{1.43}$$

The filter coefficients α_{ijbu} are computed by least-squares projection of the degradation signal $\hat{s}_{ijbu}(t) - s^{0}_{ijbu}(t)$ onto the subspace spanned by the delayed true source signals. The optimal estimator for the filter coefficients is given by $\alpha_{ijbu} = \mathbf{S}^{+}_{bu}(\hat{\mathbf{s}}_{ijbu} - \mathbf{s}_{ijbu})$ where $\mathbf{S}_{bu}$ is the matrix containing the delayed true source signals and $(\cdot)^{+}$ denotes matrix pseudo-inversion.

Step 3: Time-Domain Resynthesis

The degradation components are recomposed in each sub-band from the filtered signals using Overlap-and-Add (OLA) [101] with a synthesis window w_s such that $\sum_u w_s(t-uN)w_a(t-uN) = 1$. This can be written as:

$$e^{\text{target}}_{ijb}(t) = \sum_{u} w_s(t-uN)e^{\text{target}}_{ijbu}(t-uN) \tag{1.44}$$

$$e^{\text{interf}}_{ijb}(t) = \sum_{u} w_s(t-uN)e^{\text{interf}}_{ijbu}(t-uN) \tag{1.45}$$

$$e^{\text{artif}}_{ijb}(t) = \sum_{u} w_s(t-uN)e^{\text{artif}}_{ijbu}(t-uN). \tag{1.46}$$

Finally, the fullband signals are reconstructed by filter bank inversion of the gammatone filter bank [54]. The fullband estimated and true source signals $\hat{s}_{ij}$ and s_{ij} are also reconstructed from their gammatone-filtered version, in order to be able to account for inaudible distortions due to filter bank inversion. Those signals are then used for further processing.

Just like for the previous decompositions, energy ratios can then be computed as quality criteria:

$$\mathrm{SDR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)|^2}{\sum_i\sum_t |\hat{s}_{ij}(t) - s_{ij}(t)|^2} \tag{1.47}$$

$$\mathrm{ISR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{target}}(t)|^2} \tag{1.48}$$

$$\mathrm{SIR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t) + e_{ij}^{\text{target}}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{interf}}(t)|^2} \tag{1.49}$$

$$\mathrm{SAR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t) + e_{ij}^{\text{target}}(t) + e_{ij}^{\text{interf}}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{artif}}(t)|^2} \tag{1.50}$$

Note that the target distortion term is now called e_{ij}^{target} to better represent what it stands for. These measures are practically the same as before [135]. However, there is one little difference: The overall distortion term SDR does not rely on the decomposition anymore, but it is now defined as a ratio between the true source image and the difference between the estimate and the true source image.

1.4 Proposed improvements

With the use of ASS signals in 3D rendering scenarios, some degradations that are inaudible when listening in stereophonic contexts may become audible. Additionally, certain degradations may become more important in terms of perceived quality in the context of spatial audio. An informal listening session involving experts from the ASS and spatial audio fields was held in the framework of the i3Dmusic project to assess the perceived degradations in signals that had undergone source separation with state-of-the-art algorithms. The extracted sources were spatialised with a 3D Wave Field Synthesis (WFS) system covering the upper hemisphere at Sonic Emotion labs in Paris. For details about the implemented 3D WFS algorithm, see section 4.4. For the chosen musical excerpts, one source was extracted using state-of-the-art ASS algorithms as described in [118] *e.g.*. Most often, the lead voice was chosen as target. The extracted source was then spatialised in front of the listeners and the residue was spatialised in the back. The listeners then had to take note of what types of degradations they perceived. The outcome of the listening session were 4 families of artifacts:

- static interference
- onset misallocation
- target distortion
- musical noise-type artifacts

Even though those categories have quite some common points with the ones defined in section 1.3.3, there are some differences that will be explained in the following paragraphs.

The interference degradation presented previously is split into two types of interference to define the two proposed onset misallocation and static interference categories. Onset misallocation defines dynamic interference from other sources that is generally occurring at the onsets of some sources (either of the target or of other sources). This happens for example when the target source has an onset and during that short time period where the source separation algorithm has to re-estimate the parameters of the target source, other sources may be interfering and be extracted as part of the target. It may also happen when other sources have similar parameters to the target source for a short duration in time at a given position in a song.
Static interferences on the other hand concern interference from other sources that is less variable in time and therefore better corresponds to the interference degradation specified above.
The musical noise-type artifacts is defined to be the same as the former artifacts degradation category, *i.e.* the residual of the orthogonal projection. The only change occurs when synthesising this type of degradation as explained in section 5.2.
The target distortion degradation stays the same as already proposed by Emiya *et al.* [42], *i.e.* a scaled version of the target source.
In terms of orthogonal projection, this new definition of the degradation families calls for some adjustments to the projection algorithm. The state-of-the-art projection does not allow for a distinction between static interference and onset misallocation. All interference from other sources is included in e_{ij}^{interf}. Since the conclusions of the informal listening session suggested that onset misallocation degradations may be more disturbing than static interference in the context of spatial audio, due to the fact that they imitate sources moving from the location of the target to that of the residue, a modification of the projection algorithm is proposed.
The method to obtain the target distortion degradation e^{target} stays the same:

$$e_{ijbu}^{\text{target}}(t) = \sum_{k=1}^{I} \sum_{\tau=-L/2}^{L/2} \alpha_{ijbu,kj}(\tau) s_{kjbu}^{\tau}(t) \tag{1.51}$$

For the static interference component e^{interf} however, the filter is forced to be time-invariant (*i.e.* to not change throughout the whole temporal duration of the signal) by using the whole temporal signals for the estimation of the filter coefficients:

$$e_{ijb}^{\text{interf}}(t) = \sum_{k=1}^{I} \sum_{l\neq j} \sum_{\tau=-L/2}^{L/2} \alpha_{ijb,kl}(\tau) s_{klb}^{\tau}(t) \tag{1.52}$$

where $\alpha_{ijb} = \mathbf{S}_b^{+}(\hat{\mathbf{s}}_{ijb} - \mathbf{s}_{ijb})$.
For the onset misallocation component e^{onset}, time-variant filters are used again through the

use of the windowed signals:

$$e^{\text{onset}}_{ijbu}(t) = \sum_{k=1}^{I} \sum_{l \neq j} \sum_{\tau=-L/2}^{L/2} \alpha_{ijbu,kl}(\tau) s^{\tau}_{klbu}(t) \tag{1.53}$$

with filter coefficients estimated as $\alpha_{ijbu} = \mathbf{S}^{+}_{bu}(\hat{\mathbf{s}}_{ijbu} - \mathbf{s}_{ijbu} - \mathbf{e}^{\text{interf}}_{ijbu})$.
The artifacts degradation component can then be computed as

$$e^{\text{artif}}_{ijb}(t) = \hat{s}_{ijb}(t) - s^{0}_{ijb}(t) - e^{\text{target}}_{ijb}(t) - e^{\text{interf}}_{ijb}(t) - e^{\text{onset}}_{ijb}(t). \tag{1.54}$$

Note that time-domain re-synthesis of the interference and target degradation components has to be done before computing the artifacts degradation component.
An energy ratio can be computed for the onset degradation component just as for the state-of-the-art decomposition. The new Signal to Onset Ratio (SOR) is defined and the other ratios modified as follows:

$$\text{SDR}_j = 10\log_{10} \frac{\sum_i \sum_t |s_{ij}(t)|^2}{\sum_i \sum_t |\hat{s}_{ij}(t) - s_{ij}(t)|^2} \tag{1.55}$$

$$\text{ISR}_j = 10\log_{10} \frac{\sum_i \sum_t |s_{ij}(t)|^2}{\sum_i \sum_t |e^{\text{target}}_{ij}(t)|^2} \tag{1.56}$$

$$\text{SIR}_j = 10\log_{10} \frac{\sum_i \sum_t |s_{ij}(t) + e^{\text{target}}_{ij}(t)|^2}{\sum_i \sum_t |e^{\text{interf}}_{ij}(t)|^2} \tag{1.57}$$

$$\text{SOR}_j = 10\log_{10} \frac{\sum_i \sum_t |s_{ij}(t) + e^{\text{target}}_{ij}(t) + e^{\text{interf}}_{ij}(t)|^2}{\sum_i \sum_t |e^{\text{onset}}_{ij}(t)|^2} \tag{1.58}$$

$$\text{SAR}_j = 10\log_{10} \frac{\sum_i \sum_t |s_{ij}(t) + e^{\text{target}}_{ij}(t) + e^{\text{interf}}_{ij}(t) + e^{\text{onset}}_{ij}(t)|^2}{\sum_i \sum_t |e^{\text{artif}}_{ij}(t)|^2} \tag{1.59}$$

This proposed projection algorithm allows to separate static and dynamic interference through the use of time-invariant and time-variant filters separately.

1.5 Conclusion

The set of existing ASS performance measures is quite small, especially in the context of multichannel audio. Row ISI assumes a known time-invariant linear demixing system, which often is not the case in practice. The L_2-normalised relative square distance has the advantage of directly comparing the original signal and the estimate but has severe design flaws (*e.g.* a limited maximum value), that make it impractical for comparison between several extraction algorithms.
These drawbacks have been addressed by Vincent *et al.* [135] by introducing measures that rely on orthogonal projection of the signals onto subspaces spanned by different versions of the original source signals degraded by a delay and/or a gain. But even when taking into account

auditory-motivated characteristics, such as nonuniform filter bandwidth and nonuniform time-resolution across frequency, the energy ratios still did not always fit the perceptual salience of all components in the estimates [42]. Low-frequency components for example affect energy ratios quite largely, whereas this might not be perceived that way. Masking of low degradations by louder components is also not taken into account.

This leads to the second part of the proposal by Emiya *et al.* [42], where an auditory model is used to further assess the salience of the degradations components extracted through orthogonal projection. While this is presented in chapter 3, chapter 2 will provide an introductory review of auditory models to get a better understanding for the involved processes.
Chapter 3 will also suggest a use for the modification to the state-of-the-art decomposition that is proposed at the end of this chapter. In the context of spatial audio, an additional degradation component is proposed with an accompanying decomposition algorithm, separating static from dynamic interference from other sources.

2 Auditory modelling of subjective audio quality

2.1 Introduction

While the objective measures presented in chapter 1 might prove to be useful from a signal processing point of view, they lack a link to the subjective impression of listeners. In all use cases, a human listener will be at the end of the signal processing chain; whether it is an artist, creatively using the results of the Audio Source Separation (ASS) algorithms, or a sound engineer, trying to provide a quality product, or an end user, listening to and enjoying the derived products. In every case, the opinion of the listener is the last and most important measure of quality.

Auditory models have been developed to predict the quality of given audio excerpts. They aim at representing the behavior of human audition in the most accurate way possible, modelling different parts of auditory processing to accomplish an audio evaluation task just like a human subject would. An alternative is given by mathematically motivated models that don't make any assumptions about involved psychoacoustic processes (or only very few) and try to establish a link between given signals and the associated subjective scores via machine learning algorithms for example. In practice, most auditory models use some kind of mathematical model to map their outputs to subjective scores since the prediction is never perfect, due to their limited knowledge about the involved processes.

Without such models, the only way of persistently judging the quality of audio signals is to conduct subjective listening tests, involving a possibly large panel of listeners and extensive statistical analysis of the data that is generated. In the words of Brandenburg: "[...] extensive tests with statistical evaluation have to be done to decide whether there is an audible difference or not" [16]. This is expensive and time-consuming. Having a model of human hearing that is able to predict the outcome of such a listening test series is much more convenient. Researchers have therefore looked for a way to establish such a model and have come up with solutions fitted to specific tasks or more general approaches. A review of the involved psychoacoustic processes and how they are modelled as well as a review of auditory models for audio quality evaluation that integrate some of these processes is given in this chapter.

A general statement about the inputs to auditory models is first made. The anatomy of the ear is then linked to different well-known psychoacoustic phenomena such as masking and adaptation *e.g.*. Modelling approaches from the literature are reported for the different phenomena. In the last part, auditory models are presented that lead up to the current state of the art in ASS evaluation.

2.2 Model inputs

The first steps towards audio quality evaluation models were made when digital audio coding became more and more important in the late 1980s and the early 1990s, arising from the need of better metrics than just Signal to Noise Ratio (SNR). In audio coding, the quality reference is always given by the original uncoded signal. As will be explained in section 5.3, state-of-the-art listening tests involving ASS quality evaluation also include the original signal to give the listener a reference as to what optimal quality would be. It is therefore reasonable to assume that a reference signal is also available to an auditory model. In fact, all the models presented in the following sections use a reference (clean) signal and a processed version of the same signal as inputs. Generally, any delay and gain difference of the processed version is compensated before further processing. This prevents the model from evaluating these gains and delays as distortions when they really just are linear characteristics of the system under test and aren't evaluated as distortions by listeners (see Huber & Kollmeier's model for example [58]).

2.3 The outer and middle ear

The ear is at the core of human auditory perception. It plays the role of a transducer, "translating" mechanical vibrations of matter to electro-chemical responses that are interpretable to the brain. Even though not all processes involved are fully understood by the time this is written (especially on the neuronal processing side), the ear can be roughly divided into 3 sections:

- the outer ear
- the middle ear
- the inner ear

All of those parts play distinct roles in the hearing process and they are modelled with more or less effort.

The outer ear is composed of the pinna (the "visible" part of the ear) and the auditory canal (depicted on figure 2.1). Its main role is to allow sound to travel towards the middle ear whilst physically protecting the more sensitive parts such as the tympanic membrane.
Its structure - especially that of the pinna - also plays a major role in human sound source

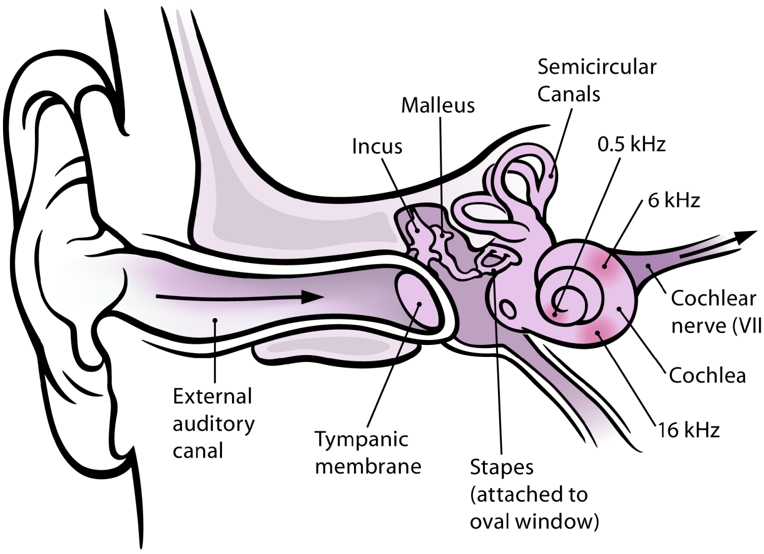


Figure 2.1 – Anatomy of the human ear, as depicted by Chittka & Brockmann [20]. The length of the auditory canal is exaggerated for viewing purposes. Source: Chittka & Brockmann [20].

localisation. Inbound sound waves are diffracted by the torso, the head and the pinnae, altering the sound pressure when comparing to the situation without a listener [108]. This spatial component will be addressed in more depth in section 4.2.

The auditory canal can be modelled as a tube of approximately 27 mm length and a cross-section of 40 mm^2 [108]. It can be approximated by a quarter-wave resonator with one open and one closed termination. As such, it presents a series of resonances at different frequencies, the first one occurring at about 3200 Hz. These resonances shape the sensitivity of the outer ear, since they correspond to frequency bands for which the amplitude of the sound pressure at the tympanic membrane is bigger than at the input of the auditory canal.

The middle ear is composed of the tympanic membrane, three little bones called incus, malleus and stapes situated in the tympanic cavity that is filled with air and connected to the nose through the Eustachian tube, and the oval window (all depicted on figure 2.1). It makes the interface between aerial acoustics (external ear) and the inner ear in which vibrations are transmitted through liquid. The arrangement of muscles and bones in the middle ear optimizes transmission between the outer and the inner ear (impedance matching). Due to its mechanical properties, the resistance to movement of the inner ear is very different from that of air. The middle ear acts as a transformer that improves sound transmission and minimizes sound reflection. Optimal transmission is achieved between 500 and 4’000 Hz [86].
It also serves as additional protection against potentially harmfully loud sounds through what is known as the “acoustic reflex” (or stapedius reflex). The muscles attached to the tympanic membrane and the stapes stiffen the transmission chain of the middle ear and therefore attenuate the amplitude of transmitted vibrations. This is possible up to certain limits, which is the reason why highly energetic sounds can still permanently damage the inner ear. At low

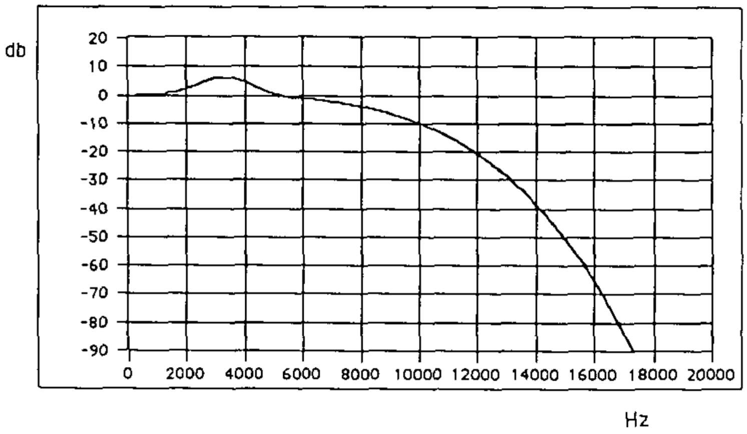


Figure 2.2 – Transfer function between the outer and the inner ear as used in different models. Source: Paillard *et al.* [97].

frequencies, the stiffness of the transmission chain is attenuating the amplitude of sound even when presented with noncritical levels and there is a decreased sensitivity of the human ear at low frequencies because of that [108].
All of these characteristics can be summarised in a low-pass transfer function as illustrated by figure 2.2 [97]. Thiede *et al.* [128] approximate this curve as

$$A(f) = -0.6 \cdot 3.64 f^{-0.8} + 6.5 e^{-0.6(f-3.3)^2} - 10^{-3} f^4, \tag{2.1}$$

with f the frequency in kHz. Most of the auditory models that take into account the influence of the outer and middle ear use such a transfer function [7, 24, 97, 120, 126–128]. Schroeder *et al.* [113] only model the effect of the stapes transmitting the velocity of the eardrum to the inner ear and they therefore use another transfer function. A simpler approach is taken by Münkner & Püschel [89], where two high-pass filters with 200 Hz and 1000 Hz cut-off frequencies are used. Karjalainen [71] directly integrates the frequency-dependent sensitivity of the ear into the frequency decomposition of the inner ear by allowing different gains for different frequency bands.

2.4 Inner ear

The inner ear is essentially composed of the cochlea (depicted on figure 2.1), which translates vibrations of the oval window to electro-chemical signals that are transmitted to the brain via the cochlear nerve (also on figure 2.1).

The cochlea's external bony walls are rigid and vibrations are transmitted to the cochlea via the oval window. While this is the "entry", the cochlea in itself is a rather complex hydrodynamic system that consists of a spiralling bone tube (2.5 turns over 3.5 cm) and several internal organs. Two membranes divide it into three chambers: the scala vestibuli and the scala

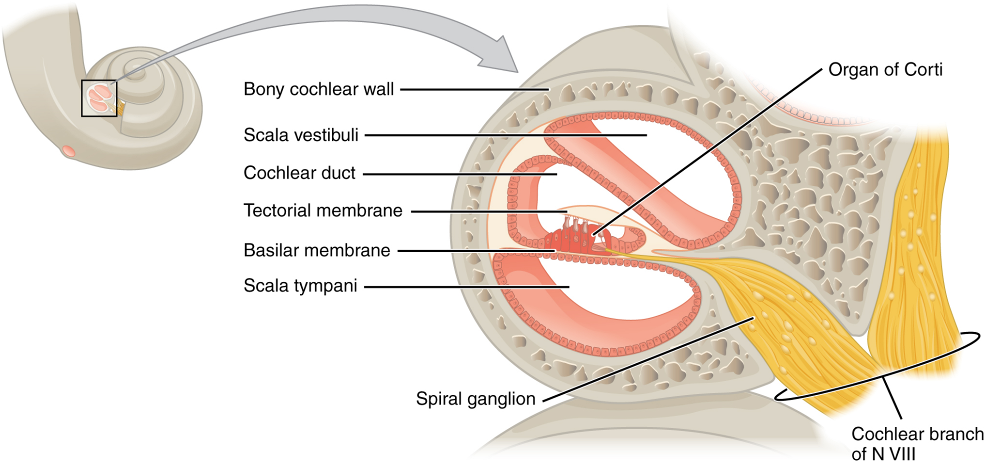


Figure 2.3 – Anatomy of the cochlea. Source: OpenStax College [95]

media, separated by the Reissner's membrane; and the scala tympani, separated from the scala media by the basilar membrane (see figure 2.3). The membrane of the oval window transmits incoming sound pressure waves to the fluids filling the cochlea, which are able to move because the other end of the cochlea is formed by another elastic membrane in the round window. This results in a pressure difference across the basilar membrane.

The organ of Corti, that supports the stereocilia (sensory hair cells) is situated on top of the basilar membrane. Those elements are at the heart of the cited transduction process. Stereocilia can be separated into external (3 rows nearer to the outer rim of the cochlea) and internal (1 row near the axis of the cochlea) categories [104]. The internal stereocilia translate the mechanical vibrations of the basilar membrane to chemical variations of their potassium concentration, which generates electro-chemical impulses in the cochlear nerve. The external stereocilia play an active role of amplification of the local displacement of the basilar membrane [108], allowing for the high dynamic range of the human auditory process. Due to spatial variations of the mechanical properties of the basilar membrane (width and thickness), different locations along the basilar membrane are sensitive to different frequency ranges as demonstrated by von Békésy [137]. Lower frequencies reach maximum amplitude further away from the oval window than higher frequencies [108]. The exact mechanical reasons for this frequency analysing behavior are still debated, but it can be noted that the inner ear decomposes an incoming acoustic signal as a function of frequency. This mapping and the fact that it is not linear will be explained in the following section.

The resulting electrical signals transmitted through the cochlear nerve are then processed by the brain. Many aspects of how this information is processed are still under investigation, but more and more insight is gained about the mechanics of auditory localisation, scene analysis, stream segregation and further processing such as speech recognition etc.

2.5 From the time domain to the frequency domain

As explained in the previous section, the cochlea performs a frequency-to-location mapping. Most psychoacoustic phenomena are therefore frequency-dependent which is why most models include a frequency decomposition as a first processing stage.

Two types of approaches can be distinguished: a Fast Fourier Transform (FFT)-based decomposition or a decomposition through a filter bank. A FFT-based decomposition has the advantage of being very efficient computationally, thanks to very well known algorithms. The disadvantage of such an approach lies in its constant time and frequency resolution. The frequency-to-location mapping process in the cochlea does not have a constant frequency resolution, nor does it have a constant temporal resolution over all frequency bands. The alternative filter bank-based decomposition yields the advantage of being able to model variable frequency resolution through the use of different filters for different frequency bands and frequency-dependent temporal resolution through the use of subsampling. However, the digital implementation may not be as efficient as that of an FFT.

2.5.1 Critical bands

The first concept that needs to be introduced in this context is that of critical bands. Critical bands have been defined by Zwicker in 1961 [146] based on the work of Fletcher [45], to account for the behavior of several psychoacoustic phenomena, such as masking *nota bene* (see section 2.8). Related to the way frequency is mapped to location on the basilar membrane, a critical band describes the bandwidth within which a secondary tone can interfere with the perception of a primary tone. Their bandwidth increases with frequency and even though their center frequency is not fixed *per se*, but has rather been found to be adapting to the content, Zwicker [146] proposed 24 standardised center frequencies for auditory filters that were adopted as a standard by the International Organisation for Standardisation (ISO), leading to the widely used Bark scale (also known as the critical-band scale or the *z* scale) that covers the entire audible spectrum. Alternatively, one could choose to use the mel scale [124], relating perceived pitch to frequency. One Bark is approximately equivalent to 100 mel [146]. The Equivalent Rectangular Bandwidth (ERB) scale [47, 87] is also based on interference (or masking) experiments but it uses rectangular band-pass filters that are computationally more convenient. A comparison between the 3 scales in addition to the linear Hz scale is shown on figure 2.4.

Note that critical bands are often approximated by third-octave bands, since they are more broadly available for measuring instruments (see the loudness calculation presented in section 2.8.2 *e.g.*). A third-octave band is defined as a frequency interval between f_1 and f_2 such that $f_2 = f_1 \times 2^{1/3}$. Center frequencies of third-octave bands have been standardised in ISO 266 [61].

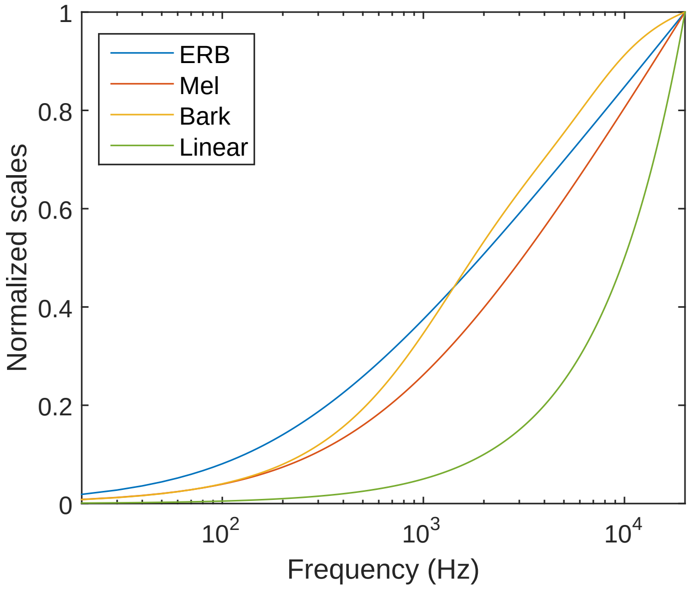


Figure 2.4 – Comparison between Mel, Bark, ERB and linear frequency scales. Since the units of the different scales are different, units are normalized.

2.5.2 FFT-based decomposition

As explained above, an FFT-based decomposition in an auditory model provides computational advantages. However, many higher-level psychoacoustic phenomena do not depend on frequency in a linear manner, but rather behave according to critical bands as presented above. Most auditory models using a short-time frequency decomposition such as the FFT or the Discrete Cosine Transform (DCT) therefore use some sort of conversion to the critical-band scale *e.g.* by grouping and averaging of the frequency components:

Soumagne *et al.* [119] use a FFT of unspecified length to compute the Frequential Signal-to-Noise Ratio (FSNR) without any psychoacoustic grouping. A short-time 20 ms sliding window is used by Schroeder *et al.* [113] to imitate characteristics of human speech perception. The frequency spectra are then mapped to critical-band densities with a mapping function $f = 650\sinh(x/7)$ where x is an approximation of the z-scale unit.

Brandenburg [16] uses a 32 ms window for the frequency decomposition. There is no explicit grouping in Brandenburg's model, but the follow-up calculations are based on a graphic procedure, where third-octave levels are implicitly mapped to a perceptual scale. The Perceptual Audio Quality Measure (PAQM) model by Beerends & Stemerdink [7] uses a 40 ms Hanning window and a mapping to the Bark scale. Colomes *et al.* [24] also use a 40 ms window and the frequency components are mapped to 620 channels on the Bark scale. The Perceptual Evaluation of Audio Quality (PEAQ) model [128] uses a FFT with 21 ms 50% overlapping Hanning windows and frequency bands are grouped by bands of 0.25 or 0.5 Bark, depending on the version of the model. For the Perceptual Evaluation of the Quality of Audio Signals (PERCEVAL) model by Paillard *et al.* [97], a Modulated Lapped Transform (MLT) is used to get

23 ms time resolution and 1024 frequency components. The MLT is similar to the DCT with overlapping windows.

2.5.3 Filter bank-based decomposition

The disadvantage of a FFT-based decomposition is its constant frequency and time resolution. The constant frequency resolution is contradictory to the frequency-to-location mapping on the basilar membrane. Since hair cells are regularly spaced but the mapping is not linear, the frequency resolution of the ear is constant on the Bark scale but not on the Hz scale. Also, since the frequency resolution of critical bands is not constant on the Hz scale, their time resolution can't be constant either. This is why certain auditory models choose a filter bank-based approach rather than a FFT.

Most models that use a filter bank for frequency decomposition choose to space filters (*i.e.* their center frequencies) equally on the Bark scale or an approximation thereof. Frequency resolution and implementations however vary: Karjalainen [71] uses 48 bandpass filters (256th-order Finite Impulse Response (FIR) filters) with a 0.5 Bark spacing. This is implemented as a matrix multiplication, meaning that the output of every channel is based on the same number of temporal samples.
Kapust [70] introduces the so-called BARK-Transform that can be seen as a set of 23 cascaded band-pass filters. They have a bandwidth of 1 critical band and they present a 27 dB/Bark steepness offband.
Stuart [126] uses a 380-band custom rounded exponential (Roex) filter bank in his auditory model "noise". The definition of Roex filters is based on psychoacoustic experiments, where the detection threshold of a tone in notched noise (broadband noise with a spectral notch around the tone) was evaluated [98]. The resulting filter shape can be approximated by two back-to-back decaying exponentials with some kind of rounding at their junction. Thiede & Kabot [127] also use a filter bank of unspecified resolution with center frequencies distributed evenly across the critical band scale and bandwidths varying accordingly.
In the Objective Audio Signal Evaluation (OASE) model by Sporer [120], a 241-band filter bank is used with center frequencies equally spaced on the Bark scale, resulting in a resolution of 0.1 Bark. Filter shapes depend on Sound Pressure Level (SPL), resulting in different offband attenuation slopes for different SPLs. Subsampling by powers of 2 is used whenever possible for the lower center frequencies to reduce computing power needs.
Hollier *et al.* [55] use a 19-band filter bank with a 1 Bark spacing using center frequencies in the range between 100 Hz and 8 kHz following the filter definition by Sekey & Hanson [115].
The advanced version of PEAQ [128] uses a linear phase 40-band filter bank that models the characteristics of the basilar membrane. Filters have regularly spaced center frequencies and constant bandwidths on an approximation of the Bark scale (between 500 Hz and 18 kHz) [65]. Filter outputs are subsampled by a factor 32.
Dau *et al.* [35] as well as Sander [110] use a 120-band filter bank with wave-digital-filters [125] resulting in a 0.2 Bark resolution.

In Münkner & Püschel's model [89], a gammatone filter bank with an unspecified number of channels is used. It is only stated that the channel with 1.5 kHz center frequency has a bandwidth of 170 Hz. Gammatone filter banks are used as an alternative to Roex filter banks, since they presents basically the same characteristics with improved computational efficiency [98, 99]. In the Perceptual Model of Audio Quality (PEMO-Q) model finally [58], a fourth-order gammatone filter bank is used with 35 bands. Gammatone filter banks are cited to have high frequency selectivity in at least some cases [110].

2.6 Rectification processing

The stereocilia translate vibrations of the basilar membrane to electro-chemical pulses through variations of their potassium concentration. In a most simple approach, one could say that the more movement, the higher the variation rate of potassium concentration. This implies a rectification process that can be translated to half-wave rectification in signal processing terms [2]. In psychoacoustic terms, rectification stands for the fact that a tone is always perceived as a constant auditory event [128]. Most auditory models include some type of rectification at one stage or another. Some include half-wave rectification with or without low-pass filtering (which models conservation of phase at low frequencies versus envelope extraction at higher frequencies) [35, 58, 89, 110], others include square-law elements [7, 71, 97, 120, 127, 128] to extract energy measures.

2.7 Logarithmic compression

The human ear is sensitive to SPL in a logarithmic fashion, hence the decibel scale to measure SPL. Some auditory models account for this by integrating some kind of logarithmic compression at one stage or another. In other models, the logarithmic compression is implied through the use of the decibel scale at the input or the output of the model. Karjalainen [71] or Paillard *et al.* [97] implement it at the end of the auditory model, whereas Colomes *et al.* [24] implement logarithmic compression to compute the excitation patterns at the output.
The PAQM model [7] does not use a logarithmic compression *per se* but the mapping from excitation E to loudness $\mathscr{L}$ is given by the following equation:

$$\mathscr{L} = k\left(\frac{E_0}{s}\right)^{\gamma}\left[\left(1 - s + s\frac{E}{E_0}\right)^{\gamma} - 1\right] \tag{2.2}$$

where s is the "schwell" factor as defined by Feldtkeller & Zwicker [44] and E_0 is the excitation at the absolute hearing threshold. The value of the exponent γ was found to be 0.23 by Zwicker who originally proposed the formula. In the PAQM model [7], its value was adjusted to optimize correlation between the model outputs and the subjective data of some database. Since $\gamma < 1$, this represents some form of a compression not dissimilar to logarithmic compression in other models. A very similar approach with a slightly different formula is taken in the model by Thiede & Kabot [127] as well as in the PEAQ model [128], whereas Stuart [126] implements

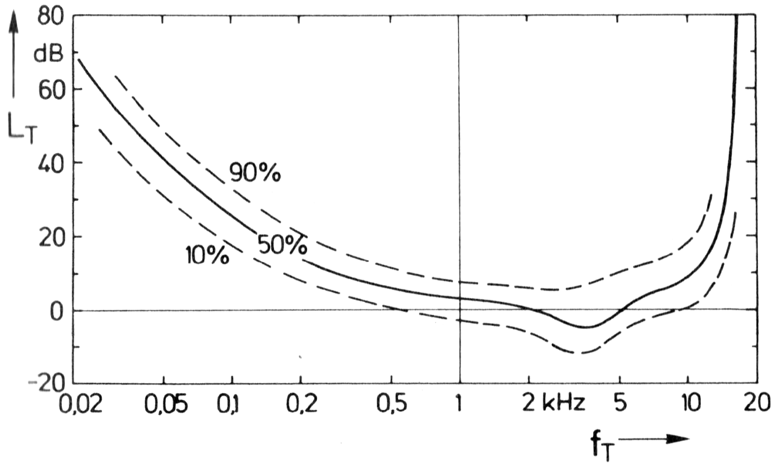


Figure 2.5 – Absolute hearing threshold L_T as a function of frequency f_T from Zwicker [147]. The solid curve corresponds to the median hearing threshold for subjects less than 25 years old and the two dashed lines correspond to the 10th and the 90th percentile. Source: Zwicker [147].

his own loudness compression.
While some models [35, 58, 89, 110] do not implement an explicit logarithmic compression stage, they implement adaptation loops (see section 2.9) to compress signals that don't vary too fast over time in a near-logarithmic way.

2.8 Masking

Masking is defined as a process whereby the threshold of audibility of one sound event (the maskee) is raised by another sound event (the masker) [14]. The amount of masking depends on the temporal position of the masker relative to the maskee (simultaneous masking when the masker and the maskee are present at the same time, temporal masking if they aren't) as well as their respective energy levels and spectra. The absolute hearing threshold also plays a role in the masking process.

2.8.1 Absolute hearing threshold

The absolute hearing threshold is defined as the sound pressure level of a sine tone, depending on its frequency, such that the tone is just noticeable in silence. It can be measured by means of psychoacoustic experiments and depends notably on the age of a listener. Older listeners tend to have a higher absolute hearing threshold at certain frequencies. For young people, the hearing threshold is the lowest for the 2-5 kHz range. It rises for low frequencies, as well as for high frequencies as shown on figure 2.5.

In auditory models, the absolute hearing threshold is often thought of as part of the masking process. Internal noise in the human ear originating from blood flow and neural activity as well as the decreased sensitivity of the ear due to the outer and middle ear transfer function results in the masking of sounds with minimal energy levels. The absolute hearing threshold is therefore often modelled as noise with a constant level that is added to every frequency band, effectively masking signals in bands with low energy [24, 71, 97, 127, 128]. Other models include the absolute hearing threshold more directly by applying a weighting when computing model outputs or internal variables. The noise loudness model [113], the PAQM model [7] and Stuart's model [126] account for the absolute hearing threshold by applying a weighting in the (specific) loudness calculation, whereas the loudness function used to compute the Noise-to-Mask ratio (NMR) [16] includes the correction in the graphic calculation method. Kapust [70] chose a similar approach in setting the absolute hearing threshold as base level for the calculation of his masking function. In Sporer's model [120], the absolute hearing threshold is directly included in the excitation filters used to construct the decomposition filter bank. The models by Münkner & Püschel [89], Sander [110], Dau *et al.* [35] or Huber & Kollmeier [58] model the absolute hearing threshold by limiting the input to the internal feedback loops to a minimal value.

2.8.2 Simultaneous masking

When two signals in the same frequency bands are presented simultaneously, the louder signal will potentially mask the signal that is less loud. Figure 2.6 shows 3 cases of an increase in audibility threshold around a bandpass noise of 60 dB at 3 different frequencies (0.25 kHz, 1 kHz and 4 kHz) with a bandwidth of 1 critical band. It clearly shows that if a signal is to be heard in the same frequency band, it will have to be nearly as loud as the noise. Note that the audibility threshold does not reach the level of the noise (60 dB) and signals with a lower level in the same band as the noise might therefore be heard.

Not only does a signal affect the audibility threshold of other signals in the same frequency band, but also in adjacent frequency bands. This is due at least partly to the fact that the basilar membrane does not respond to a pure tone in a single location, but rather in a region, resulting in an auditory filter. The width of the audibility threshold curves in figure 2.6 give an indication of that phenomenon. The increase in audibility threshold extends above and below the bandwidth of the masking noise. In figure 2.7, where the audibility threshold curves are depicted for different levels L_G of a masking noise centered at 1 kHz with a bandwidth of one critical band, it can be clearly seen that the louder the masking noise, the more it affects especially higher frequency bands.

The consequence of such masking patterns is that parts of complex signals (such as music) can be masked by other components of the same signal. This psychoacoustic property is leveraged in perceptual compression algorithms such as MP3 for example where inaudible parts are simply removed. It is also used in coding applications where the spectrum of the quantisation noise is shaped in a way that it's masked by the useful signal. Auditory models

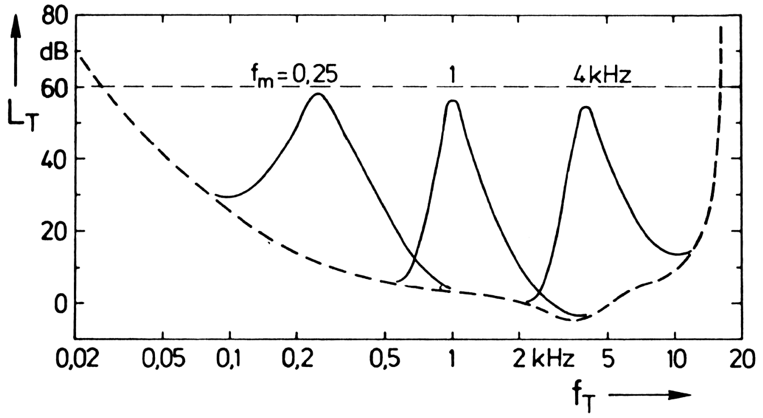


Figure 2.6 – Effective auditory threshold in the presence of a band-pass noise centered at different frequencies (0.25 kHz, 1 kHz and 4 kHz) with a bandwidth of 1 critical band with a level of 60 dB from Zwicker [147]. The difference to the absolute hearing threshold (dashed line) is known as masking. Source: Zwicker [147].

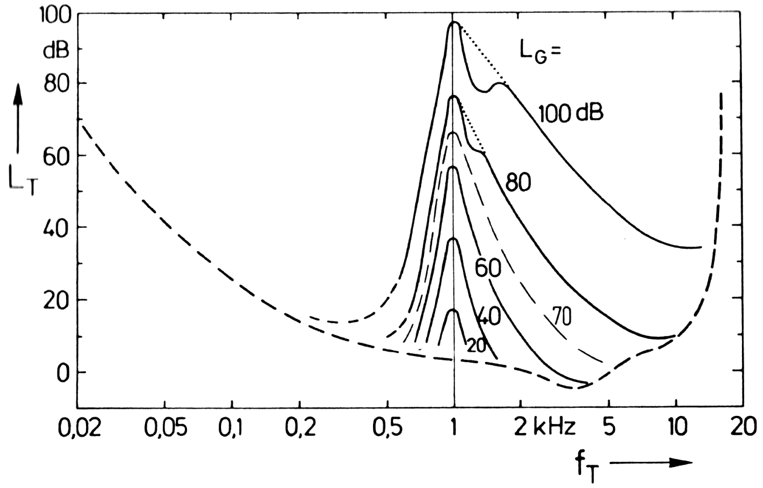


Figure 2.7 – Effective hearing threshold in the presence of a band-pass noise centered at 1 kHz with a bandwidth of 1 critical band at different levels L_G from Zwicker [147]. Source: Zwicker [147].

need to account for it since it's an inherent property of human audition.

In Schroeder *et al*'s model [113], masking towards other frequency bands is modelled by convoluting the spectrum that was transformed to the pitch domain with a spreading function. This spreading function is level independent and has a lower slope of +25 dB per critical band and an upper slope of -10 dB per critical band. Simultaneous masking is then modelled via a simple weighting that is applied in the calculation of noise in presence of a speech signal.

Brandenburg proposed a metric called NMR in 1987 that simulates the masking function in the human ear [16]. For the audio coding application described by Brandenburg, the short-term spectra of the signal and of the quantisation noise are computed and a masking function is calculated based on the absolute hearing threshold as a function of frequency, the masking by the signal within each band and the masking by the signal from lower to higher frequencies. Based on these properties, Zwicker [147] derived a masking function that is used in the calculation of the loudness of stationary noise and was adopted in DIN 45631 [39] and ISO 532:1975 [60]. Based on a graphical procedure using third-octave levels, this function is computed for the signal to be encoded in Brandenburg's application [16] and it is then compared to the spectrum of coding noise. Whenever the energy of the noise surpasses the value of the computed masking function, a masking flag is set to signal potentially audible quantisation noise.

In Paillard *et al.*'s model [97], masking towards higher and lower frequencies is modelled by a dispersion filter that is applied to localised basilar energy distributions gained from the frequency decomposition. The shape of this filter is that of a double-sided decreasing exponential with time constants taken from experiments reported by Feldtkeller & Zwicker [44].

A similar approach is taken in the PAQM model [7] where pitch components are convolved with a spreading function. The different components in every frequency band due to these smearing processes are then added according to

$$M_{\text{composite}} = \left(\sum_{i=1}^{n} M_i^{\alpha} \right)^{1/\alpha}, \qquad \alpha < 2, \tag{2.3}$$

where M_i is the energy of the different contributions and α is a compression power. The FFT-based version of PEAQ [128] also uses a spreading function, where a constant lower slope of 24 dB/Bark is used and an upper slope that depends on frequency. Different contributions are then added nonlinearly using a power-law model.

In Kapust's model [70], intra-band masking is evaluated taking into account the tonality of the signal, since masking depends on the type of signal, and a measure of masking that are both empirically determined. Masking towards higher and lower frequencies is taken into account by applying a spreading function in the form of a triangle in the Bark domain with a fixed lower slope of 27 dB/Bark and a higher slope that depends on the SPL. The Perceptual Objective Measure (POM) model [24] also uses a spreading function, but in the shape of a triangle with a rounded top.

Auditory models using filter banks [35,58,126–128] generally directly model masking properties of the ear via the overlapping bands of the filter bank.

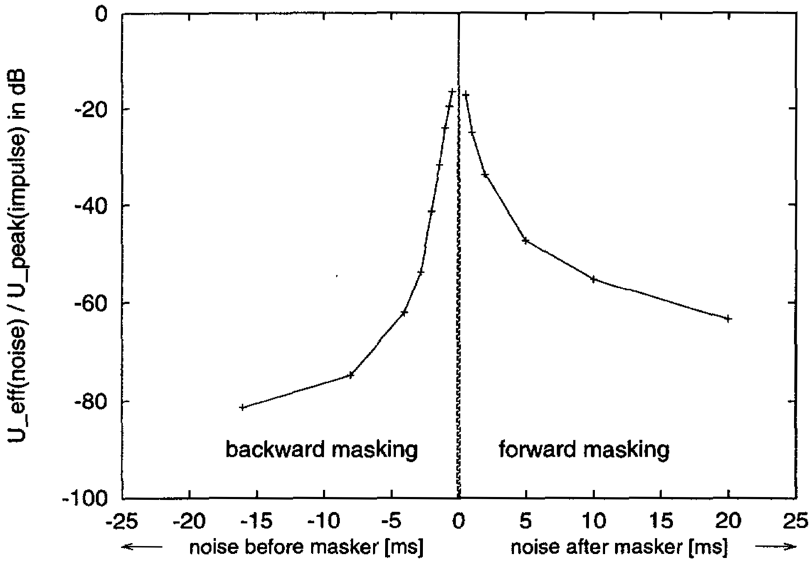


Figure 2.8 – Temporal masking of Gaussian noise by a Gaussian impulse as illustrated by Sporer [120]. Source: Sporer [120].

2.8.3 Temporal masking

While simultaneous masking occurs when a masker is present throughout the duration of the maskee, temporal masking occurs when the maskee is present at another time than the masker. Two situations can be distinguished: If the masker occurs before the maskee, there is a short time period where the audibility threshold of the maskee is raised because of the presence of the masker. This is called "forward masking" or "post-masking". One explanation of forward masking would be that the stereocilia need some time to recover after a period of stimulation [108]. The same is happening if the maskee precedes the masker and it's called "pre-masking" or "backward masking".

The duration and the amount of temporal masking depends on the delay between the masker and the maskee, the level of the masker, as well as the spectral and temporal characteristics of both signals [86]. Figure 2.8 shows an example of Gaussian noise being masked by a Gaussian impulse. Noise occurring a few milliseconds before the impulse and a few after the impulse is masked by it.

In Karjalainen's model [71], pre- and post-masking is modelled by a nonlinear low-pass filter with a 100 ms time constant. A 10 ms delay in post-masking is however reported but rated as noncritical for their application. The same nonlinear filtering approach for post-masking is taken by Sporer [120], but with a time constant of only 3 ms.

In PAQM [7], forward masking is modelled by spreading the energy of the pitch components in one time frame to the next overlapping frame after dampening it exponentially according to a frequency-dependent time constant. Based on what is explained in an earlier paper [56], Hollier *et al.*'s model [55] probably also uses some kind of dampened carry-over of the energy from one frame to the next. The exponential approach is judged invalid by Kapust [70] based on the work of Zwicker [147]. In Kapust's model [70], post-masking is therefore modelled through a normalised decay D depending on the duration of the masker d and time t (both expressed in ms):

$$D = 1 - \frac{1}{1.35} \arctan\left(\frac{t}{13.2d^{0.25}}\right) \tag{2.4}$$

In Thiede and Kabot's model [127], time-domain smearing is applied to model pre- and post-masking. This is implemented by averaging the outputs of the auditory filters over a sliding squared cosine window with a length of 8 ms and then applying a first order low-pass filter. Time constants depend on the filter center frequency. PEAQ [128] also models pre-and post-masking through filtering, but two filter stages are used. The first one is a 8 ms raised cosine shaped FIR filter to account for pre-masking, whereas the second one is a first order Infinite Impulse Response (IIR) low-pass filter with frequency-dependent time constants to model post-masking (50 ms at 100 Hz, 4 ms at least for high frequency bands).

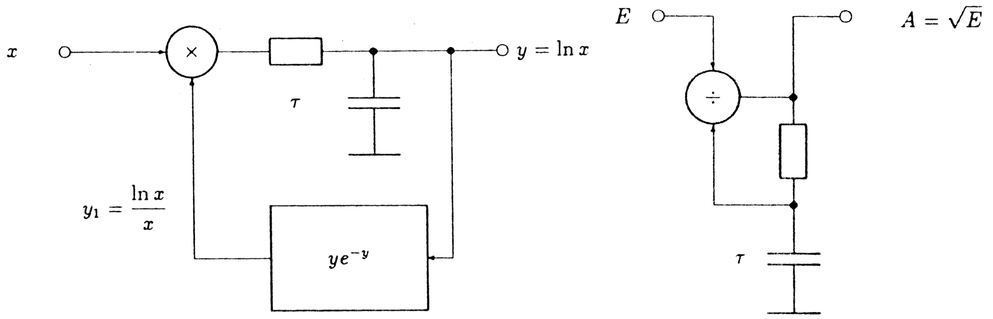
In the models by Münkner & Püschel [89], Sander [110], and Dau *et al.* [35], the feedback loops used for adaptation modelling also implicitly model forward masking. Their charge decays over time after the input drops and piecewise model the downward slope of forward masking.

2.9 Adaptation

Auditory adaptation is the decrease in auditory response to a constant stimulus over time. It is linked to hair cells (or other nervous cells along the auditory pathway [105]) reaching an equilibrium between the energy that they use for the transmission of information and the metabolic energy that is available to them [86]. In terms of perceived loudness, this results in decreasing loudness until a steady state is reached.

After trying different approaches, Püschel proposed to model adaptation through the use of one or several feedback loops, where an adaptive gain is fed back to the input [105]. Two feedback loop topologies were proposed originally: one with a nonlinear adaptive gain (in the form of the logarithmic element in the feedback loop) with a first-order low-pass filter (see figure 2.9a) and a second one where the value at the input of a low-pass filter is fed back as a denominator to the input (see figure 2.9b). The output of the first topology is equal to the logarithm of the input in steady state, whereas the output of the second one corresponds to the square root of the input. Applying multiple loops of the second topology in series, the output will be the 2^nth root of the input, approaching a logarithmic law.

The advantage of such an implementation is that it also models temporal masking through the time constants of the low-pass filters and obviously logarithmic compression. The first implementation with the logarithmic compression was found to converge to the steady state too quickly [105], which is why the second topology was proposed. Five consecutive feedback loops were used in the end, to approach the logarithmic function well enough. Time constants with different values enable the loops to reach steady states at different times, enabling some loops to transmit the signal uncompressed, while others are already in a steady state. Since



(a) Adaptation loop with nonlinear adaptive gain feedback (b) Adaptation with first-order low-pass filter feedback

Figure 2.9 – Adaptation loop topologies as proposed by Püschel [105]. Figures by Püschel [105].

human audition integrates information for about 200 ms [105], a corresponding low-pass filter can be added after the last feedback loop. In the original proposal, the time constants of the low-pass filters of the feedback loops were linearly spaced between 5 ms and 500 ms.

This way of modelling adaptation was followed by several models, sometimes with some adaptation of the values of the time constants of the feedback loops [35, 58, 89, 110].

2.10 Modulation processing

The temporal resolution of the ear is finite. Periodic variations in amplitude (temporal modulation) can only be detected up to a certain modulation rate. Modulation detection experiments have shown that modulation rates above 1 kHz can't be detected.

Below 16 Hz, the detection threshold is independent of modulation rate because of the amplitude resolution of the ear [86]. In that range, modulation is associated with the perception of rhythm, with modulations rates of 3-4 Hz corresponding to the rate of syllables or words in speech [79].

Above 16 Hz, amplitude modulation is called roughness. The sensitivity of the ear to amplitude modulation gradually increases as a function of modulation rate up to a maximum between 40 and 70 Hz [79], before decreasing again and reaching 0 at 1 kHz [86].

Sensitivity to amplitude modulation may be approximated by a low-pass filter with a cut-off frequency of 50 Hz [79]. Some auditory models use a modulation filter bank to model the analysis of envelope periodicity in human audition [58, 89, 110].

2.11 Audio quality evaluation by auditory models

Auditory models aiming at the prediction of the quality of a given audio item model some or most of the phenomena presented above. As stated in the introduction of this chapter, the first auditory models were developed for the evaluation of compressive digital audio coding. More recently, such models have also been used in a broader context, but the task of audio quality evaluation has stayed the same (at least for the presented models). The following sections present some models that lead up to the current state of the art in ASS evaluation.

2.11.1 Perceptual Evaluation of Audio Quality (PEAQ)

PEAQ was adopted as an International Telecommunication Union (ITU) recommendation in 1998 [65] and consists of a joint development of the authors of the predominant auditory models of the time. While this modelling approach differs in a few ways from the current state of the art in ASS evaluation, it illustrates how the phenomena that were evoked above can be integrated into an auditory model.

As described by Thiede *et al.* [128], two versions of PEAQ are available. A basic version, based on an ear model using a FFT for frequency decomposition that maximizes processing speed, and an advanced version maximising prediction accuracy and based on both a FFT-based ear model and an ear model using a filter bank were proposed. The respective block schemes of both ear models can be found on figures 2.10 and 2.11.

The FFT-based ear model uses 50% overlapping 2048-sample Hann windows. This corresponds to 21 ms temporal resolution at 48 kHz sampling frequency. Rectification is achieved by discarding phase information of the obtained spectra. The outer and middle ear are modelled by applying a filter, which is translated to a spectral weighting function. After grouping coefficients by 0.25 or 0.5 Bark depending on wether the basic or the advanced PEAQ model is used, a frequency-dependent offset is added to model internal noise. The energy dispersion along the basilar membrane is then approximated by first using a filter with a constant lower slope of 24 dB/Bark and a level- and frequency-dependent upper slope on every frequency group. After this, the different contributions from different frequency bands are added nonlinearly using a power-law model. Backward masking is not taken into account due to the poor time resolution of the FFT. Forward masking is modelled with a first-order IIR filter, smearing the frequency components in time. However, in order to preserve onsets, components of the resulting internal representation that are smaller when filtered than when unfiltered are replaced by their unfiltered value.

The filter bank-based ear model uses a linear-phase filter bank. The bandwidths and the level-dependent slopes of the 40 filter pairs directly model the characteristics (frequency resolution) of the basilar membrane. Filter bands are summed using a frequency weighting to preserve exponential decay on the slopes. To cope with changing levels over time, the variation of the level-dependent slope is limited using a low-pass filter. Rectification is achieved by computing the instantaneous energy of the signal. Temporal masking is modelled through a very short FIR for backward masking and a longer IIR filter for forward masking. Internal noise is modelled through outer- and middle-ear transfer functions and the addition of a frequency-dependent offset.

One original feature of this model is that linear and nonlinear degradations are treated separately. This is done by adapting the internal representation of the signal under test to the reference over time, based on components that are present in both signals.

The PEAQ model outputs a whole set of variables that are derived from the internal representations. This includes a measure of envelope modulation, modulation difference, partial noise loudness, audible linear distortion, NMR, signal bandwidth, detection probability and the error harmonic structure. These outputs are then mapped to subjective scores from listening tests using an artificial neural network. The article presenting the PEAQ model [128] does not report any goodness of fit and simply states that “both versions are superior to previously existing measurement methods”.

2.11.2 Dau *et al.*

Whereas most approaches that were developed in the early stages of auditory modelling aimed at the prediction of the quality of (mostly low bit-rate) coding algorithms, there were a few that aimed at predicting the quality degradation induced by any kind of distortion. Such an approach is taken by Dau *et al.* [35], since they compare a signal including a masker and a test signal to the masker alone.

The first stage of their model is a time-frequency decomposition, as seen in other models, via a 120-band filter bank. After that, half-wave rectification and low-pass filtering at 1 kHz follows. The low-pass filtering simulates the preservation of envelopes for higher frequencies. Auditory adaptation is then implemented by feedback loops following Püschel [105] and Münkner & Püschel [89]. A stationary input to such a feedback loop will charge the output over time, rising the divisor and therefore slowing down the charging. This compresses stationary inputs with a near-logarithmic rate, whereas rapid fluctuations are processed more linearly.
Not only do the feedback loops model adaptation, they also model forward masking. If the input of the loops reaches zero, the charge of the loops will decay exponentially, reaching the system’s initial state after a given time constant. Additionally, a lower limit to the input of the adaptation stage simulates internal noise, accounting for the absolute hearing threshold. Five feedback loops in series with different time constants were used in this model. After the adaptation loops, the signal is low-pass filtered at 8Hz, resulting in the internal representation of the signal.

The model described by Dau *et al.* [35] is designed to estimate detection thresholds. Corresponding subjective experiments are most often forced-choice experiments. In such experiments, a listener typically has to choose which signal of two choices has been processed in some way, as opposed to the other one that was not. The processed signal is very easily detectable at first (largely suprathreshold) and the processing is then attenuated until the listener is not able to choose correctly anymore. Successively decreasing and increasing the processing with adaptive step sizes allows to determine the detection threshold.
Because of that experimental structure, the model proposed by Dau *et al.* [35] supposes that a template of the test signal can be constructed from the difference between the internal

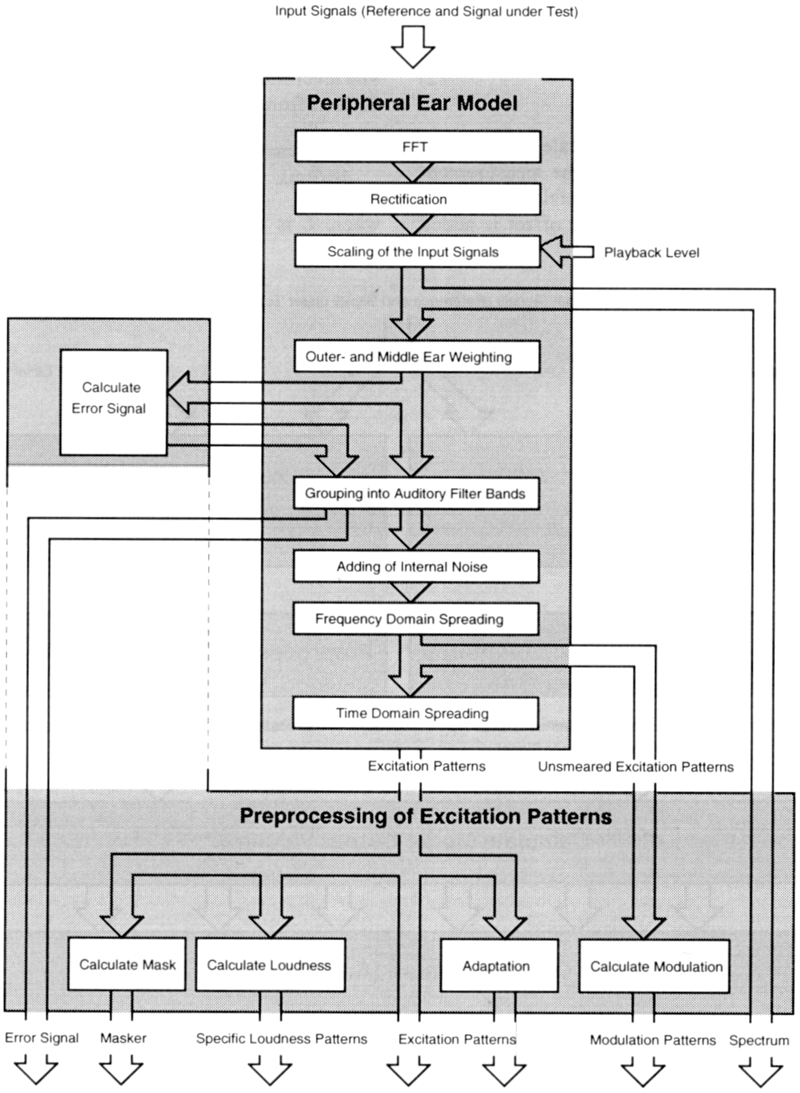


Figure 2.10 – Block scheme of the FFT-based version of PEAQ by Thiede *et al.* [128]. Source: Thiede *et al.* [128].

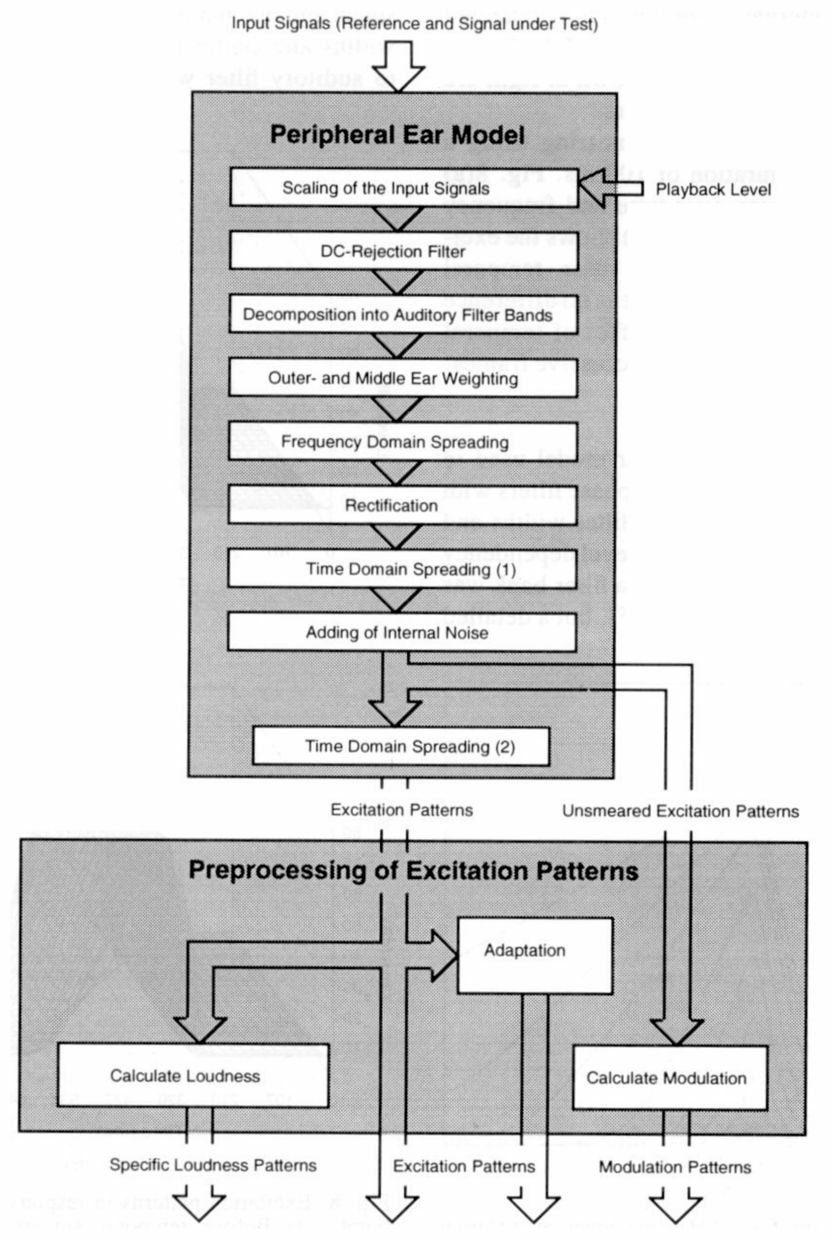

Input Signals (Reference and Signal under Test)
Peripheral Ear Model
Scaling of the Input Signals
Playback Level
DC-Rejection Filter
Decomposition into Auditory Filter Bands
Outer- and Middle Ear Weighting
Frequency Domain Spreading
Rectification
Time Domain Spreading (1)
Adding of Internal Noise
Time Domain Spreading (2)
Excitation Patterns
Unsmeared Excitation Patterns
Preprocessing of Excitation Patterns
Adaptation
Calculate Loudness
Calculate Modulation
Specific Loudness Patterns
Excitation Patterns
Modulation Patterns

Figure 2.11 – Block scheme of the filter bank-based version of PEAQ by Thiede *et al.* [128]. Source: Thiede *et al.* [128].

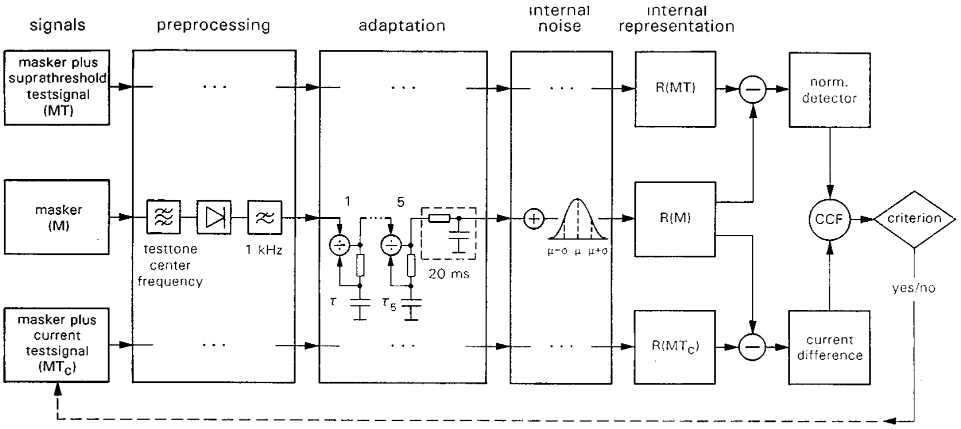


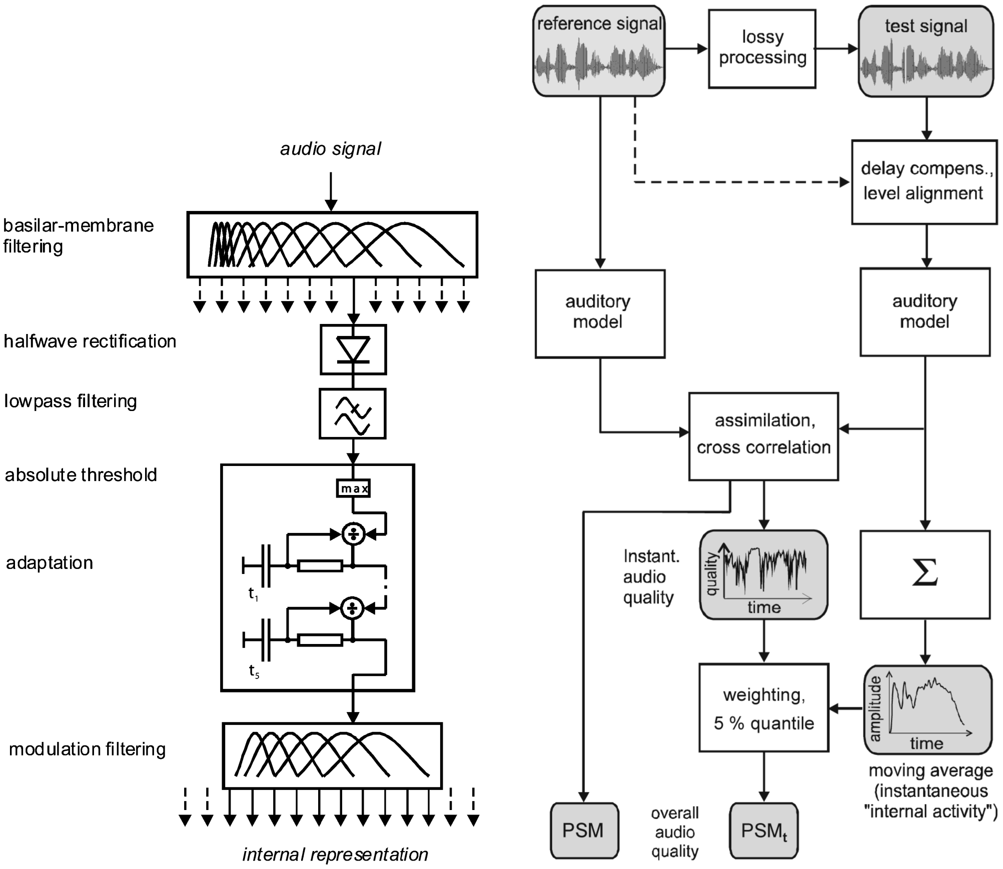
Figure 2.12 – The block diagram of Dau *et al.*'s model [35]. Source: Dau *et al.* [35].

representation of the masker plus test signal at a suprathreshold level and the internal representation of the masker alone. This template is then used for detection at lower levels through correlation. The detection process is implemented using an optimal detector. The block diagram of the model is depicted in figure 2.12.

Results of this model are reported in a companion paper to the model [36]. Multiple experiments are reported together with the output of the model. For simultaneous masking, predictions are generally within 3 dB of the experimental values. The model does however not match subjective data when the test signal is too long and further optimisation of the adaptation loops is suggested.
A corrected version of the adaptation loops was used for forward masking experiments, where results of the model generally matched experimental data. For short masker durations, some divergence between the model and experimental data occurs. This, once again, was blamed on the missing optimisation of the adaptation loops.
A model derived from Dau *et al.*'s proposition was later used to model auditory separation (the ability of human audition to concentrate on one sound in a mix of others even in monaural conditions) with modifications to the adaptation loops [110]. Hansen & Kollmeier also used a similar model topology to predict speech quality [50].

2.11.3 Perceptual Model of Audio Quality (PEMO-Q)

Building on all the work from previous models presented in the last sections, Huber & Kollmeier presented a model aimed at general audio quality prediction called PEMO-Q [58]. Based on an adapted version of Dau *et al.*'s model [35], a 35-band fourth-order gammatone filter bank is used for time-frequency decomposition. Center frequencies are regularly distributed on the ERB scale between 235 and 14500 Hz. Half-wave rectification and 1 kHz low-pass filtering are applied as in Dau *et al.* [35] as well as the 5 adaptation loops with time constants between 5 and 500 ms. A modulation filter bank is then applied with constant 5Hz filter bandwidths



(a) Block diagram of the PEMO-Q auditory model (b) Block diagram of the PSM calculation

Figure 2.13 – Block diagrams of the auditory model of PEMO-Q and of the whole method, including PSM calculation [58]. Figures by Huber & Kollmeier [58].

up to 10 Hz center frequency and logarithmic center frequency spacing above 10 Hz with a constant quality factor of 2, overlapping at -3 dB. For center frequencies above 10 Hz, the Hilbert envelope is extracted to model some loss of information. In total, 8 bands were used in the modulation filter bank with center frequencies between 5 and 129 Hz. Downsampling is used to reduce storage space of the resulting internal representations. The block diagram of this auditory model is given in figure 2.13a.

In post-processing of the internal representations, the internal representation of the signal under test y is partially assimilated to that of the reference x when it has smaller absolute values:

$$\tilde{y}_{tfm} = \begin{cases} \frac{(y_{tfm}+x_{tfm})}{2}, & |y_{tfm}| < |x_{tfm}| \\ y_{tfm}, & |y_{tfm}| \geq |x_{tfm}| \end{cases}, \tag{2.5}$$

where t, f, and m denote the subscripts for the temporal frame, the filter band and the modulation filter sub-band respectively. This takes into account the assumption that missing components in the signal under test are less disturbing than additive components. The cross-correlation coefficient r_m between the two internal representations is then computed in every modulation channel before weighting them according to the normalised mean squared values of the corresponding modulation channel and summing them up into the final quality measure PSM:

$$\mathrm{PSM} = \sum_m w_m r_m, \text{with} w_m = \frac{\sum_{t,f=1}^{N,M} y_{tfm}^2}{\sum_{t,f,m'=1}^{N,M,L} y_{tfm'}^2} \tag{2.6}$$

where N, M, and L denote the number of temporal frames, filter bands and modulation sub-bands respectively. This measure denotes, on a scale between -1 and 1, how much the signal under test is different on a perceptual level from the reference.

A second model output is also calculated by computing the PSM measure over sections of 10 ms instead of the entire length of the signal. After that, this instantaneous quality measure PSM(t) is weighted by the moving average of the internal representation of the test signal, resulting in a rough approximation of instantaneous loudness. The second output PSM$_t$ is then computed by calculating the fifth percentile of it. The PSM$_t$ is cited as being independent from the type of input signals [58]. The block diagram of the whole PEMO-Q model is given by figure 2.13b.

A nonlinear transformation is applied to the PSM$_t$ measure to map it to subjective scores from different databases. A correlation coefficient of about $r = 0.9$ was achieved between the subjective scores and the objective scores for codec quality evaluation experiments predicted by the PEMO-Q model. The only flaw that is mentioned is that the 10 ms time constant used for cross-correlation is short with respect to forward masking. But since real-world signal were used, sharp offsets are rarely observed, due to reverberation, and forward masking plays a minor role in perceived quality. When compared to the PEAQ model, PEMO-Q outperformed PEAQ in almost all given settings.

2.12 Conclusion

Much work has been done in the area of psychoacoustically motivated auditory models. More and more complete models are able to integrate more and more psychoacoustic phenomena, getting even closer to what the input to the auditory nerve and the more central stages of auditory processing must be. The results seem to confirm the approach as subjective scores and a wide range of psychoacoustic experiments are modelled pretty accurately with the presented models.

The models presented in this chapter give an understanding of the state of the art in auditory modelling and shall support the choices that are made throughout this work. Note that the models presented here are mostly used for audio coding quality evaluation. A whole series of models exist that evaluate speech coding quality. Even though they often share at least parts of the components of the auditory models, they weren't presented here mainly because the topic of this thesis is ASS quality evaluation, which is very similar to audio coding quality evaluation. The interested reader may however find speech coding quality evaluation models in the literature based on the current ITU recommendation [66].

In the view of all the results from psychoacoustic modelling and the models that are presented in this chapter, it may seem fit to use such an auditory model to predict the perceived quality of ASS signals which should not be very dissimilar to the prediction of the quality of audio coding algorithms. In fact, this approach is taken in Perceptive Evaluation Methods for Audio Source Separation (PEASS), where the PEMO-Q model is used to predict the perceptual salience of the different components of ASS degradations [42]. This model linking ASS and perceptual modelling will be presented in the next chapter.

3 Evaluation of audio source separation with an auditory model

3.1 Introduction

The objective Audio Source Separation (ASS) evaluation algorithm presented in chapter 1, that uses the decomposition of the degradations into specific components, lacks the link to subjective evaluation data. The presented measures provide an objective quality scale, that is somewhat related to the opinion of listeners who are presented with the signals that are submitted to the ASS evaluation algorithm. However, as stated by Emiya *et al.*, the filters that are used do not match the frequency resolution of the human ear, the low frequencies affect the energy ratios much more than the human ear and auditory masking of degradations is not taken into account [42]. In practice, correlation coefficients between the energy ratios and subjective data vary widely between -0.16 and 0.85 [42], which is not optimal. This may be resolved through the use of a more psychoacoustically motivated approach as it has been taken in the auditory models presented in chapter 2. These models aim at reproducing the characteristics - if not whole parts - of the human auditory process. Results are often clearly related to subjective data gathered through psychoacoustic experiments, which is why such models are used more and more.

For ASS, the state-of-the-art link between objective ASS evaluation and subjective data is given by the Perceptive Evaluation Methods for Audio Source Separation (PEASS) presented by Emiya *et al.* in 2011 [42]. An introduction to this model is given in section 3.2 to provide an overview of the state of the art. The common problem of mapping objective data from an auditory model to subjective data from psychoacoustic experiments is then presented in section 3.2.1 to introduce the optimisation process that the state of the art has undergone in the past. The modifications to objective ASS evaluation introduced in chapter 1 are then proposed as an extension to PEASS in section 3.3. The performance of the extended version is assessed using existing subjective data and conclusions and perspectives are drawn as for the further use of this model.

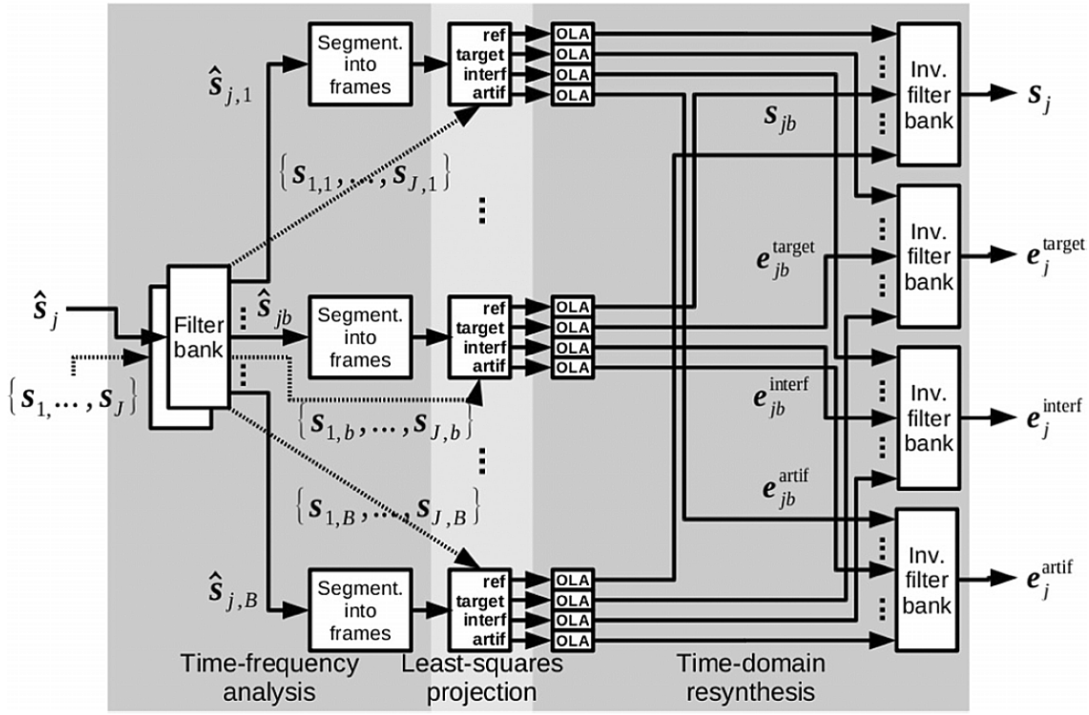


Figure 3.1 – The perceptually-motivated decomposition algorithm used in PEASS to decompose estimated source signals $\hat{\mathbf{s}}_j$ into degradation components $\mathbf{e}_j^{\text{target}}, \mathbf{e}_j^{\text{interf}}$ and $\mathbf{e}_j^{\text{artif}}$ by comparing them to the clean source signals $\{\mathbf{s}_1,...,\mathbf{s}_j\}$. For reference, the clean source signal $\mathbf{s}$ is also put through the decomposition algorithm. Source: Emiya *et al.* [42].

3.2 Perceptive Evaluation Methods for Audio Source Separation

In 2011, Emiya *et al.* presented the PEASS model, along with a subjective test protocol [42]. The PEASS model is the combination of the decomposition presented in section 1.3.5 and the Perceptual Model of Audio Quality (PEMO-Q) model by Huber & Kollmeier [58] presented in section 2.11.3.

First, the estimated source signals $\hat{s}_{ij}$ provided by some ASS algorithm are compared to the true source signals (the reference) s_{ij} and their difference is decomposed into 3 components as presented in section 1.3.5, where i and j index the I channels and J sources of a mixture.

$$\hat{s}_{ij}(t) - s_{ij}(t) = e_{ij}^{\text{target}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t). \tag{3.1}$$

This decomposition follows a perceptually-motivated approach through the use of a gammatone filter bank and least-squares projection (see section 1.3.5 for details). Figure 3.1 gives an overview of the involved decomposition process.

The subjective salience of each of the components is then assessed through the use of the Perceptual Similarity Measure (PSM) provided by the PEMO-Q model. This is achieved by comparing the estimated source signal with the reference minus the degradation component that is being assessed:

$$\begin{aligned} q_j^{\text{overall}} &= \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j) \\ q_j^{\text{target}} &= \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j - \mathbf{e}_j^{\text{target}}) \\ q_j^{\text{interf}} &= \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j - \mathbf{e}_j^{\text{interf}}) \\ q_j^{\text{artif}} &= \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j - \mathbf{e}_j^{\text{artif}}) \end{aligned} \tag{3.2}$$

The PEASS model was accompanied by a subjective experiment about ASS evaluation to validate the modelling approach. 20 participants rated 80 sounds coming from 10 mixtures. In each mixture, the 8 test sounds consisted in 4 “real-world” sounds from real ASS algorithms, 3 low anchors containing one of 3 synthetic degradations each and 1 hidden reference. All of those sounds were rated according to 4 different task:

1. rate the global quality compared to the reference
2. rate the quality in terms of preservation of the target source
3. rate the quality in terms of suppression of other sources
4. rate the quality in terms of absence of additional artifacts

This resulted in 4 different subjective scales for every test item. These scales do not match the scales of the similarity measures from the model. The similarity measures q_j^{overall}, q_j^{target}, q_j^{interf} and q_j^{artif} from the model are therefore taken as inputs for a nonlinear mapping stage that combines the similarity measures or a subset thereof to get objective scores that correlate with the subjective scores on the subjective grading scales from the accompanying experiment. Since the subjective experiment consisted of 4 tasks related to the different degradation components, a set of 4 objective measures are computed, namely Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS) and Artifact-related Perceptual Score (APS) as illustrated by figure 3.2. The nonlinear mapping is performed through the use of an Artificial Neural Network (ANN) with sigmoidal activation functions. Details about the mapping, ANNs and their optimisation are given in the next section.

The final PEASS model achieved a prediction accuracy (measured through Pearson's linear correlation coefficient) of 0.61 for the OPS. A monotonicity rating of 0.55 and a consistency rating of 0.87 are also reported for the OPS. Monotonicity is measured by Spearman's rank correlation (the linear correlation coefficient between the ranks of the objective and subjective data after sorting) and consistency is defined as $1 - R_o$ where R_o is the ratio of sounds and subjects for which the prediction is further away from their subjective score than twice the standard deviation of the subjective scores over all subjects for that sound.

The PEASS model underwent an optimisation process presented by Vincent in a subsequent paper [131]. The optimisation of the nonlinear mapping and of the internal parameters of the

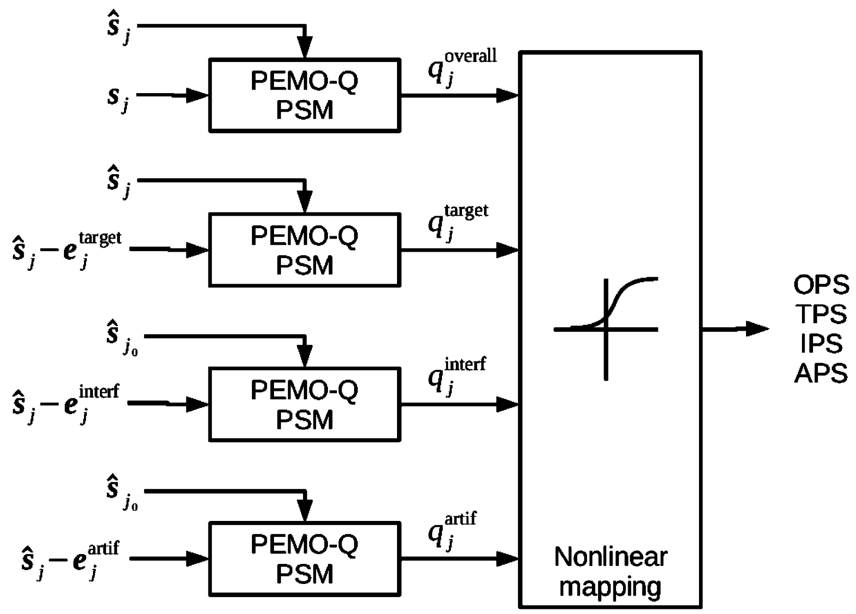


Figure 3.2 – Mapping stage of PEASS. The perceptual similarity measures q_j^{overall}, q_j^{target}, q_j^{interf} and q_j^{artif} are combined and a nonlinear mapping is applied to get objective scores that match the subjective scores from the ASS quality assessment experiments. Source: Emiya *et al.* [42].

model allowed for the improvement of the accuracy to reach 0.909 on average over all 4 tasks.

3.2.1 Mapping objective scores to subjective evaluation scores

Mapping in auditory models

Even though auditory models and perceptually-motivated objective measures are able to model the human auditory process to some extent, the “universal” auditory model does not exist (yet). State-of-the-art measurement techniques have made available more and more physical data over the years, bringing within reach from the sound field at the eardrum up to data about neural patterns in the auditory nerve and more complex activation processes in higher auditory neural centers. However, the human auditory process is yet to be fully understood, hence the active research fields in auditory scene analysis, stream segregation, etc. Objective data about rather complex cognitive processes such as quality evaluation, even when transformed by an auditory model, will therefore almost never match subjective data to a satisfactory level. In the best case, a linear relationship between the two sets of data exists and a regression analysis through a linear least-squares projection for example can be performed, giving a simple link between the two. When this is not the case, a more general approach may be taken with nonlinear mappings such as ANNs or other machine learning algorithms.

In a general manner, it can be said that the less linear the relationship between the subjective data and the objective data is, the more computationally complex the mapping algorithm (or the training thereof) is. On the other hand, the more psychoacoustic phenomena are taken into account by an auditory model, the more computationally complex it is, but also, the more linear the relationship between the objective data and the subjective data may be. There is therefore a tradeoff between the computational complexity of the auditory model and the mapping algorithm. At one end of this tradeoff scale, so-called deep learning algorithms mapping the raw sound pressure signals to subjective scales do not make any psychoacoustically motivated *a priori* assumptions (for a review, see *e.g.* Deng & Yu [38]). On the other extreme end of the scale, the hypothetic "complete" auditory model as it is drafted for example by Blauert *et al.* [15], could model all auditory processes without the need for any subsequent mapping. Most of the models presented in chapter 2 are somewhere in between these two, by applying some results from psychoacoustics as *a priori* information and then adding a mapping layer to match subjective scales from experiments.

Herre *et al.* [52] used a linear transformation to map Noise-to-Mask ratio (NMR) scores to subjective scores. The PEMO-Q model uses a mix between a linear regression and a hyperbolic regression function to match subjective scores [58]. In the Perceptual Evaluation of the Quality of Audio Signals (PERCEVAL) [97] and the Perceptual Audio Quality Measure (PAQM) [7] models, internal parameters are optimised to achieve optimal correlation. The Distortion Index (DIX) model [127] uses a sigmoidal (*i.e.* nonlinear) mapping function to match subjective scores. Hollier *et al.* suggest either a sigmoidal mapping function or a clustering algorithm similar to ANNs or a mix between those methods [57]. The Perceptual Evaluation of Audio Quality (PEAQ) model uses an ANN to map the model output variables to subjective scores [128], just like the PEASS model presented above [42].

Linear regression

One of the most simple forms of mapping is given by linear regression. A linear regression model written in matrix notation has the form

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \epsilon, \tag{3.3}$$

where $\mathbf{y}$ is the vector containing the target variable (subjective scores in the case of a objective to subjective scores mapping), $\mathbf{X}$ is the matrix containing the predictors (the objective scores, *i.e.* model output variables), $\hat{\beta}$ contains the estimated regression parameters and ϵ contains the error (the part of the target variable that could not be predicted).
The most common estimator for $\hat{\beta}$ is given by ordinary least squares. It minimizes the sum of squared residuals (*i.e.* $\sum_{i=1}^{I} \epsilon_i^2$). The estimated value of the regression parameters $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.4}$$

where the $(\cdot)^T$ denotes transposition. This estimator is unbiased and consistent if the variance of ϵ is finite and if ϵ is uncorrelated with the columns of $\mathbf{X}$ [74]. It is very useful when a linear relationship between the predictors and the target variable seems to exist. This is the estimator that is used with the binaural modelling approach in the two experiments presented in chapters 6 and 7.

Artificial Neural Network (ANN)

When a linear relationship does not exist, nonlinear mapping functions may be used. One such function is provided by ANNs. Inspired from human neural interconnection and activity, simple ANNs have the general structure shown on figure 3.3. Note that this is the most simple feedforward structure and much more complex designs exist. Every input-layer neuron transmits the available data (the predictors) to all neurons of the first hidden layer. The hidden neurons apply a (generally nonlinear) transformation to the sum of weighted inputs and transmit their output to either the next hidden layer or to the output layer. In the PEASS model [42], an ANN with one hidden layer of n_{neur} sigmoids is used to map the 4 objective scores OPS, TPS, IPS and APS to the subjective scale used in the experiments. The resulting mapping is given by

$$f_r(\mathbf{q}) = \sum_{k=1}^{n_{\text{neur}}} v_{rk} g(\mathbf{w}_{rk}^T \mathbf{q} + b_{rk}), \tag{3.5}$$

where $\mathbf{q}$ denotes the input features, v_{rk} the output weight, $\mathbf{w}_{rk}$ the vector of input weights and b_{rk} the bias of sigmoid k for a given task r. The sigmoidal activation function g is given by

$$g(x) = \frac{1}{1+e^{-x}}. \tag{3.6}$$

This implementation supposes a linear output layer with the output neurons not using a sigmoidal activation function. In a later optimised version of PEASS, the output neurons also used a sigmoidal activation function and an additional log-mapping of the input features from $[-1,1]$ to $\mathbb{R}$ of the form $q_j^k \leftarrow \log((1+q_j^k)/(1-q_j^k))$ was used [131]. The parameters of an ANN (*i.e.* the weights and biases) are learned by training. A set of known responses are used to estimate the ANN's parameters. However, if all available data is used, overfitting may occur. Overfitting designates the case when the resulting mapping is very accurate in predicting the subjective data from the training set, but rather inaccurate in predicting any other subjective data. A lack of generalisation is the consequence, meaning that the given ANN can't be used to predict subjective data from new objective data. To avoid overfitting, cross-validation is used. In a cross-validation setting, part of the available data is left aside during training. This data is then used to determine the prediction accuracy of the trained ANN for new data.

In the PEASS model [42], a 200-fold cross-validation setting is used. For each fold, the subjective scores of 19 out of 20 subjects for 9 out of 10 mixtures are used as training data. The testing is then performed on the data of the remaining subject in the remaining mixture for the 4 real-world mixtures. The number of sigmoids n_{neur} was adjusted between 1 and 8 to maximise accuracy (*i.e.* the correlation between the subjective and the objective data). Note that in the later optimised version, all real-world sounds and all anchors but only one reference were used

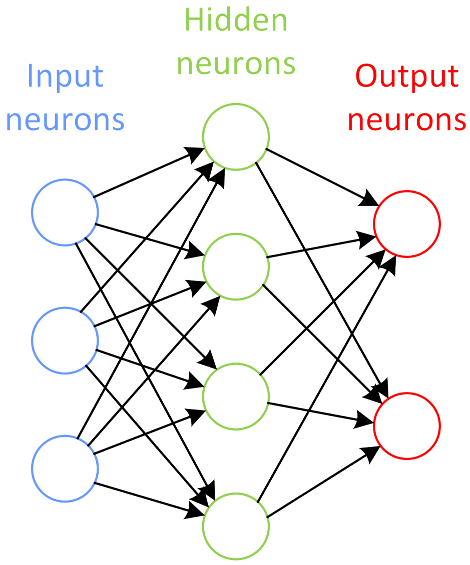


Figure 3.3 – General structure of an ANN. A series of input neurons (blue) distributes the (weighted) inputs to the first hidden layer. Each hidden neuron (green) combines the weighted inputs, transforms them (usually through a sigmoidal activation function) and transmits them to either the next hidden layer or to the output neurons (red).

in training and testing in order to avoid a bias towards the higher values where the references are.

This ANN approach from PEASS has been kept to test the extended decomposition as proposed in section 1.4 in combination with PEASS as presented in the following section.

3.3 PEASS with extended decomposition

The extended degradation decomposition method presented in section 1.4 has been implemented in PEASS, in the prospect of improving the performance of PEASS in the context of 3D audio. Following equations (1.54) and (3.1), the output of the extended decomposition stage can be written as

$$\hat{s}_{ij}(t) - s_{ij}(t) = e_{ij}^{\text{target}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{onset}}(t) + e_{ij}^{\text{artif}}(t). \tag{3.7}$$

The perceptual similarity of the new onset degradation component $e_j^{\text{onset}}(t)$ is then computed using the PSM$_t$ measure just like for the other degradation components in equation (3.2):

$$q_j^{\text{onset}} = \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j - \mathbf{e}_j^{\text{onset}}). \tag{3.8}$$

The resulting feature q_j^{onset} can then be used in the mapping stage, *i.e.* as input to the log-mapping and/or the ANN.

The performance of the extended PEASS model is expected to be at least as good as the state-of-the-art PEASS model for the case without spatialisation. If used in the 3D audio context, it is expected to perform at least as good as the state-of-the-art model as well. Onset misallocation degradations that were probably attributed to either the interference degradation component or the artifact degradation component are now modelled separately. If, as expected, the onset misallocation degradation affects the quality much more in 3D audio than in stereo audio, this might lead to improved performance of the extended PEASS model over the state-of-the-art model.

3.3.1 Parameter optimisation for the PEASS database

As briefly discussed above, the original PEASS model proposed by Emiya *et al.* [42] later underwent an optimisation process as reported by Vincent [131]. The prediction accuracy of the model with optimised mapping and internal parameters was increased from 0.53 (mean over all tasks) to 0.909. This may be partly due to the fact that for the optimised version, all anchors as well as one reference were used in the training and testing data, whereas in the original version, no references or anchors were used. The original version also reports scores for the case where not individual scores, but the Mean Objective Score (MOS) (*i.e.* the mean over all subjects) was used. In that case, the mean accuracy over all tasks improved to reach 0.73. In the case were the MOS was used but anchors and references were included in the training and testing data, the mean accuracy over all tasks even reached 0.86. In the optimised version, the MOS is always used with all anchors and some references. It is therefore suspected that the dramatic improvement in prediction accuracy is not only due to the parameter optimisation, but also to the use of the MOS and of more signals for which the participants are expected to reach an agreement in scoring. This is confirmed by Vincent in the optimisation study, since the baseline performance including anchors and one reference and using the MOS is reported as being 0.738. The parameter optimisation therefore accounts for an increase of 0.171 in prediction accuracy, which is still far from negligible.

As presented by Vincent in the optimisation study [131], there is a set of 11 parameters that can be modified in the PEASS model. Due to the complexity of the model, the formulation of a complete optimisation problem is difficult - if not impossible. The optimisation was therefore performed in 3 stages. The first stage addresses the mapping stage (ANN and optional log-mapping) and the modulation processing, the second stage addresses the similarity measure PSM_t and the last optimisation stage addresses the computation of the internal representations in PEMO-Q.

First stage: Mapping and modulation processing

This stage addresses the nonlinear mapping, with the selection of the optimal feature vector $\mathbf{q}_j$ and parameters for the ANN. This includes the choice of using a sigmoidal activation function for the output neurons or not (n_{lay} = 2-layer network, versus n_{lay} = 1.5-layer network [131]) and the number of neurons in the hidden stage n_{neur} (between 1 and 8 in the original PEASS

version). The optimal parameters were found to be a $n_{\mathrm{lay}} = 2$-layer ANN (including a sigmoidal activation function for output neurons) with log-mapping applied to the input features. The optimal number of neurons and the optimal feature vector vary as a function of the task.
The choice of using a low-pass filter for the modelling of the modulation processing or a filter bank version is also included in this optimisation stage. With a low-pass filter and optimal ANN parameters, the prediction accuracy is reported to attain 0.882. This reflects the better choice of feature vector, the use of the log-mapping and of the output activation function as opposed to the original version.

Second stage: PEMO-Q similarity measure

The second stage of the optimisation process included the selection of the optimal percentile p used in the PSM$_t$ computation, the partial assimilation ratio α, the framelength for the time-varying correlation l_{corr} and the framelength for the Root mean square (RMS) computation l_{amp}.
Optimal parameters were reported to be $\alpha = 0.25$, $l_{\mathrm{corr}} = 100$ ms, $l_{\mathrm{amp}} = 1$ s and $p = 0.5$. The prediction accuracy further improves to 0.898 with those optimal parameters. Note that all of these values are different from the original implementation of PEASS.

Third stage: PEMO-Q internal representations

The third optimisation stage adjusts the internal PEMO-Q parameters relative to the internal representations. Those parameters include the width of the gammatone filter bank at the input of the model, given as the minimal and maximal frequencies f_{min} and f_{max} respectively; the absolute hearing threshold a_{thresh} and the maximum amplitude ratio r_{max} of rapid changes that are emphasised by the adaptation loops.
Optimal values for these parameters are given by the default $f_{\mathrm{min}} = 235$ Hz, $f_{\mathrm{max}} = 14500$ Hz and $r_{\mathrm{max}} = +\infty$ and by $a_{\mathrm{thresh}} = 10^{-6}$. Prediction accuracy attains 0.909 in average over all tasks after this final optimisation stage.

Fourth stage: Framelength of the decomposition filter

The same optimisation process has been followed for the PEASS model with the extended decomposition. In addition to the parameters optimised in the basic version, the framelength for the filtering process l_{filt} of the decomposition was considered in a last optimisation stage. A set of five different framelengths were considered: 75 ms, 150 ms, 300 ms, 500 ms and 1000 ms with the default framelength being $l_{\mathrm{filt}} = 500$ ms. Also, the number of possible feature vectors was increased from 8 to 16 since one additional feature is available, resulting in more possible combinations. This is also why the maximal number of neurons n_{neur} was increased to 10 instead of 8 (twice the number of features).

Table 3.1 – Optimal feature vector $\mathbf{q}_j$ and number of hidden neurons n_{neur} for every subjective scale in the extended PEASS model

	$\mathbf{q}_j$	n_{neur}
OPS	$[q_j^{\mathrm{overall}}, q_j^{\mathrm{target}}, q_j^{\mathrm{onset}}]$	2
TPS	$[q_j^{\mathrm{overall}}, q_j^{\mathrm{target}}, q_j^{\mathrm{onset}}]$	7
IPS	$[q_j^{\mathrm{overall}}, q_j^{\mathrm{interf}}, q_j^{\mathrm{onset}}]$	3
APS	$[q_j^{\mathrm{overall}}, q_j^{\mathrm{target}}, q_j^{\mathrm{interf}}, q_j^{\mathrm{artif}}]$	2

Results

Results of the optimisation process indicate that the extended PEASS version achieved the same prediction accuracy as the optimised basic version with an average of 0.90 over all tasks. The optimal parameters were the same as the ones for the optimised basic version, except for the number of neurons n_{neur} and the feature vector $\mathbf{q}_j$ for every task which were adapted to the new features.

The optimal framelength for the decomposition filter was found to be $l_{\mathrm{filt}} = 500$ ms when considering the average prediction accuracy over all 4 subjective scales. The detailed impact of l_{filt} is depicted on figure 3.4. The performance of TPS and APS seems to decrease monotonically with l_{filt}, while the performance of OPS seems to increase with l_{filt}. IPS on the other hand presents a peak at l_{filt} before decreasing for $l_{\mathrm{filt}} = 1000$ ms. The average accuracy therefore stagnates more or less up to $l_{\mathrm{filt}} = 500$ ms before decreasing. Note that variations in accuracy of ± 0.01 can be explained by the random initialisation of the ANN training. Since the average accuracy was slightly better for $l_{\mathrm{filt}} = 500$ ms than for $l_{\mathrm{filt}} = 75$ ms, the former was chosen as optimal value, since it also reduces the computational cost of the filtering process.

The feature vectors $\mathbf{q}_j$ and the number of hidden neurons n_{neur} associated with every subjective scale is given in table 3.1. The onset misallocation degradation component is used in the OPS as well as in the TPS and IPS scales. This might be explained by the fact that onset misallocations may be perceived as target distortions in the case where the onset of the target source is allocated to the residue and as interference when onsets of other sources are allocated to the target. Note that the subjective experiment from PEASS did not consider any onset misallocation degradations in the anchors. They may however have occurred for the real-world test sounds.

3.4 Conclusion

The PEASS model provides a way of integrating psychoacoustic knowledge into objective ASS evaluation. The decomposition components of the error from ASS presented in chapter 1 are used as inputs to an auditory model and their perceptual salience is assessed. The resulting features are mapped to a subjective scale through a nonlinear function provided by a 2-layer ANN. In the optimal case, this allows for a prediction accuracy of 0.909 with the

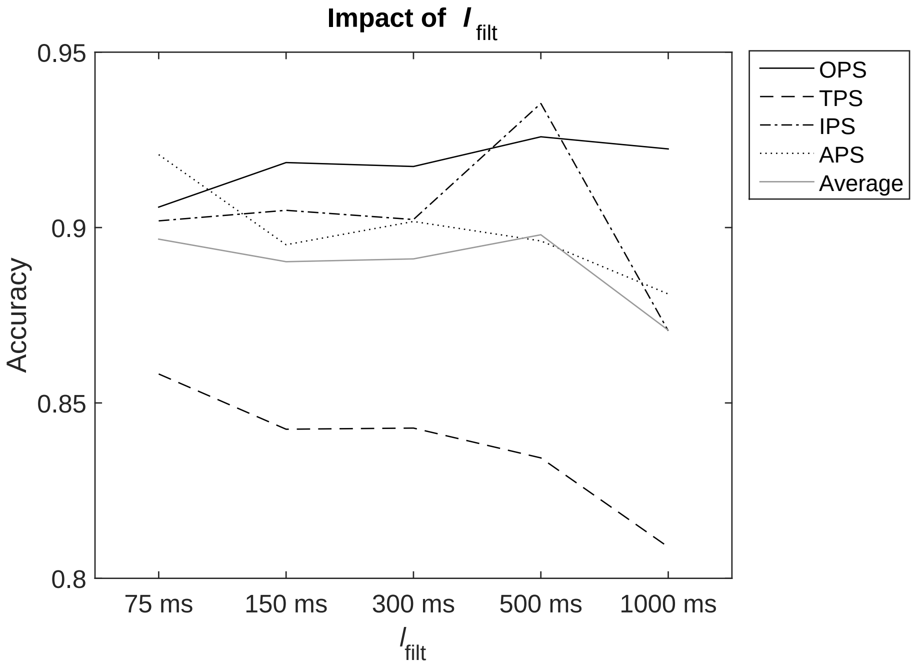


Figure 3.4 – Accuracy as a function of the framelength l_{filt} for the four different subjective scales as well as the mean performance given the optimal parameters for the rest of the PEASS model.

state-of-the-art model.

The same accuracy is achieved with the subjective PEASS data when using the extended decomposition presented at the end of chapter 1 with an additional onset misallocation degradation component. Since an additional parameter and an additional feature were introduced in the extended model, an optimisation process was followed to gain insight about the optimal parameter combination. Results show that the onset misallocation degradation component is used after optimisation in three of the four subjective scales from the PEASS subjective experiment.

The fact that the same performance is achieved in the case without spatialisation paves the way to the use of the extended model in the 3D audio application case. In fact, onset misallocation degradations are thought to be important in the 3D audio case, where a misallocated onset may result in a rapid switching of one source from one location to another, increasing its impact on perceived quality.

The PEASS model in itself is however computationally very complex and adding a degradation component does not simplify it. Moreover, there is no psychoacoustic knowledge about spatial hearing that was integrated into the model. This may not work in its favor in 3D audio scenarios, but it will provide a base score against which alternative models can be evaluated.

4 3D Audio

4.1 Introduction

We live in a 3-dimensional world. Translated to the acoustical domain, this means that sounds are produced in a precise spatial location. They then propagate through and interact with a 3D environment, before reaching the ears of a listener. Their brain analyzes the inbound sound waves as presented in chapter 4.2 and correlates the extracted cues with visual information to gain an understanding of the source of the sound, its location and their surroundings.
While this occurs very naturally in everyday situations, it's less obvious in a setting where a 3D sound scene has to be reconstructed artificially. The cues used by human audition then need to be introduced into the reproduced sound signals, or the intended effect will be missed.

In practice, playback is controlled through the use of appropriate electro-acoustical means such as headphones or loudspeakers, allowing for an appropriate control of the synthetic auditory cues. Stereophony, as reproduced through loudspeakers or headphones, is a first step towards such a scenario, through the use of level differences between the loudspeakers. However, the set of possible locations of sound sources in stereophonic reproduction is limited to the line connecting two speakers, even though this may be somewhat extended through the use of techniques such as Vector Base Amplitude Panning (VBAP) [102].
Other techniques, such as binaural synthesis, Wave Field Synthesis (WFS) or Higher Order Ambisonics (HOA) aim at a "natural" reconstruction of the sound field inbound to the listener's ears, in the sense that the artificial sound field should be as similar as possible to the sound field that would occur for the same scene in real life, *i.e.* in free-field listening.

For auditory experiments, a controlled environment needs to be achieved to avoid changing conditions from one listener to another and in order for the research to be as reproducible as possible. This chapter will give an overview of the 3D audio techniques that were used to present listeners with such an environment for the experiments conducted to validate the proposed models. First, a short introduction to spatial listening will be given, followed by the presentation of binaural synthesis and of a practical 3D WFS formulation and its perceptual validation that was conducted as part of this thesis.

4.2 Spatial hearing

While the models presented in chapter 2 model the ear roughly from the ear canal to the auditory nerve, spatial hearing is related to the physics of the incoming sound wave in the way it interacts with the human body. This includes the outer ear and how the information that is impregnated on the signal during that process is extracted and processed.

Spatial hearing is essentially related to the ability of localising sound sources in a three-dimensional environment. Many of the mechanisms involved in spatial hearing are tied to the use of both ears - binaural hearing. Even though people with hearing impairments may have some spatial hearing capability, this is out of the scope of the presented work. All of the following explanations and assumptions apply to people with normal hearing.

4.2.1 Basic physical description

Outer ear

As described in section 2.3, the outer ear is composed of the pinna (the “visible” part of the ear) and the auditory canal (depicted on figure 2.1). Its main role is to allow sound to travel towards the middle ear whilst physically protecting the more sensitive parts such as the tympanic membrane. The structure of the outer ear however - especially that of the pinna - plays a major role in human sound source localisation. Inbound sound waves are diffracted by the torso, the head and the pinnae, altering the sound pressure when comparing to the situation without a listener [108]. Since diffraction depends on the size of an obstacle compared to the wavelength of incoming sound, the effect depends on frequency. This enables the localisation of sound sources, because different directions of arrival will result in different obstacles on the acoustic pathway between the sound source and the tympanic membrane.
For every direction of arrival, a pair of Head Related Transfer Functions (HRTFs) can be established by measuring the transfer function between the sound pressure without the influence of a listener’s presence and the sound pressure at the listener’s ears. Since every listener has a different physiology, HRTFs are individual.
Spatial hearing is commonly assumed to be related to 3 cues in the acoustic pressure signals at the ears:

- Interaural Intensity Difference (IID) (which is related to the Interaural Level Difference (ILD))
- Interaural Time Difference (ITD) (which is related to the Interaural Phase Difference (IPD))
- HRTF high-frequency spectral cues

The following sections give an overview of those cues and their respective use in spatial hearing. Even though a basic explanation will be provided, the interested reader may refer to

Blauert [14] for a very complete review.

Interaural Intensity Difference (IID)

The IID measures a difference in intensity (or level in the case of ILD) of the acoustic wave between the two ears. This is mainly due to two reasons: the difference in location of the two ears and the head being a natural obstacle in the acoustic pathway between a sound source and the ear. Since the two ears are spatially separate, sound that does not originate from a sound source on the median plane (the vertical plane cutting the head from front to back) will have a longer propagation path towards the ear that is further away from the source. Since the sound intensity is inversely proportional to the square of the distance to a sound source, this results in a "natural" IID between both ears. Additionally, the presence of the head accentuates the effect for the ear that is the furthest away from the sound source. The smaller the wavelength with respect to the size of the head, the more pronounced this effect is [108]. If $\underline{A}(f) = |A(f)|e^{-j\phi}$ denotes the Fourier transform of the interaural transfer function $a(t)$, its magnitude relates to the IID; and the ILD relates to $20\log|A(f)|$ [14].
The IID is used essentially for azimuthal source localisation since lateral displacement of a sound source significantly changes the mean IID throughout the frequency range, while vertical displacement essentially affects the very high frequency range.

Interaural Time Difference (ITD)

If a sound source is not in the median plane, the acoustic pathway between the source and an ear is of different length for each ear. Due to the finite celerity of sound, this results in a Time-Difference of Arrival (TDOA) of the sound waves between the two ears. This is also referred to as ITD. In stationary regime, this corresponds to a phase delay between the signals at both ears, that is called IPD [108]. Since there is an ambiguity when the IPD surpasses π, this can only be used for frequencies where the difference in length of the acoustic pathway is less than a wavelength. Rossi [108] estimates this frequency limit to be around 800 Hz.

When modelling ITD, different approaches with different complexities can be adopted. The simplest model would be to approximate the head by a sphere with radius r. If a sound source is placed on a unit circle in the horizontal plane at an azimuth θ, the resulting TDOA τ_{ITD} can be expressed as a function of the cited parameters and the sound celerity c as

$$\tau_{\text{ITD}} = \frac{r}{c}(\theta + \sin(\theta)). \tag{4.1}$$

The resulting ITD as function of θ is illustrated on figure 4.1 in comparison with actual measured values. Larcher *et al.* [80] formulated the same model for a sound source on a unit sphere, where the elevation of the sound source is given by ϕ:

$$\tau_{\text{ITD}} = \frac{r}{c}(\arcsin(\cos\phi\sin\theta) + \cos\phi\sin\theta) \tag{4.2}$$

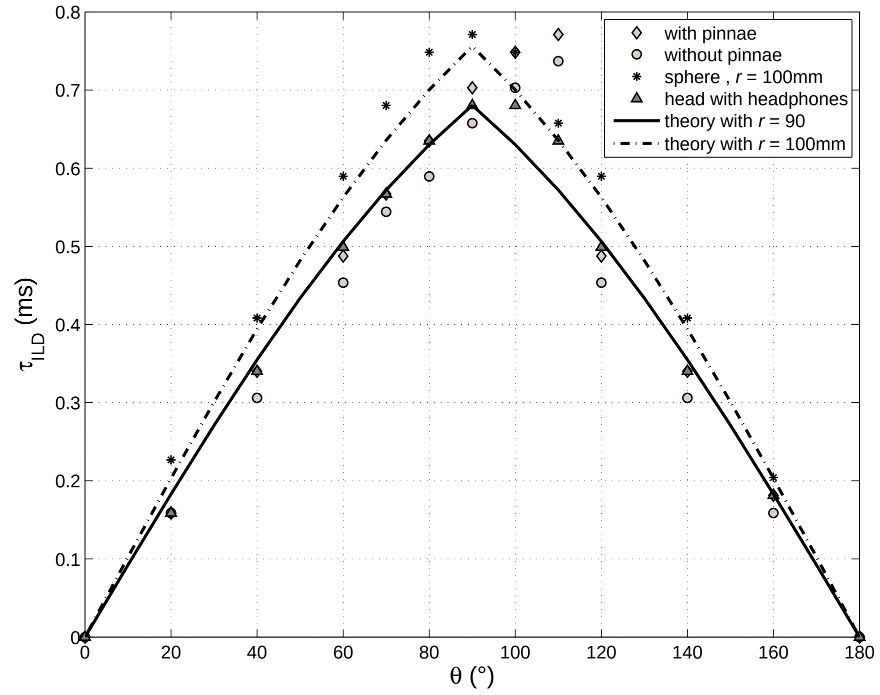


Figure 4.1 – ITD (τ_{ITD}) as a function of azimuth (θ) after eq. (4.1) adapted from Rossi [108].

There is virtually no upper limit on the complexity of head models used to approximate TDOA, but as shown by Rossi [108] and Larcher *et al.* [80], a model as simple as a sphere already gives a very good match between approximated and measured ITDs.

Head Related Transfer Function (HRTF) high-frequency spectral cues

In the high frequencies, the diffraction, dispersion and reflections due to the pinnae - and more generally of the head - result in peaks and valleys in the frequency spectrum of the signal at the eardrum above approximately 1'500 Hz [14] that are different for every ear and for every source position. These irregularities provide localisation cues to the listener [92] (sometimes called directional bands) and learning proves to play an important part in localisation using high-frequency HRTF cues [92].

4.2.2 Binaural unmasking

A psychoacoustic phenomenon that is clearly related to spatial audio is spatial unmasking. Spatial unmasking or spatial release from masking designates the decrease in effectiveness of a masker when it's spatially separated from the maskee. When binaural hearing is implied, one speaks about binaural unmasking. Spatial unmasking may for example be encountered in

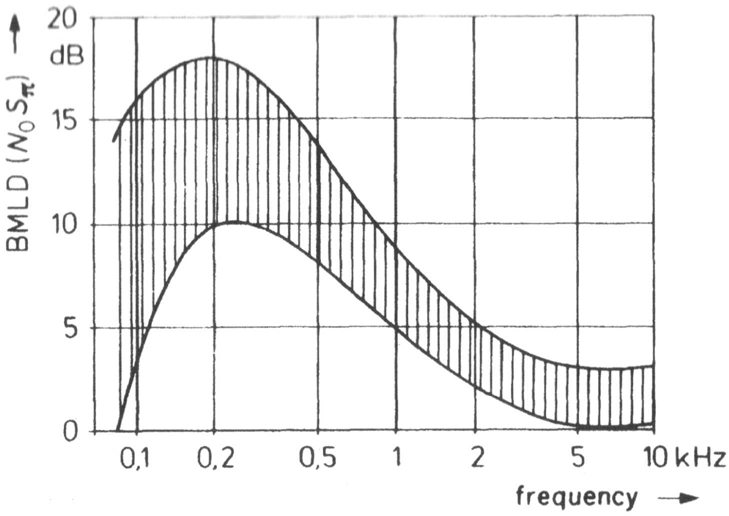


Figure 4.2 – BMLD for $N_0 S_\pi$ of a sinusoidal signal in broadband noise as a function of frequency as reported by Blauert [14]. The data is a summary from 6 different authors. Source: Blauert [14].

the so-called “cocktail party effect” [19] where a single speaker among many other speakers may be easily understood in binaural listening conditions because of its location, whereas they are much more difficult to understand when one ear is plugged [14].

The extent of this effect for different types of signals and situations is measured as the Binaural Masking Level Difference (BMLD) given by the difference between the masked threshold in monotic (or sometimes diotic) conditions and the masked threshold in binaural hearing conditions. As every masking phenomenon, the BMLD depends on the spectral and temporal properties of the signals. In the literature, the following notation is used: N and S stand for the noise and the target signal respectively. They are followed each with indices indicating the type of presentation. m stands for monotic presentation, 0 for diotic presentation without interaural phase delay, ϕ or π for diotic presentation with interaural phase difference ϕ or π, τ for diotic presentation with an ITD of τ and u for diotic presentation with binaurally uncorrelated signals.

Blauert [14] summarizes a few results from BMLD experiments: Figure 4.2 summarizes the research of 6 authors for a sinusoidal signal in broadband noise for $N_0 S_\pi$. The BMLD reaches a maximum of 10-15 dB around 200 Hz and then falls off with increasing frequency. For pulsed signals, the BMLD seems to be maximal for ITDs around 1.5-2 ms [14]. Other results suggest that the BMLD also depends on the Sound Pressure Level (SPL) and on the degree of interaural correlation of the noise [14].

In free-field listening conditions, the reference is not given by monotic conditions, but by the situation where both sources come from the same direction. Experiments summarised by Blauert [14] suggest that the BMLD increases for signal azimuths up to about 6 dB around 45° and slightly decays between 60° and 90° by about 2 dB.

In a context of Audio Source Separation (ASS), spatial unmasking might be important for the case where the estimated target is spatialised in a different location than the estimated residue. Degradations that might have been masked in the presence of the residue (masker) in the same location might be unmasked and therefore audible when moved away from it.
Shinn-Cunningham [116] postulated that there are 3 components contributing to spatial unmasking: Energetic (better-ear) effects, binaural processing and spatial attention.

Better-ear hearing

Spatial unmasking may be partly explained by so-called better-ear hearing [116]. When a source is to one side of the head, the closest ear receives more energy from that source due to the obstacle of the head as explained in section 4.2.1. This has been shown as being very helpful for the understanding of speech for example [17] and for the detection of speech-like signals [11].

Binaural processing

Even though better-ear hearing may decrease the masked threshold in some situations, it may decrease even further when using binaural processing. As Shinn-Cunningham points out, binaural models that use short-time Interaural Cross-Correlation (IACC) (like Shinn-Cunningham & Kawakyu [117] *e.g.*) detect the presence of two spatially separate sounds with way more ease than two spatially coincident sounds [116].

Spatial attention

The third component to spatial unmasking is spatial attention, which is related to informational masking. When the target and the masker are statistically similar, spatial unmasking happens in a way that cannot be explained simply by better-ear hearing. Shinn-Cunningham therefore postulates that a difference in perceived spatial locations increases informational unmasking, contributing to spatial unmasking [116].

4.3 Binaural synthesis

Binaural synthesis designates the techniques used to reproduce spatial sound at a listener's ears (typically via headphones) without the existence of recorded 3D signals (as opposed to binaural reproduction in general, where recordings can be used). Based on the description of a sound scene to be created, signals are synthesised for the right and left channels containing binaural cues for every sound source. A review of binaural technologies was written by Nicol [93].

Most often, HRTFs are used for binaural synthesis. In that case, the pair of HRTFs (or their temporal counterpart: Head-Related Impulse Responses (HRIRs)) that are the most appropriate for a given source location are convolved with the source signal and the resulting signals are played back through headphones. However, there are a few points that need to be addressed in that scenario. The HRIRs that are used, are often recorded using microphones at the opening of the ear canal or at the ear drum. Because signals can't be played back at these locations by practical means, the shift in location needs to be compensated for, as well as the transfer function of the emitter (*i.e.* the headphones) [93]. This is done through inverse filtering by a Headphone Transfer Function (HpTF). Proper HpTF measurements are quite difficult however, as pointed out by Nicol [93], due to variations in the positioning of the headphones between listening sessions and individual pinna morphology. While the latter can be addressed by using individual HpTFs, the former has no real solution. Usually, a set of HpTF's measured from different positioning settings is averaged as an approximation [93]. Additionally, to be precise about the location of the HRTF measurement point, the same apparatus should be used for HRTF measurements and for HpTF measurements [93]. This may not always be doable, especially if HRTFs from a pre-recorded database are used.

Depending on the accuracy prerequisites of the reproduction, a choice about the encoded cues can also be made. As presented in chapter 4.2, not all binaural cues provide the same information and therefore only selecting one or several cues (*e.g.* only ITD or ILD) can provide a reproduction that is accurate enough for certain needs. For example, since ITD and/or ILD provide cues in the horizontal plane, a rough approximation for binaural synthesis can be achieved by adding only those cues and neglecting high-frequency spectral cues to the source signals. However, this adds the risk of In-Head-Locatedness [93] and front/back confusions. For the experiment presented in chapter 6, where only two sound source locations in the horizontal plane had to be synthesised, the choice was made to use only ITD as spatial cue and use a HpTF to compensate for the transfer function of the headphones.

4.4 Wave Field Synthesis (WFS)

WFS is a sound field reproduction technique that enables the accurate reproduction of spatio-temporal properties of target sound sources in an extended listening area [9]. The classical formulation of WFS, often referred to as $2\ \frac{1}{2}$ D WFS [122], considers that virtual sources, loudspeakers and listeners are all located in the same horizontal plane, thus limiting WFS to 2D reproduction. From an "intuitive" point of view, the sound field produced by a primary sound source can be reproduced by a distribution of secondary sources as illustrated on figure 4.3.

While a 3D formulation of WFS has been proposed in the literature [91, 122], it does not account for any practical constraints as the $2\ \frac{1}{2}$D WFS does. Usual $2\ \frac{1}{2}$D implementations use a loudspeaker spacing of 10 to 20 cm which implies thousands of loudspeakers when extended to 3D.

A practical formulation for 3D WFS and its perceptual validations is presented in the following

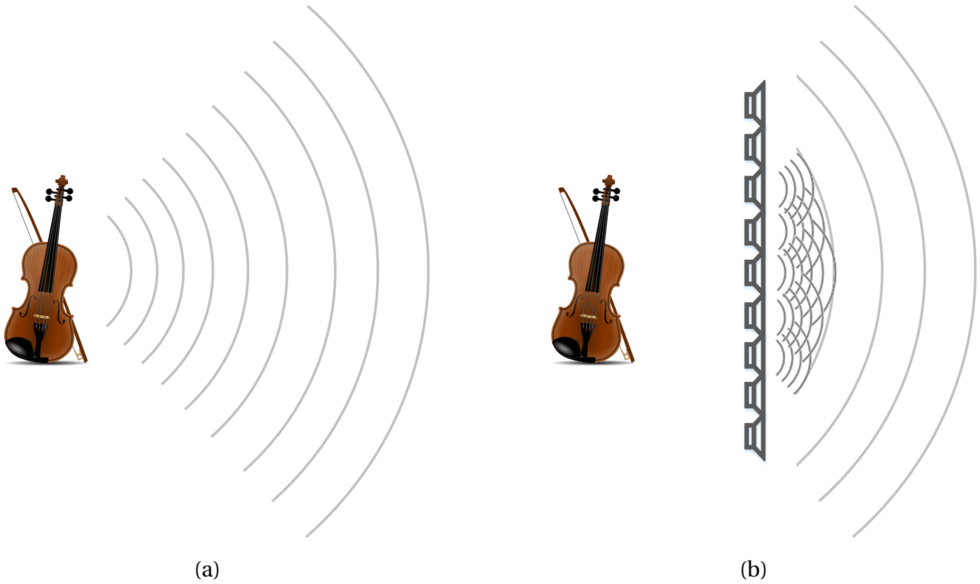


Figure 4.3 – Principle of WFS reproduction: The sound field of a primary sound source (a) can be reproduced using a distribution of secondary sound sources (b).

sections. This work was published in a journal article [107] and preliminary results were published in two conference presentations [30, 31]. Note that the 3D WFS formulation proposed here was not developed as part of this thesis, as it is the work of project partners of the i3Dmusic project (the framework of this thesis). The subjective validation of the height rendering was however conducted as part of this thesis.
In the following, bold letters refer to vectors and ω is the angular frequency.

Kirchhoff Helmholtz Integral

Wave Field Synthesis, as a boundary-based sound field reproduction technique, is based upon approximations of the Kirchhoff-Helmholtz integral [9, 136]. The Kirchhoff-Helmholtz integral provides a direct solution for reproducing arbitrary sound fields in a source free subspace V such that the pressure $P(\mathbf{x})$ at any point $\mathbf{x}$ of V can be expressed as:

$$P(\mathbf{x},\omega) = -\oint_{\partial V} P(\mathbf{x_0},\omega)\frac{\partial G(\mathbf{x}|\mathbf{x_0},\omega)}{\partial \mathbf{n}} - G(\mathbf{x}|\mathbf{x_0},\omega)\frac{\partial P(\mathbf{x_0},\omega)}{\partial \mathbf{n}}\, dS_0, \tag{4.3}$$

where $P(\mathbf{x_0},\omega)$ is the acoustic pressure at the boundary ∂V, the complementary subspace of Ω_R, on $\partial\Omega$; $\mathbf{n}$ is the inward normal vector to ∂V and G is the free field Green's function in three dimensions:

$$G(\mathbf{x}|\mathbf{x_0},\omega) = \frac{e^{-j\frac{\omega}{c}|\mathbf{x}-\mathbf{x_0}|}}{4\pi|\mathbf{x}-\mathbf{x_0}|}. \tag{4.4}$$

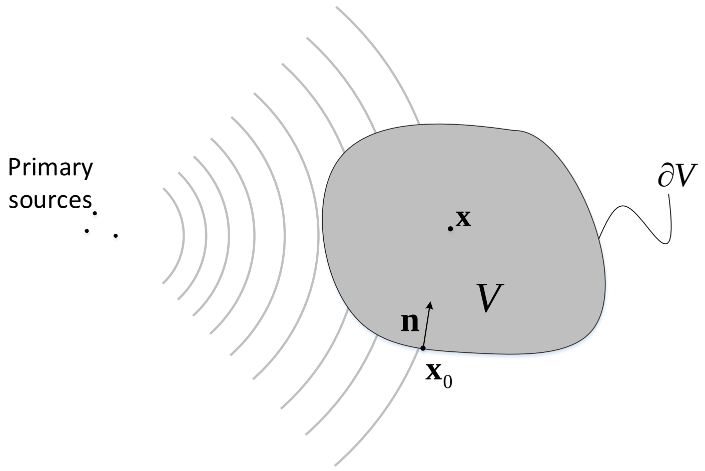


Figure 4.4 – Geometry for the Kirchhoff Helmholtz integral

The geometry of the space described above is illustrated on figure 4.4.

According to eq. (4.3), Kirchhoff-Helmholtz integral based sound field reproduction requires a continuous distribution of both omnidirectional and dipolar secondary sources located on the boundary ∂V. The so-called 3D formulation of WFS [91, 122] realizes a first simplification by selecting only omnidirectional sources:

$$P(\mathbf{x},\omega) \approx -2\oint_{\partial V} a(\mathbf{x_S},\mathbf{x_0})G(\mathbf{x}|\mathbf{x_0},\omega)\frac{\partial P(\mathbf{x_0},\omega)}{\partial \mathbf{n}}dS_0, \tag{4.5}$$

where $a(\mathbf{x_0})$ is a rectangular windowing function that selects only a subset of omnidirectional sources. Spors *et al.* specify that ∂V must be convex to prevent the unwanted components from re-entering the reproduction volume V [122].

In Wave Field Synthesis, the target sound field is often described as emitted by a so-called primary point source located at $\mathbf{x_0}$. In this case, the windowing function is expressed as:

$$a(\mathbf{x_S},\mathbf{x_0}) = \begin{cases} 1 & \text{if}\langle \mathbf{x_0}-\mathbf{x_S}, \mathbf{n}(\mathbf{x_0})\rangle > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4.6}$$

The driving function $D_{3D}(\mathbf{x_0},\omega)$ of the omnidirectional secondary sound source located at $\mathbf{x_s}$ is thus given as:

$$D_{3D}(\mathbf{x_0},\omega) = -2a(\mathbf{x_S},\mathbf{x_0})\frac{(\mathbf{x_0}-\mathbf{x_S})^T\mathbf{n}(\mathbf{x_0})}{4\pi|\mathbf{x_0}-\mathbf{x_S}|^2} \times \left(\frac{1}{|\mathbf{x_0}-\mathbf{x_S}|}+\frac{j\omega}{c}\right)e^{-j\frac{\omega}{c}|\mathbf{x_S}-\mathbf{x_0}|}\hat{S}_{sw}(\omega), \tag{4.7}$$

where $\hat{S}_{sw}(\omega)$ is the source signal. Assuming the primary source is located in the far field of all secondary sources ($\frac{1}{|\mathbf{x_0}-\mathbf{x_S}}| \ll \frac{j\omega}{c}$) and neglecting the dependency to the source signal, the driving filter $U_{3D}(\mathbf{x_0},\omega)$ can be expressed as:

$$U_{3D}(\mathbf{x_0},\omega) = W(\mathbf{x_S},\mathbf{x_0})F_{3D}(\omega)e^{-j\frac{\omega}{c}|\mathbf{x_S}-\mathbf{x_0}|}, \tag{4.8}$$

where $W(\mathbf{x_S},\mathbf{x_0})$ is a gain factor, $F_{3D}(\omega)$ is a secondary source location independent filter and the last term corresponds to a delay, expressed in the frequency domain, that depends on the distance between the primary source and the considered secondary source. The proposed formulation is thus very similar to known formulations of WFS for horizontal reproduction, so-called $2\,\frac{1}{2}$ D WFS, except that the filter $F_{3D}(\omega)$ exhibits a 6 dB per octave slope in contrast to the 3 dB per octave slope of the filter $F_{2D}(\omega)$ used for $2\,\frac{1}{2}$ D WFS [29].

This approach uses two approximations of the original Kirchhoff-Helmholtz integral that limit the rendering quality of the target sound field. First, the restriction to omnidirectional sources imposes a windowing of the secondary source distribution, thus introducing artifacts due to diffraction. These artifacts affect the sound field in a similar way to $2\,\frac{1}{2}$ D WFS. Second, the far field approximation used for the derivation of eq. (4.8) is valid mostly at high frequencies or if the primary source is located far from all secondary sources.
It is worth mentioning that this formulation of 3 dimensional WFS issued from the literature is only valid for a continuous distribution of omnidirectional sources (*i.e.* the entire surface ∂V acts as a continuum of monopolar sources). This can not be achieved in real world conditions where sound sources (loudspeakers) are discrete and present in a finite number. A practical formulation which accounts for these constraints is therefore needed. The following sections show approximations which are made to enable practical WFS.

Spatial Sampling

Any practical formulation of WFS in either two or three dimensions must include a step of spatial sampling of the secondary source distribution. In $2\,\frac{1}{2}$ D WFS, this step is simply realised by considering that loudspeakers are regularly spaced and by applying a compensation gain that equals the loudspeaker spacing in meters [130].
It is proposed here to perform a decomposition of the boundary ∂V into smaller surfaces ∂V_i such that each surface is associated with one loudspeaker only. The surface integral in equation (4.5) can then be approximated as a finite sum. The equivalent driving filter for loudspeaker i is thus expressed as:

$$U_{3D}(\mathbf{x_i},\omega) = \frac{S_i}{S}W(\mathbf{x_S},\mathbf{x_i})\hat{F}_{3D}(\mathbf{x_i},\omega)e^{-j\frac{\omega}{c}|\mathbf{x}-\mathbf{x_i}|}, \tag{4.9}$$

where S_i is the surface of ∂V_i, S is the surface of ∂V, and $\hat{F}_{3D}(x_i,\omega)$ is a modified version of the filter $F_{3D}(\omega)$ accounting for the spatial sampling. Above the so-called spatial aliasing frequency (Nyquist frequency of the spatial sampling process), the loudspeakers are not interacting in the same way as at lower frequencies and the compensation filter should be modified. This is also true for $2\,\frac{1}{2}$ D WFS [121].
The exact definition of the modified filter $\hat{F}_{3D}$ is beyond the scope of this work and has not

been disclosed by Sonic Emotion. The decomposition of the surface into smaller surfaces that are attached to a given loudspeaker may be done using triangulation methods for arbitrary surfaces or using simple sampling rules for regular loudspeakers setups and simple shapes (sphere, shoe box, ...). However, the exact calculation is not detailed here.

The effect of spatial sampling on perceived sound quality has been already addressed in $2\,\frac{1}{2}$ D WFS. Spatial sampling creates physical inaccuracies in the synthesised sound field that may lead to perceptual artifacts such as localisation bias [111, 130], increase of source width [123], sound coloration for fixed [142] and moving listeners [37]. The audibility of these artifacts for a given loudspeaker configuration mostly depends on the frequency content of the sound material [37, 111, 123]. The study presented in section refsec:elevation-perception-study aims at investigating the audibility of these artifacts in terms of localisation performance.

Simplification Strategies for 3D WFS

The previous section introduced general driving filters for 3D WFS that can be used with arbitrary loudspeaker distributions over a closed surface. The following section now proposes methods that enable the reduction of the number of loudspeakers so as to achieve 3D WFS in a practical manner.

Sampling Strategy

Methods for 3D sound reproduction such as VBAP [102] and HOA [34] often consider loudspeaker distributions that have similar density over the horizontal and the vertical dimension. In particular, HOA is best reproduced with a spherical distribution of loudspeakers with a regular sampling, even though modern approaches allow for arbitrary loudspeaker distributions (as proposed by Ahrens & Spors *e.g.* [1]).
However, the localisation capabilities of humans are known to be very different for sources located in the horizontal plane compared with sources located in elevation as explained in section 4.5.1. Therefore, it is proposed to account for this limitation by using a higher density of loudspeakers in the horizontal plane than in the vertical plane.

Reducing Loudspeaker Surface

The total number of loudspeakers can be further reduced by limiting the size of the loudspeaker surface. Such incomplete loudspeaker arrays are often used in $2\,\frac{1}{2}$ D WFS (finite-length linear arrays, U-shaped, ...). There are two main consequences of such a reduction:

- diffraction artifacts may occur but are known to cause limited perceptual artifacts [130],
- the positioning of virtual sources has to be limited in such a way that they remain visible within an extended listening area through the opening of the limited loudspeaker array. The corresponding source visibility area can be easily defined using simple geometric

criteria [27].

It is therefore possible to limit the size of the loudspeaker array for 3D WFS in a similar way by considering an open surface that may span the locations in which it is physically possible to put loudspeakers in the installation. The loudspeaker surface can be further defined by considering the subspace where virtual source positioning is required, according to the application.
In most applications, it is not possible to put loudspeakers at low elevations because they are either masked by other people in the audience or because it is simply not practical to do so. Therefore, the proposed formulation focusses mostly on loudspeaker distributions that target the reproduction of virtual sources above and around the listener. This is not a limitation of the proposed method, but rather a choice for reducing the number of required loudspeakers.

Reduction of Spatial Sampling Artifacts

Various methods for the reduction of spatial sampling artifacts have been proposed in the literature using either spatial bandwidth reduction [130], partial de-correlation of loudspeakers at high frequencies [28], stereophonic reproduction at high frequencies [142], or reducing the number of active speakers for increasing the spatial aliasing frequency in a preferred listening area [29]. All these techniques have been defined for horizontal reproduction only.

It is proposed here to extend to 3D WFS the technique proposed by Corteel *et al.* [29] for 2 $\frac{1}{2}$ D WFS. It targets the improvement of reproduction accuracy in a preferred listening area. A simple modified loudspeaker driving filter $\widehat{U}_{3D}$ can be expressed as:

$$U_{3D}(s_p, \mathbf{x_S}, \mathbf{x_i}, \omega) = \frac{S_i}{S} W(s_p, \mathbf{x_S}, \mathbf{x_i}) \times \hat{F}_{3D}(s_p, \mathbf{x_S}, \mathbf{x_i}, \omega) e^{-j\frac{\omega}{c}|\mathbf{x}-\mathbf{x_i}|}. \tag{4.10}$$

In this formulation, the origin of the coordinate system corresponds to a reference listening position located within the preferred listening area. The parameter s_p can be used to control the size of the preferred listening area around the reference position reducing the number of active loudspeakers as can be seen in eq. (4.10) and [29]. This parameter is denoted “spatial precision control” here, since this parameter affects spatial precision as will be seen in the following experiments. This parameter may be expressed in % for practical implementation. It can either be a design choice of the system installer or a parameter offered to the user of the system.
For $s_p = 0\%$, all loudspeakers of the original 3D WFS driving function in eq. (4.9) are used. This setting is referred to as “Low” spatial precision in the experimental part. Higher percentages of this parameter can be used for concentrating the rendering on a lower number of loudspeakers located around the direction of the virtual sound source. The “High” spatial precision setting of the experimental part corresponds to a value of $s_p = 70\%$ where a large number of loudspeakers remain active (see section 4.5.2).

4.5 Perception of elevation in a practical WFS implementation

4.5.1 Introduction

Virtual audio environments are generated through different techniques which generally aim at reproducing a target sound field at the listener's ears as accurately as possible. Well known techniques include WFS [9], Ambisonics [46] and VBAP [102] as well as other derived sound field control techniques (see *e.g.* Corteel [27] or Kolundžija *et al.* [77]). Every method presents its advantages and drawbacks in terms of localisation accuracy of the reproduced sources, bandwidth, listening area, number of loudspeakers, etc. The study presented here investigates the performance of the WFS implementation presented in section 4.4 in terms of a listener's ability to localize sound sources in the median plane. Performance is measured by means of localisation accuracy, localisation precision and response time for two different seating positions.

Sound Source Localisation Evaluation

Sound source localisation performance can be measured in different ways. When using an absolute localisation protocol (*i.e.* pointing at the perceived location of a source) localisation judgement data can be modelled as normal distribution, as explained by Letowski & Letowski [82]. Two types of performance indices can then be distinguished as for all Gaussian processes: accuracy and precision. Accuracy describes the location of the distribution relative to a reference, corresponding to the constant error component [82]. When applied to localisation error data, it corresponds to a localisation bias, *i.e.* a difference between the mean of the distribution and the actual location of the sound source.
Precision however describes the spread of the distribution, corresponding to the random error component. Although other quantities may be considered, the standard deviation of the distribution gives a good measure of precision [82].

If a relative localisation task is used (*i.e.* compare the locations of two sources which are presented sequentially based on a forced-choice protocol), one may also consider a threshold based on a psychometric function to measure localisation acuity. Blauert for example defines localisation blur to be the Minimum Audible Angle (MAA). For a given initial source position, this corresponds to the angular distance for which 50% of the participants noticed a change in source location when moving the source away from its initial position [14]. It is however not clear if localisation acuity and localisation precision are related [88], making a comparison of results difficult. The results summarised by Blauert [14] however give a good indication of the variation in human localisation performance as a function of the direction of incidence of an auditory event. Tables 4.1 (horizontal localisation) and 4.2 (vertical localisation) show localisation blur data as reported by Blauert [14]. Localisation performance is therefore best in front of the listener, decreasing for lateral azimuths and for increasing elevation.

Response time was also used to measure localisation performance by Guillon [49]. Participants

Table 4.1 – Localisation blur Δ_ϕ in the horizontal plane as a function of the azimuth θ for white-noise pulses as summarised by Blauert [14]

θ	Δ_ϕ
0°	±3.6°
90°	±9.2°
180°	±5.5°
270°	±10.0°

Table 4.2 – Localisation blur Δ_ϕ in the median plane as a function of the elevation ϕ in front ($\theta = 0°$) and in the back ($\theta = 180°$) for continuous speech by a familiar person as summarised by Blauert [14]

θ	ϕ	Δ_ϕ
0°	0°	±9°
0°	36°	±10°
0°	90°	±13°
180°	36°	±15°
180°	90°	±22°

were asked to turn their nose towards the perceived location of a sound source while the position of their head was tracked. The time until a stable position was reached was measured and analysed as performance index. The measurement of this cue should therefore allow to gain an additional insight on the localisation performance of human listeners in 3D WFS.
Note that all of the evaluation methods cited above refer to the human ability to locate one or several sound sources. In any situation where localisation performance is measured, an audio rendering system which may introduce additional errors (i.e. with its own accuracy and precision) is involved (see *e.g.* the limitations found by Blanco-Martín *et al.* [12]). When measuring localisation performance, the measured error components are therefore always the results of two phenomena: human localisation uncertainty and rendering acuity of the audio system. Depending on the experiment (*e.g.* when using a single point-like sound source), the audio system may play only a minor role and be ignored. In the context of WFS however, the phenomena are not separable and sum up in the measured error components.

In the following sections, the focus lies on vertical localisation, since the novelty of the employed spatialisation technique is to enable 3D rendering in WFS. Such a system will however never be employed in an environment where all listeners are at fixed positions without moving their head. Head movement was therefore not restricted. The participants could rely on a full set of localisation cues, potentially giving them an increased localisation accuracy when comparing to studies with a fixed head position (see Blauert [14] for a review). Given the localisation task and the chosen protocol, localisation accuracy and precision as well as the response time are analysed.

Reporting Method

Sound source localisation experiments may be biased depending on the chosen reporting method. Different methods have been applied in the literature. Oral reporting, such as the "absolute judgement" technique used by Wightman and Kistler [141] have the disadvantage of necessitating extensive training. Another solution may be head-tracking and asking the participants to turn their head towards the location at which the sound source is perceived such as employed by Makous and Middlebrooks [84]. This may however introduce errors due to the lag which is introduced by headtracking devices and it may be quite uncomfortable for locations in the median plane. Listeners also have no way of knowing if they are actually pointing to the desired location since no feedback whatsoever is provided. Oldfield and Parker previously had used a special gun to point at perceived locations [94] and photography to record the answers of the participants. Besides being impractical for the rear quadrant (listeners had to aim through their head), such visual pointing methods have the disadvantage of possible mismatch between the visual and auditory modalities [76].
More recently, Pulkki and Hirvonen used an auditory pointer in the form of a loudspeaker mounted on an arm which could be moved [103]. This method has the advantage of providing immediate feedback and not being multimodal. This method was later extended to virtual audio sources by Bertet *et al.* who used a virtual source as auditory pointer which can be controlled by the participant [10]. The task was then to align the pointer location to the perceived location a physical sound source (loudspeaker). Given the advantages of this method and the similarity of the task at hand, the present study uses this reporting method, submitting the participants to a source location matching task with a physical reference loudspeaker used as target (see section 4.5.2).

Objectives

The present study aims at gaining an insight into the spatialisation performance of a practical implementation of a 3D WFS algorithm. 3D WFS systems provide a mean of placing virtual sources all around and especially above a target listening area. The vertical localisation performance in a 3D WFS setup is investigated by conducting a localisation experiment. Based on WFS theory which states that ideally any virtual source position may be synthesised accurately and based on the fact that the proposed simplifications will introduce some degradations while conserving a certain degree of precision and accuracy, it is expected that the participants are able to discriminate between several source elevations and that there is no influence of a participant's position on localisation performance. Moreover, the chosen rendering algorithm introduces a spatial precision parameter, which should additionally increase localisation performance by reducing the number of active loudspeakers as long as several loudspeakers remain active (see section 4.5.2 for the choices made). Performance is also measured by means of the response time of the participants, giving an additional cue on the difficulty of the task at hand. It is expected that the more accurately a source location is perceived, the quicker the localisation task will be accomplished. Localisation precision is also expected to increase when the spatial precision is decreased.

4.5.2 Method

WFS System

A WFS setup was installed in a listening room of 6.70 x 6.80 x 2.60 m. The mean reverberation time of the room was measured to be about 0.25 s and flat below 5.3 kHz and decaying for higher frequencies to reach 0.18 s at 16 kHz, which is similar to studio conditions. The background noise level of the room was measured to be approximately 23 dB(A) (1 second integration period, averaged over 2x 10 minutes of measurement). 3 of the walls are coated with absorbing materials (mineral wool covered with tissue), the floor is entirely covered with carpet and the ceiling is acoustically treated. The fourth wall contains windows. Even though no measurements for early reflections which potentially influence perceived location [106] were made, the configuration of the room and the covering of the windows with heavy curtains should minimize the influence of the room.

The WFS rendering system is composed of 24 ELAC 301.2 loudspeakers, which are distributed as illustrated on figure 4.5: two horizontal rows 9 and 7 loudspeakers at heights 0 m and 1.20 m respectively relative to the position of a listener's head (brightest grey rows) and a ceiling over which the remaining eight loudspeakers are distributed in two other rows (darker grey rows). The loudspeaker setup therefore covers an azimuthal range of roughly 90° ($-45° \leq \theta \leq 45°$) and an elevation range of 90° $\left(0° \leq \phi \leq 90°\right)$ in front of the listener ((θ, ϕ, r) being spherical coordinates).

Eight additional loudspeakers, not contributing to the WFS are mounted on the setup to serve as potential targets (white squares on figure 4.5). That said, all WFS loudspeakers can also be used separately and serve as target as well. To avoid any visual influence, the setup was hidden by acoustically transparent curtains.

The 3D WFS algorithm is implemented on a Sonic Wave 1 3D sound processor[1], which delivers the loudspeaker driving signals to 4 sonic emotion M3S amplifiers through a RME ADI-648 MADI to ADAT converter. All software components, commands and stimuli were generated with MATLAB® on a PC connected to a MOTU HD-896 soundcard.

The sound processor allows for a low shelf, a high shelf and 3 parametric equalizers on every output (*i.e.* every loudspeaker). Manual measurement of the output spectrum at the center of the setup for each loudspeaker in combination with these equalizers was used to compensate for room coloration.

The set of possible virtual sources is located at a constant distance of $r = 5.4$ m with respect to the center of the system and could be controlled in elevation with a precision of $\sim 1.67°$. In this study, all sources are considered as being located on the median plane, at an azimuth of 0°.

[1] http://www.sonicemotion.com/professional

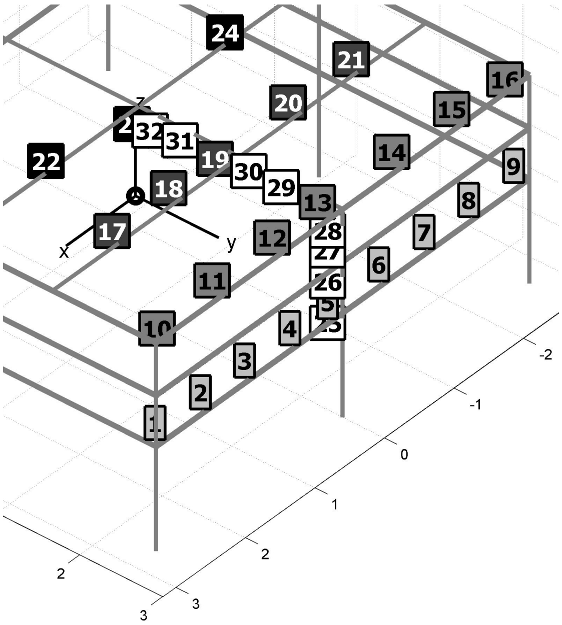


Figure 4.5 – Loudspeaker setup at EPFL - Squares are loudspeaker positions whereas lines are aluminium tubes of the rack stand. The black spot represents the position of a participant's head at a centered position.

Since the implemented method allows different values of the spatial precision parameter, two settings are used: in a first setting, there is no restriction in spatial precision ("low" precision), resulting in spatially broad perceived virtual sources, whereas in the second setting ("high" precision), spatially precise rendering is targeted. The "high" precision setting was determined considering a preferred listening area of 2 m diameter around the center point of the installation. Figure 4.6 provides the number of "active" speakers depending on the virtual source elevation and spatial precision setting. It can be seen that in the "low" spatial precision setting, 15 or more loudspeakers contribute to the virtual source rendering for nearly all source elevations. The number of active speakers is only related to the source visibility criterion. For the "high" spatial precision setting, the number of active speakers remains large: around 10 between 0 and 35 degrees, around 7 up to 60 degrees and gradually reducing to 1 around 90 degrees (the "voice of god" speaker). It should be noted that the number of active speakers is given as the number of speakers having a relative level between the level of the loudest speaker at the given source elevation and 15 dB below that. It can be regarded as the number of speakers that significantly contribute to the sound field reproduction.

Experimental Task

Since visual or motional reporting of perceived location is subject to sensory bias, an auditory pointer was used as employed by Bertet *et al.* [10]. The task of the participant therefore consisted in matching the perceived location of a pointer source (rendered with 3D WFS)

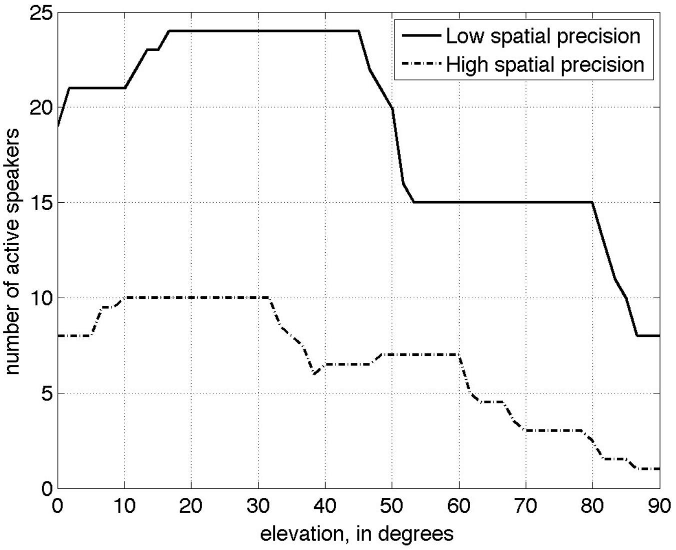


Figure 4.6 – Number of active speakers depending on target elevation for high and low spatial precision settings. The number of active speakers corresponds to the number of speakers having a driving signal level between that of the loudspeaker receiving the maximum level for a given source position and 15 dB below that.

with the perceived location of one of the target sources (physical reference loudspeaker). The pointer source could be moved in elevation with the arrow keys of a computer keyboard by increments of 1.67° between $\phi = 0°$ and $\phi = 90°$. The participant was free to switch between the target and the pointer sources and had no time limit to fulfill the matching task. They could store the pointer elevation by pressing "Enter" on the keyboard as a confirmation of their estimate.

Stimuli

Amplitude-modulated pink noise was used as stimuli for both target and pointer sources. By employing time-varying broadband noise, maximum localisation cues should be provided to the participant to minimize confusion, since localisation has been shown to improve with increasing bandwidth (see *e.g.* King & Oldfield [75]) and different modulation frequencies enable the participant to distinguish between the two stimuli. The target signal was modulated at $f_{\mathrm{mod,target}}$ = 15 Hz whereas the pointer signal was modulated at $f_{\mathrm{mod,pointer}}$ = 20 Hz. The amplitude modulation depth was d_{mod} = 50% in both cases.

In order to minimize the influence of timbre during the matching task, in addition to equalising the loudspeakers, the target signal was high-pass filtered using a second-order Butterworth filter with $f_{3\mathrm{dB}}$ = 500 Hz. The two stimuli therefore could be easily distinguished and the participants could not rely on timbre to match the locations. To avoid any additional bias, the two stimuli were subjectively adjusted to present equal loudness.

Experimental Design

The experiment was split into two parts, differing by the listening position of the participant. In the first part, the listener was seated at the origin of the coordinate system (center of the setup, see figure 4.5), facing the loudspeaker setup. For the second part, the listening position was moved 1 meter to the left, but the listener's orientation was kept constant. Each part was composed of 5 runs. In each run, 10 trials (5 target elevations x 2 spatial precision settings) were presented in random order. The initial elevation of the pointer source was randomly set for each trial (*i.e.* each target/pointer pair).
To prevent edge effects (*i.e.* bias in the perceived location due to the sound field not being rendered completely when the virtual source is on the edge of the valid rendering domain), 5 central loudspeaker positions were tested as targets, defined by their elevation: $\phi_{target} = \{14°, 26°, 36°, 43°, 58°\}$ corresponding to loudspeaker numbers $\{27, 13, 30, 19, 31\}$ on figure 4.5. Two of the target sources therefore were part of the WFS system (numbers below 25) and the three others weren't.

Each participant was instructed to feel free to move their head. At the beginning of the experiment, each participant had to complete at least one training trial to understand the task. The two parts took place at different times (3-4 months apart), but with the same panel of participants. Within each part, there was no break between runs, but the participant was free to have a break during the experiment once. Each part of the experiment took approximately 30 minutes per participant and the participants needed 25.9 seconds per trial on average to complete the elevation matching task.
11 participants, 2 women and 9 men between ages 22 and 38 (M = 28.4, SD = 5.6), took part in the study. The panel was composed of master and PhD students, as well as post-doc researchers at EPFL, including two of the authors of the paper publishing this experiment [107]. 7 of them had already heard the spatialisation system in a different context. None of the participants was compensated in any way for the experiment. They all reported normal hearing although no audiometric measurement was made.
The experimental design in this case was a repeated measures design with 4 factors: the target elevation (5 levels), the spatial precision setting (2 levels), the seating position (2 levels) and the repetition (5 levels).

4.5.3 Results

Analysis

The pointer source elevation at the end of each trial was recorded. Additionally, the history of pointer source movement over time was recorded for each trial. 3 measurements out of 1100 in the available data set were discarded during post-screening, because participants pressed the "Enter" key twice and therefore skipped one trial.

In a first part, an analysis of the variance of the localisation data is conducted to test the data

Table 4.3 – Statistical analysis of effects on matched source position

Effect	F	p
REFS	619.2	$< .001$
PREC	0.3	.571
POS	10.9	$< .005$
TIME	2.5	$< .05$
REFS×PREC	9.7	$< .001$
REFS×POS	3.0	$< .05$
PREC×POS	0.1	.758
REFS×TIME	1.0	.431
PREC×TIME	0.1	.981
POS×TIME	1.0	.421

for influences of the following factors: listening position POS ("centered", "1m to the left"), target source number REFS (27, 13, 30, 19, 31), spatial precision PREC ("low", "high") and repetition TIME (1,2,3,4,5). The dependent variable is the matched source elevation.

To perform the analysis, a mixed model is considered with all factors being modelled as fixed effects (see appendix A for a short introduction to linear mixed models). Since the order of presentation of the different runs is random, the covariance matrix is assumed to have a compound symmetry structure. Covariance is therefore assumed being the same between any two measurements and variance being the same for every measurement. A model with a covariance matrix having a heterogeneous compound symmetry structure was also considered. Such a model would allow for different variances for every measurement while keeping the constant covariance assumption. However, it showed no improvement over the previous model. The value of the Akaike Information Criterion (AIC) was slightly smaller in the second case, but the value of the Bayesian Information Criterion (BIC) was larger. Since the BIC penalizes the estimation of a too large number of parameters, the simpler model with the compound symmetry covariance matrix structure was kept. For all pairwise comparisons that are made during the analysis, the Sidak correction is used to account for multiple comparisons. The significance level is set to $\alpha = 0.05$.
Main effects and two-by-two interactions are tested. The results of this analysis are given in table 4.3.

Localisation Accuracy

The analysis reveals that the source number (*i.e.* the target source elevation) REFS has a significant effect on the mean reported source elevation ($p < .001$). This means that different elevations were globally reported for different reference loudspeakers, confirming the good functioning of the rendering method. Pairwise comparisons using the Sidak correction to account for multiple comparisons are made between reference loudspeaker levels to test if

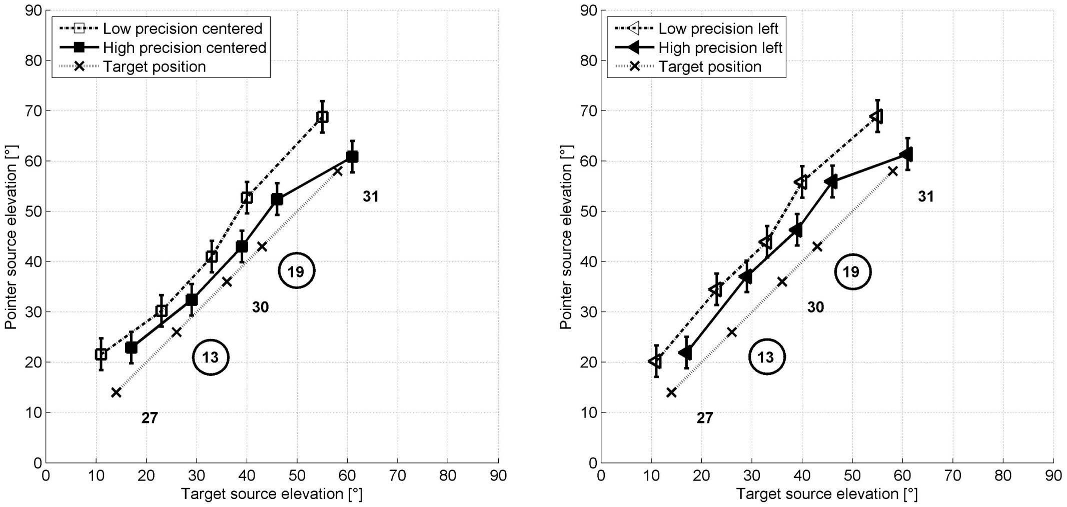


Figure 4.7 – Localisation accuracy: Estimated marginal means and 95% confidence intervals of the matched elevation data at the centered listening position (left) and at the listening position 1m to the left (right). Circled numbers correspond to loudspeaker positions which are part of the WFS system.

matched source location levels are well distinguished one from each other in every situation (seating position and spatial precision combination). The differences prove to be statistically significant in all situations. Significance levels are $p < .001$ for all differences, except for the difference between sources 19 and 31 with the "high" spatial precision at the left seating position, where significance level is $p < .05$. Five levels of target elevations between 14° and 58°, even for inter-elevation differences as small as 7° (between 36° and 43°) could therefore be distinguished in any situation. The estimated marginal means and the corresponding 95% confidence intervals are shown on figure 4.7 for the centered and the left listening positions.

A significant effect of the listening position POS is also reported ($p < .005$). Participants globally set a 2.0° higher elevation ($p < .005$) when they are seated at the left listening position compared to the centered position. No main effects are reported for the spatial precision PREC ($p = .571$). The analysis also shows a significant effect of the repetition number factor (TIME, $p < .05$). Participants globally report lower matched source locations as the repetition number increases. Pairwise comparisons however show that the difference between matched source locations is significant only between repetitions 3 and 5 ($p < .05$). Participants set the virtual source 2.86° lower during the 5th repetition than during the 3rd repetition. All other comparisons are not statistically significant ($p > .15$).

The analysis also reveals that two interactions are statistically significant. The first one is between the reference source number and the spatial precision (REFS×PREC, $p < .001$). A quick inspection of figure 4.7 shows that the slope of the curves is getting smaller for source 31 when using the high spatial precision, whereas it seems to remain constant when using the

Table 4.4 – Statistical analysis of effects on mean standard deviation of matched source position

Effect	F	p
REFS	0.6	.664
PREC	20.7	$< .001$
POS	0.8	.379
REFS×PREC	3.5	$< .01$
REFS×POS	3.1	$< .05$
PREC×POS	1.7	.191
REFS×PREC×POS	1.4	.248

low precision. This is confirmed by the fact, that there is a significant difference between the matched elevations for target source 31 when using the "high" spatial precision, as compared to using the "low" spatial precision. The difference in matched elevation is 7.7° for that case ($p < .001$). Other pairwise comparisons were not statistically significant ($p > .05$).

The second significant interaction is between reference source number and seating position (REFS×POS, $p < .05$). This would translate into different slopes of the response curves between seating positions if the spatial precision parameter was ignored in figure 4.7. Pairwise comparisons of matched source locations between left and centered seating position were not all significant which prevents any further comment. All other two-by-two interactions were not statistically significant.

Another fact that may be worth mentioning is that there is a systematic bias in the matched source elevation with respect to the target elevation. The average matched elevation being 8.2° higher than the corresponding reference source. If broken down by seating position, the bias is 7.2° for the centered listening position and 9.2° for the left listening position.

Localisation Precision

Localisation precision is given by the standard deviation (SD) for each participant / target / precision combination, which was computed as a new dependent variable. To evaluate the impact of the spatial precision parameter on the localisation precision, a new analysis is performed with this new dependent variable with target source number REFS, listening position POS and spatial precision PREC as factors. Main effects are computed as well as all possible interactions. The parameters of the model remain the same as for the analysis on the reported localisation.

The results reported in table 4.4 show that there are 3 significant fixed effects at the .05 level. Firstly, the spatial precision significantly contributes to enhance localisation precision ($p < .001$). Estimated marginal means reveal that when the spatial precision is set to "high" (estimated mean SD: 6.3°), the SD is 2.3° smaller on average than for a "low" setting (estimated

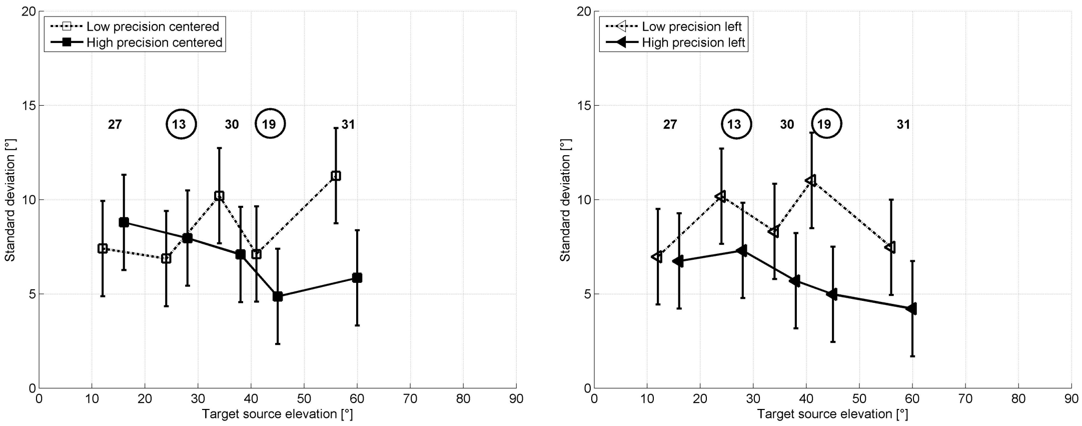


Figure 4.8 – Localisation precision: Estimated marginal means and 95% confidence intervals of the standard deviation data at the centered listening position (left) and the listening position 1m to the left (right). Circled numbers correspond to loudspeaker positions which are part of the WFS system.

mean SD: 8.7°) and the difference between both is significant at the .001 level.

The two other statistically significant effects are two interaction effects: reference source number with spatial precision (REFS×PREC, $p < .01$) and reference source number with listening position (REFS×POS, $p < .05$). For the first of these two interaction effects, a comparison by pairs (splitting up the effect for each reference source number and comparing between “low” and “high” spatial precision) reveals that the difference in SD is strong for the highest 3 reference sources ($p < .05$) and not significant for the lowest 2 reference sources. This interaction can also be seen on figure 4.8, where the curves following the results of the “low” spatial precision setting globally have a different slope than those who follow the results of the “high” setting.

On the other hand, even though the interaction effect between listening position and reference source number is statistically significant, a comparison by pairs (splitting up the effect for each reference source number and comparing between left and centered listening positions) shows only a significant difference in SD for the highest reference source ($p < .05$). All other pairs do not show statistically significant SD differences.

Response Time

The same analysis that was run on the reported matched locations (section 4.5.3) was run on the response time RESPTIME dependent variable. Table 4.5 reports the main effects and two-by-two interactions.

Three effects are shown to have a statistically significant influence on the average response time: the spatial precision (PREC, $p < .001$), the repetition (TIME, $p < .05$) and the interaction between the reference source number and the spatial precision(REFS×PREC, $p < .05$).

Table 4.5 – Statistical analysis of effects on response time

Effect	F	p
REFS	1.2	.289
PREC	37.9	$< .001$
POS	0.8	.367
TIME	3.3	$< .05$
REFS×PREC	3.3	$< .05$
REFS×POS	0.7	.591
PREC×POS	0.0	.842
REFS×TIME	1.0	.391
PREC×TIME	0.6	.684
POS×TIME	1.0	.425

In this study, a low spatial precision setting is potentially detrimental to accurate localisation, which is confirmed by the variation of the participants' response times as a function of the spatial precision setting. A pairwise comparison shows that the mean response time decreases by 4.3 s ($p < .001$) from 28.0 s to 23.7 s when using the "high" rather than the "low" spatial precision setting. This is further illustrated by figure 4.9 where estimated marginal means and 95% confidence intervals are shown for both settings. The curves with the high spatial precision setting are globally below the curves with the low spatial precision setting.

The effect of the repetition number is also statistically significant ($p < .05$). The response time therefore globally decreases with the number of repetitions. However, the mean response time difference is significant only between repetition times 1 and 3 ($p < .01$).

The last significant effect is the interaction between the reference source number and the spatial precision (REFS×PREC, $p < .05$). This is also illustrated on figure 4.9, where the slopes of the curves between the "low" and the "high" spatial precision differ.

All other effects are not significant. There is therefore no influence of the reference source number ($p = .289$) or of the seating position ($p = .367$) on the mean response time.

4.5.4 Discussion

General Discussion and Localisation Accuracy

The first observation that can be made on the results is that the implemented 3D WFS method allows to properly discriminate 5 target elevations between 14° and 58°, even for inter-elevation differences as small as 7° (between 36° and 43°). This confirms the spatial resolution of the method in a first approximation. However, there seems to be a systematic bias between the matched and the actual target positions. On average, mean values of the reported virtual source positions are 7.2° higher than the real target positions for the centered listening position and 9.2° higher for the left listening position. This bias cannot be explained in terms of the positions of the loudspeakers which contribute to the WFS array. If reported source

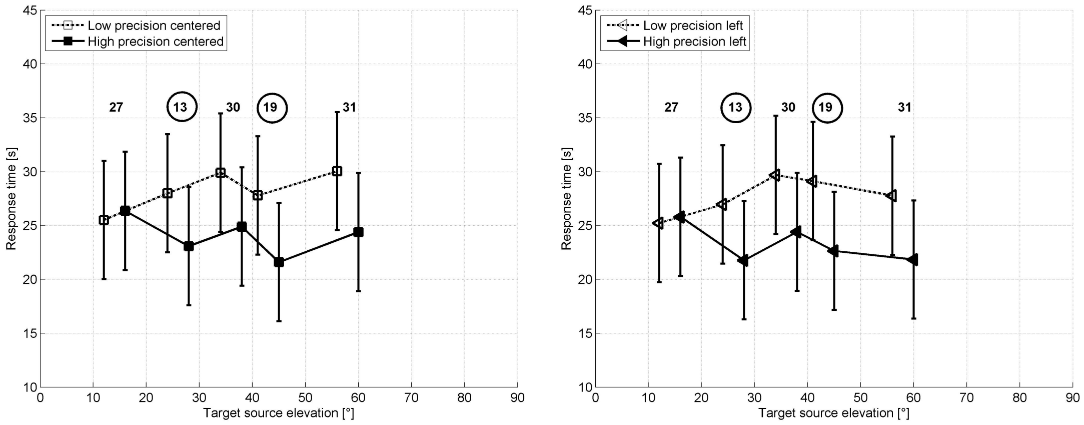


Figure 4.9 – Response time: Estimated marginal means and 95% confidence intervals of the response time data at the centered listening position (left) and the listening position 1m to the left (right). Circled numbers correspond to loudspeaker positions which are part of the WFS system.

locations were biased towards the nearest loudspeaker for example, the bias would disappear for virtual source positions which correspond to the position of a loudspeaker contributing to the WFS. This is however not the case for loudspeakers 13 and 19 which are circled on figure 4.7 where the "zero bias" line corresponds to the dotted diagonal line. The WFS system must therefore introduce this constant error, but this could be easily compensated for.
It has to be noted that the 5 levels of elevation are discriminated at both listening positions. The difference in elevation between listening positions can be ignored in practical implementations, because a localisation accuracy shift of 2° when moving over a distance of roughly one fourth of the total system width seems more than reasonable and barely noticeable in practice.

A second observation is that the interaction between the reference source number and the spatial precision parameter is always statistically significant. This can be readily explained by the setup geometry and the spatial precision parameter definition. The WFS setup is constructed such that there are fewer loudspeakers on the uppermost layers than for the lower layers. The spatial precision parameter further reduces the number of active loudspeakers. The combination of both results in few loudspeakers that are active when a high elevation is to be rendered with a high spatial precision. The matched source locations should therefore be quite precise and present almost no bias at high elevations, whereas at low elevations, the effect is less present. This is also expected to enhance localisation precision and reduce response time. This is measured by said interaction and can be seen across the results. It is noticeable however, that even though only a small number of loudspeakers may be active at high elevations with high spatial precision settings, localisation results do not vary significantly even if the virtual source location does not match the positions of a physical loudspeaker of the rendering system (*e.g.* source #19).

There also seems to be a small learning effect, since the response time and the systematic bias are both slightly reduced with increasing number of repetitions. However, the analysis shows that the performance increase is not important.

Virtual Sound Source Localisation Performance

When comparing the results to other studies, a difference has to be made between free-field localisation with physical sound sources and localisation of virtual sound sources. For physical sound source localisation, former studies report best accuracy in the frontal quadrant. Oldfield and Parker *e.g.* report 6° or less azimuthal error in the horizontal plane and 8° or less elevation error in the median plane in the frontal quadrant when using broadband white noise and a manual pointing reporting method (pointing with a gun whilst blindfolded) [94]. The limitations of a rendering system however will influence the resolving ability of human audition and therefore the results of localisation experiments. Virtual sound source synthesis can be implemented using different techniques, such as WFS, amplitude panning (vector-based (VBAP) or simple stereo phantom source imaging), Ambisonics or even binaural synthesis. Vertical localisation performance varies depending on the proposed technique and the experimental setup. Moreover, since virtual sound scenes are rarely directly compared to the original sound scene (when it exists), localisation accuracy may not be the most relevant performance index. Localisation precision on the other hand does not depend on direct comparison of two sound scenes and may therefore be a more meaningful performance index when comparing the presented results with other studies.

De Bruijn [37] studied vertical localisation using a visual pointing task, comparing vertical localisation accuracy using a dense vertical WFS array (12.5 cm spacing) against phantom source imaging (lower- and uppermost loudspeakers of his WFS array) with speech stimuli for his study. A standard deviation of ~ 7° is reported when employing the dense WFS array. Phantom source imaging was shown to be nonrobust for vertical localisation, results being close to random for small listening distances where loudspeakers appear to be spaced by more than 60 degrees in elevation. Similar degrees of localisation precision are obtained here, but with distances between loudspeakers which are much greater (smallest distance is ~ 54 cm in the horizontal dimension and ~ 105 cm in the vertical dimension).

Chung *et al.* proposed a combination of WFS and vertical amplitude panning [22]. Two horizontal WFS arrays were used to generate a third virtual WFS array by vertical amplitude panning, which was intended to generate the targeted sound field. This approach seems interesting since the number of loudspeakers is greatly reduced as it is with the proposed method. However, the results suggest that vertical localisation is very poor with vertical panning between the two horizontal WFS arrays even though the stimuli were pink noise bursts presenting all necessary cues.

Pieleanu conducted an extensive study about horizontal and vertical localisation for first- and second-order Ambisonics in her masters thesis [100]. She reports a mean localisation error of up to 13.5° and a SD of around 10° in the median plane depending on experimental conditions

when using pink noise bursts. She found no dependency of the localisation accuracy on the Ambisonics order when comparing first and second order Ambisonics.
For HOA, attempts have been made in reducing the number of loudspeakers and in tackling other practical constraints as well. One of the proposals is Mixed-Order Ambisonics (MOA) systems, as for example by Käsbach *et al.* [73]. While no localisation experiments were conducted in the cited paper, the subjective tests focusing on spatial resolution, clarity and distance perception seem to show good results. Travis reviewed the basics of the MOA technique regarding elevation rendering [129], but the simulated systems present elevation localisation errors that may easily surpass 10°.

With a mean localisation accuracy of 7 – 9° and a mean localisation precision of 6 – 9°, the results of the conducted study tend towards the performance that has been shown for physical sound source localisation. When compared to other spatialisation systems, the study shows similar performance to the dense WFS loudspeaker array and outperforms reported localisation precision of other WFS implementations and Ambisonics systems as well as phantom source imaging. Moreover, the study reports results at two listening positions which prove to show similar performance where most studies in the literature only provide results at an ideal listening position (“sweet-spot”).

Response Time as Localisation Performance Measurement

Response time measurements are a quite recent development in the field of localisation performance assessment. In fact, most, if not all cognitive tasks are subject to a so-called speed-accuracy tradeoff [96]. The more time a participant is given to accomplish a task, the more accurate a participant’s response will be and inversely, the quicker the response has to be given, the less accurate it is. Even if no instructions are given about the time in which a task has to be accomplished, a participant will make such a tradeoff which can be hypothesised to be just below optimal accuracy. So the assumption can be made that the combination between achieved accuracy and response time can give important information about the underlying difficulty of the task. Previous studies showed that factors which are potentially detrimental to accurate localisation (such as nonindividualised head-related transfer functions by Chen [18] and high system latencies by Yairi *et al.* [143]) increase the localisation response time.

In the presented study, an interesting observation is made when comparing the results while using the two different spatial precision settings. Not only a quicker response time is achieved when using a high spatial precision setting (4.3 s decrease), but a better localisation precision is also reported (better by 2.3°). In terms of a speed-accuracy tradeoff (*i.e.* a time-precision tradeoff in this case), these results are expressed as two points lying on different curves as illustrated in figure 4.10. No units are given, since the actual shape of the curves has not been measured during the experiment. The general form of the curves is however inferred from the theory and the results reported by Pachella [96]. No time limit is given for the task, so each point is situated just below optimal precision, but since the attained precision is different, two different curves have to be hypothesised. The optimal tradeoff would be located in the upper

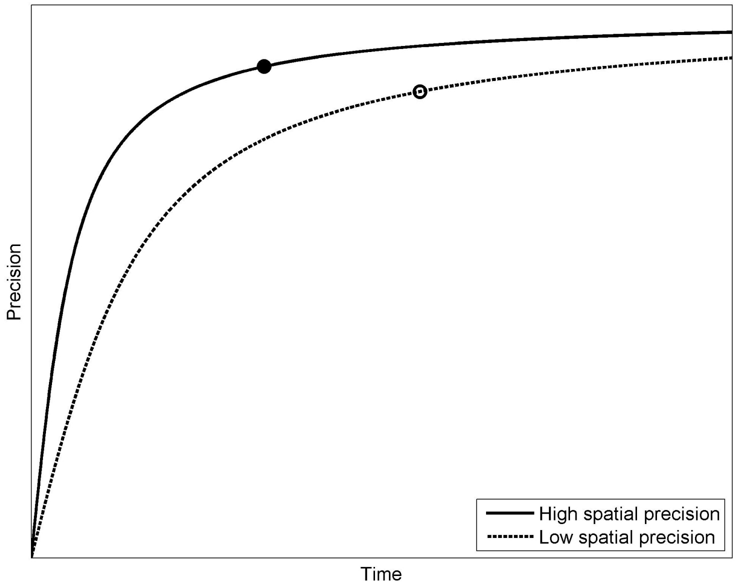


Figure 4.10 – Hypothesised time-precision tradeoff curves corresponding to the two spatial precision settings with the two points representing the reported results.

left corner, attaining optimal precision in a very small time. Since a high spatial precision gives better source localisation precision and smaller response times at the same time, the localisation task can be assumed to be accomplished with more ease than with a low spatial precision.

Lastly, it must be noted once again that head movement was not restricted during the experiment, allowing the participants to move their head freely when listening to the stimuli. Dynamic binaural cues were therefore exploitable by the participants, but this was the case for both spatial precision settings and there should therefore be no bias coming from that fact. It may influence the global mean response time across all settings but not the measured difference in mean response times when alternating between the two settings of the precision parameter.

Is it still Wave Field Synthesis?

The proposed technique relies on two fundamental properties of Wave Field Synthesis although it is using a significantly smaller number of loudspeakers than in conventional WFS for the same installation size.
First, it is derived from the Kirchhoff-Hemholtz integral, using a description of the target sound field at the boundaries of a reproduction subspace. As illustrated in section 4.4, the proposed technique follows similar approximations:

1. selection of a reduced portion of the surface using a 3D source visibility criterion,

2. selection of omnidirectional sources only,
3. discretisation of the line/surface.

The proposed method offers a more general discretisation of the surface allowing for irregular loudspeaker distributions. It also proposes an additional weighting of the loudspeakers so as to improve the rendering in a target listening area using an extension of the technique proposed by Corteel *et al.* [29] for 2 $\frac{1}{2}$ D WFS.
Second, it could be shown that the proposed method preserves localisation accuracy within an extended listening area. The proposed method does not realize a perfectly valid physical reproduction. However, the restriction to a horizontal linear array in 2 $\frac{1}{2}$ D WFS does not preserve the attenuation of the natural sound field and the sound field is not accurately reproduced above the aliasing frequency either. A large portion of the audible bandwidth therefore remains, where the sound field is only reproduced in a plausible way with limited localisation artifacts even in conventional WFS. The proposed method goes only one step further but proves to provide reliable localisation cues in height at two distinct listening positions separated by one meter.

4.6 Conclusion

The ASS evaluation techniques and the auditory models presented in the previous chapters of this thesis were designed for monaural processing. While they provide good results for situations where spatial audio is not considered, they may fail to do so in other conditions. Binaural models might therefore be considered, integrating psychoacoustic knowledge about spatial processing by human audition. This chapter provides an introduction to spatial hearing and associated phenomena. Binaural cues such as ITD and ILD were introduced in order to give a better understanding of the assumptions that are made in the design of the binaural model presented in chapter 5. Binaural unmasking was also introduced. It is a psychoacoustic phenomenon distinctive to binaural listening conditions and will therefore have to be considered for a binaural model as well.

3D audio synthesis techniques were also presented with an introduction to binaural synthesis and a more detailed presentation of 3D WFS. Both of them will be used in the subjective experiments presented in chapters 6 and 7 respectively. A practical implementation of 3D WFS that was developed by project partners during the i3Dmusic project (the framework of this thesis) is presented. The implemented formulation for 3D WFS enables precise spatial rendering of sound sources while addressing known problems of this reproduction technique. The validity of the proposed formulation is then confirmed by a sound source location matching experiment in the median plane that was conducted as part of this thesis. Participants are asked to match the perceived vertical location of a reference loudspeaker with a WFS virtual source pointer. Localisation performance is investigated using localisation accuracy, localisation precision and response time as performance cues.
Localisation accuracy is shown to be good, with 5 levels of elevation being discriminated for

2 listening positions and 2 spatial precision settings. A systematic bias of 8.2° in elevation is found, but this can be easily compensated if good absolute localisation is required. Even though not expected, the listening position is shown to influence the perceived location of a source. However, the change is only about 2° which can be neglected.
Localisation precision is shown to be about 6° - 9° even though only 24 loudspeakers are employed for the WFS system. As expected, the implemented spatial precision parameter increases the localisation precision by 2.3°. The benefits of this parameter are also shown in the response time analysis, where quicker response times were achieved with a higher spatial precision setting, implying a simpler localisation process.
Localisation performance is therefore judged to be good when compared to other studies with denser loudspeaker arrays or other spatial reproduction techniques. This supports the implemented 3D WFS technique as a serious alternative to other state-of-the-art spatialisation methods.

The next chapter bridges the gap between state-of-the-art ASS evaluation models and 3D audio by using knowledge about spatial unmasking presented in this chapter to introduce a new binaural ASS evaluation model based on Perceptual Model of Audio Quality (PEMO-Q) presented in chapter 2. A pre-study in binaural rendering conditions is also presented, using new ASS degradation synthesis algorithms.

5 ASS evaluation in 3D audio

5.1 Introduction

While the models presented in chapter 2 lead to the Perceptive Evaluation Methods for Audio Source Separation (PEASS) that performs very well in Audio Source Separation (ASS) evaluation as presented in chapter 3, it doesn't take into account binaural phenomena such as spatial unmasking and may therefore lose some of its accuracy in the 3D audio context (see chapter 4). Moreover, it is rather complex from a computational point of view.

This chapter presents the tools that were developed as preliminary steps towards the evaluation of ASS in the 3D audio context. First, algorithms simulating ASS degradations are presented as a way of providing reproducible and controlled experimental conditions, *i.e.* stimuli that contain a controllable amount of degradations without involving real-life ASS algorithms. A psychoacoustic pre-study in binaural conditions that was conducted to get user feedback and to assess the validity of the proposed algorithms is then presented. In a last part, an objective ASS evaluation model is presented. Not only is it simpler computationally, it also integrates knowledge about binaural hearing and will therefore be a valuable asset for ASS evaluation in the 3D audio context.

5.2 Synthetic ASS degradation: description of the methodology

If a model for ASS evaluation in the context of 3D audio is to be built, subjective studies including some test signals are unavoidable. As explained in section 1.4, the main families of degradations caused by ASS algorithms were identified as being

- static interference
- onset misallocation
- target distortion

- musical noise-type artifacts

If the relative importance of these degradation types is to be assessed, the stimuli that are used in the experiments need to contain controlled and reliable amounts of the different degradation families, which raises the need for synthesis algorithms to generate them. Real-life ASS algorithms may or may not introduce some amount of some or of all of those degradations, depending on the chosen sound excerpts. Thiede *et al.* summarize: "[...] that the part of the test signal most susceptible to artifacts may be only a short part of the total duration" [128]. This makes the selection of algorithms and excerpts cumbersome for subjective studies. Artificial ASS degradation synthesis algorithms have therefore been developed. Since many ASS algorithms are nondestructive (*i.e.* the sum of all estimated sources in a channel equals the original channel), the requirements for the synthetic degradation algorithms were to be nondestructive as well. The proposed algorithms have been formulated for stereophonic source material, which is the most common use case.

5.2.1 Static interference

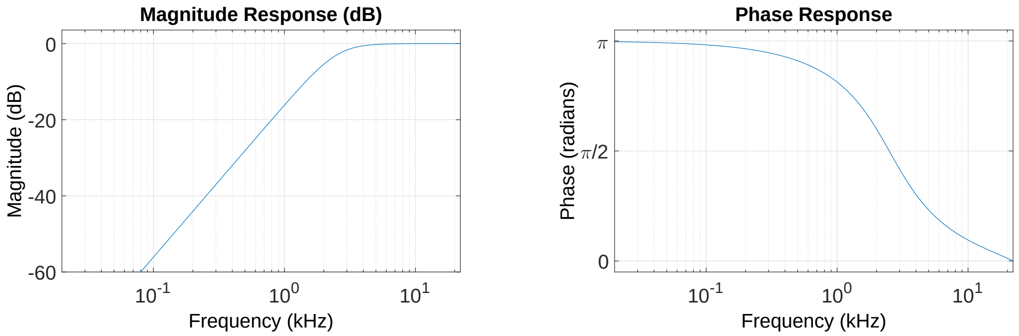
The first family of degradations which was identified is characterised by sources or parts of sources being falsely allocated to the estimated signals. This type of degradations has been proposed before by Vincent *et. al.* [134]. The estimated target therefore contains parts of the sources in the residue and / or the estimated residue contains parts of the target. However, the amount of each of the exchanged sources in the other estimated signal does not vary over time, as opposed to the onset misallocation degradation presented below. To simulate such a behavior, the following algorithm has been implemented: Based on an amount argument $a_{\text{interference}}$, the simulated estimated target and residue signals $\tilde{x}$ and $\tilde{y}$ are calculated from the reference target and residue signals x and y as

$$\begin{cases} \tilde{x} = x + a_{\text{interference}}\, y \\ \tilde{y} = y(1 - a_{\text{interference}}) \end{cases}, \quad \text{with } 0 \leq a_{\text{interference}} \leq 1. \tag{5.1}$$

5.2.2 Onset misallocation

The second family of degradations is characterised by onsets of sources in the residue being allocated to the estimated target and/or onsets of the target source being allocated to the estimated residue. Contrary to the interference degradation family, the amount of each source being switched between residue and target depends on the waveforms of the signals and therefore strongly varies over time. Note that if not stated otherwise, all processing steps below are applied to both channels if stereo source material is used.

To model this type of degradations, the reference signals x and y are first high-pass filtered at 2.5 kHz with a 4th-order Butterworth high-pass filter designed using the MATLAB® `butter` function (see the Bode diagram on figure 5.1). However, to prevent degradations due to the



(a) Magnitude response

(b) Phase response

Figure 5.1 – Response of the high-pass Butterworth filter used in the onset misallocation synthesis algorithm

filter phase, zero-phase filtering is used by applying the filter in forward and backward direction (implemented by using the MATLAB® `filtfilt` function), which squares the magnitude of the filter's transfer function and doubles its order. In the frequency domain, this can be expressed as:

$$\begin{cases} X_{\mathrm{hp}}(j\omega) = |H_{\mathrm{hp}\,2.5\mathrm{kHz}}(j\omega)|^2 X(j\omega) \\ Y_{\mathrm{hp}}(j\omega) = |H_{\mathrm{hp}\,2.5\mathrm{kHz}}(j\omega)|^2 Y(j\omega) \end{cases}, \quad \text{with } 1 \leq n \leq N \tag{5.2}$$

where $H_{\mathrm{hp}\,2.5\mathrm{kHz}}$ is the high-pass filter transfer function. The high-pass filtered signals $x_{\mathrm{hp}}[n]$ and $y_{\mathrm{hp}}[n]$ can then be computed through the inverse Fourier transform. The rationale for this filtering is that in practice, the onsets that are mostly being falsely allocated are percussive noises such as hi-hats or snare drums which dominate in the high-frequency range. The filtering allows a better focus on this type of noises, while preserving their temporal location. Note that zero-phase filtering is not causal and can't therefore be implemented in any type of online scenario.

The Hilbert envelopes x_{env} and y_{env} of the resulting signals are then computed by low-pass filtering the amplitude of their Hilbert transform at 20 Hz. The low-pass filter is once again a 4th-order Butterworth filter and zero-phase filtering is applied, squaring its amplitude and doubling the order.

$$\begin{cases} x_{\mathrm{env}}[n] = \mathscr{F}^{-1}(|H_{\mathrm{lp}\,20\mathrm{Hz}}(j\omega)|^2 \mathscr{F}(|\mathscr{H}\{x_{hp}[n]\}|)) \\ y_{\mathrm{env}}[n] = \mathscr{F}^{-1}(|H_{\mathrm{lp}\,20\mathrm{Hz}}(j\omega)|^2 \mathscr{F}(|\mathscr{H}\{y_{hp}[n]\}|)) \end{cases}, \quad \text{with } 1 \leq n \leq N \tag{5.3}$$

where $\mathscr{H}$ denotes the Hilbert transformation, $\mathscr{F}$ is the Fourier transform and $H_{\mathrm{lp}\,20\mathrm{Hz}}$ is the low-pass filter in the frequency domain.

The resulting signals contain information about the location of the onsets. To extract this information, the signals are then differentiated with respect to time and half-wave rectified to reject the negative part of the derivative, ignoring the decays.

$$\begin{cases} x_{\text{rect}}[n] = \max(x_{\text{env}}[n+1] - x_{\text{env}}[n], 0) \\ y_{\text{rect}}[n] = \max(y_{\text{env}}[n+1] - y_{\text{env}}[n], 0) \end{cases}, \quad \text{with } 1 \leq n \leq N-1 \tag{5.4}$$

Note that this reduces the signal length by 1 sample, which is not convenient for later processing steps. A sample with 0 value is therefore appended to the rectified signals.

$$\begin{cases} x_{\text{rect}}[N] = 0 \\ y_{\text{rect}}[N] = 0 \end{cases} \tag{5.5}$$

The signals are then normalised:

$$\begin{cases} x_{\text{norm}}[n] = \frac{x_{\text{rect}}[n]}{\max\limits_k(\{x_{\text{rect,L}}[k], x_{\text{rect,R}}[k]\})} \\ y_{\text{norm}}[n] = \frac{y_{\text{rect}}[n]}{\max\limits_k(\{y_{\text{rect,L}}[k], y_{\text{rect,R}}[k]\})} \end{cases}, \quad \text{with } 1 \leq n \leq N, \tag{5.6}$$

where the L and R subscripts designate the left and right channels of the signals respectively. A threshold is then applied to reject information from very small envelope fluctuations.

$$\begin{matrix} x_{\text{thr}}[n] = \begin{cases} x_{\text{norm}}[n] & \text{if} \quad x_{\text{norm}}[n] \geq T \\ 0 & \text{otherwise} \end{cases} \\ y_{\text{thr}}[n] = \begin{cases} y_{\text{norm}}[n] & \text{if} \quad y_{\text{norm}}[n] \geq T \\ 0 & \text{otherwise} \end{cases} \end{matrix}, \quad \text{with } 1 \leq n \leq N \tag{5.7}$$

where T is a constant threshold that was arbitrarily fixed to -40 dB with respect to the maxima of the signals.

After normalisation and applying a threshold to reject insignificant information, the locations of local maxima are extracted, resulting in peak vectors p_x and p_y. The only criteria being applied to the peak search is that two consecutive onsets must be separated by at least 90 ms, since this was arbitrarily chosen as being the length of an onset. With that constraint, about 11 onsets can be detected per second.

$$\begin{matrix} p_x[n] = \begin{cases} 1 & \text{if} \quad n = \arg\max\limits_l(x_{\text{thr}}[l]) \quad \forall l : n - L_{\text{onset}} \leq l \leq n + L_{\text{onset}} \\ 0 & \text{otherwise} \end{cases} \\ p_y[n] = \begin{cases} 1 & \text{if} \quad n = \arg\max\limits_l(y_{\text{thr}}[l]) \quad \forall l : n - L_{\text{onset}} \leq l \leq n + L_{\text{onset}} \\ 0 & \text{otherwise} \end{cases} \end{matrix} \quad \text{with } 1 \leq n \leq N \tag{5.8}$$

where L_{onset} is the number of samples corresponding to the onset length of 90 ms.

If stereo source material is used, the peak location vectors of the left and right channels of the residue signal $p_{y,\text{L}}$ and $p_{y,\text{R}}$ are then compared and peaks that fall within an arbitrarily fixed offset tolerance window of ± 75 ms are considered as being synchronous (centered in the stereo panorama). Their location is kept for further processing while other peaks are rejected.

This can be formulated as extracting the residue peak vector p_y according to

$$p_y[n] = \begin{cases} 1 & \text{if} \quad \sum_{l=n-L_\text{offset}}^{n+L_\text{offset}} p_{y,\text{L}}[l] + p_{y,\text{R}}[l] > 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{with } 1 \leq n \leq N \tag{5.9}$$

where L_offset denotes the number of samples corresponding to the offset tolerance of 75 ms. The synchronicity between the target and the residue is then taken into account by weighting the target's peaks according to their distance to the next peak in the residue. To do so, the target peak signal is convolved with a gaussian window $g[n]$ of 150 ms length and multiplied element by element with the residue onset peaks, resulting in a "detection probability" for every peak in the residue:

$$p_\text{sync}[n] = (p_x * g)[n] p_y[n], \quad \text{with } 1 \leq n \leq N \tag{5.10}$$

where $(*)$ denotes convolution. Note that $0 \leq p_\text{sync}[n] \leq 1$. If stereo material is used, this is done for both channels of the target. Target onsets are considered to be always centered in the stereo panorama for this work, because the target was always the voice of the singer. This may need to be revised for future applications. If stereo material is used, the overall peak vector containing weighted possible onset location candidates p_weighted can then be extracted:

$$p_\text{weighted}[n] = \max(p_{\text{sync,L}}[n], p_{\text{sync,R}}[n]), \quad \text{with } 1 \leq n \leq N \tag{5.11}$$

The P remaining candidates are then sorted by value and the greatest $P \times a_\text{onset}$ are kept and their weight set to 1, resulting in selected peak vector $p_\text{selected}[n]$. Other peaks are discarded by setting their weight to 0. a_onset is the amount argument associated with this family of degradations and its value is between 0 and 1.

The selected peaks vector p_selected is then used to construct an onset signal o_x from the target signal x by convoluting half of a Tukey window of length $2L_\text{onset}$ (half the window therefore having length L_onset) with the peaks vector. The Tukey window is a tapered cosine window. It is a rectangular window at its center and increasing and decreasing cosine portions at its beginning and end respectively. It can be denoted

$$w[n] = \begin{cases} \frac{1}{2}\left\{1 + \cos\left(\pi\left(\frac{2n}{r(2L_\text{onset}-1)} - 1\right)\right)\right\} & \text{for } 0 \leq n \leq \frac{r(2L_\text{onset}-1)}{2} \\ 1, & \text{for } \frac{r(2L_\text{onset}-1)}{2} \leq n \leq (2L_\text{onset}-1)\left(1 - \frac{r}{2}\right) \\ \frac{1}{2}\left\{1 + \cos\left(\pi\left(\frac{2n}{r(2L_\text{onset}-1)} - \frac{2}{r} + 1\right)\right)\right\} & \text{for } (2L_\text{onset}-1)\left(1 - \frac{r}{2}\right) \leq n \leq 2L_\text{onset} - 1 \end{cases}, \tag{5.12}$$

where r is a parameter denoting the ratio of cosine-tapered section length to the entire window length. r is defined to be equal to 0.1 here. Only the second half of $w[n]$ is used here, which is denoted $t[n] = w[l]$, where $l = L_\text{onset} + n$ and $0 \leq n \leq L_\text{onset} - 1$. This allows for a smooth fadeout of the onsets. The resulting convolved peaks vector is then multiplied element by element with the target:

$$o_x[n] = (t[n] * p_{\text{selected}}[n])x[n], \quad \text{with } 1 \leq n \leq N \tag{5.13}$$

The simulated target and residue signals $\tilde{x}[n]$ and $\tilde{y}[n]$ can then be computed by simply subtracting the onset vector from the target and adding it to the residue:

$$\begin{cases} \tilde{x}[n] = x[n] - o_x[n] \\ \tilde{y}[n] = y[n] + o_x[n] \end{cases}, \quad \text{with } 1 \leq n \leq N \tag{5.14}$$

This onset misallocation synthesis algorithm may seem rather arbitrary. In fact, onset detection is an active research area and multiple methods based on different characteristics of the signal are available. For a tutorial, see Bello *et al.* [8] for example. While the general scheme of onset detection algorithms presented by Bello *et al.* [8] is followed here, in-depth investigation of onset detection is out of the scope of this work. The presented algorithm was adjusted "by ear" to imitate onset misallocation degradations produced by ASS algorithms. It is by no means a way of detecting every onset in a piece of music and further improvement may be obtained by integrating onset detection research into it.

5.2.3 Target distortion

The target distortion degradation is characterised by changes in the timbre of certain sources. This is simulated by low- and high-pass filtering both the target and the residue signals and exchanging the high-pass filtered parts in a similar fashion to what was proposed by Emiya *et al.* as low target distortion anchor [42]. For the presented studies, this has been implemented with Butterworth high- and low-pass filters of order 16. The cut-off frequency $a_{\text{target dist}}$ is the amount argument of this degradation. Zero-phase filtering is employed, doubling the order of the initial filters and squaring the magnitude of the transfer function. First, both target and residue signals x and y are filtered to obtain x_{hp}, x_{lp}, y_{hp} and y_{lp}. The simulated estimated signals $\tilde{x}$ and $\tilde{y}$ are then computed as

$$\begin{cases} \tilde{x} = x_{lp} + y_{hp} \\ \tilde{y} = y_{lp} + x_{hp}. \end{cases} \tag{5.15}$$

In the presented studies, $a_{\text{target dist}}$ varied between 500 and 22500 Hz.

5.2.4 Musical noise-type artifacts

The musical noise-type artifacts degradation is characterised by the exchange between the target and the residue of small portions of signal which are concentrated in time and frequency. It is synthesised by exchanging single time/frequency bins between the spectrograms of the signals. A random selection as done by Emiya *et al.* [42] is however not applicable, since it does not sound as desired. For this reason, a slightly more elaborate selection method was chosen. As in the onset misallocation synthesis algorithm, if not stated otherwise, the same processing is applied to the left and right channels when using stereo material.

First, filtered versions x_f and y_f of the target x and the residue y respectively are computed. The filter is an ITU-R BS.468-4 [62] type weighting filter with impulse response h_{468} designed using a Least P-norm Infinite Impulse Response (IIR) implementation in MATLAB®. This weighting was specifically developed to measure the audibility of noise in audio and seems therefore better suited for noise artifacts than the more popular A-weighting that was fitted to the loudness of single tones. In the filtered signals, clearly audible parts of the signal should therefore be accentuated, while others are attenuated.

$$\begin{cases} x_f[n] = h_{468}[n] * x[n] \\ y_f[n] = h_{468}[n] * y[n] \end{cases}, \quad \text{with } 1 \leq n \leq N \tag{5.16}$$

In a second step, the spectrograms of the filtered target x_f, of the filtered residue y_f and of their sum signal $s = x_f + y_f$ are computed for every channel. To compute the spectrograms, the short-term discrete Fourier transform is used with a 46 ms half-overlapping sine-window w as proposed in PEASS [42]. Since these spectrograms are only used for the selection of audible time/frequency bins, phase is discarded and only the magnitude is kept.

$$\begin{cases} X_f(r,k) = \left|\sum_{m=0}^{N-1} x_f[rR+m]\,w[m]e^{-j\frac{2\pi}{L}km}\right| \\ Y_f(r,k) = \left|\sum_{m=0}^{N-1} y_f[rR+m]\,w[m]e^{-j\frac{2\pi}{L}km}\right| \\ S_f(r,k) = \left|\sum_{m=0}^{N-1} (x_f[rR+m] + y_f[rR+m])\,w[m]e^{-j\frac{2\pi}{L}km}\right| \end{cases} \tag{5.17}$$

where L is the length of the window in samples, $R = L/2$ is the hop size of the analysis frame, $0 \leq k \leq L-1$ indexes the quantised frequency vector, $0 \leq r \leq (\lfloor (N-L)/R \rfloor - 1)$ is the analysis frame index and N is the length of the signal.

Since the goal is to exchange coefficients between the residue and the target, centered coefficients need to be selected when using stereo source material, because in the presented work, the voice of the lead singer was always selected as target. In almost every modern mix, the lead voice is positioned at the center of the stereo panorama, hence the search for centered coefficients. This criterion is implemented by rejecting coefficients that are below a certain threshold in at least one channel. In that way, coefficients that have at least a certain amount of energy in both channels remain. The threshold T_{center} was arbitrarily set to -40 dB in this case.

$$\begin{aligned} X_{\text{centered,L/R}}(r,k) &= \begin{cases} X_{f,\text{L}}(r,k) & \text{if } X_{f,\text{L}}(r,k) \geq T_{\text{centered}} \text{ and } X_{f,\text{R}}(r,k) \geq T_{\text{centered}} \\ 0 & \text{otherwise} \end{cases} \\ Y_{\text{centered,L/R}}(r,k) &= \begin{cases} Y_{f,\text{R}}(r,k) & \text{if } Y_{f,\text{L}}(r,k) \geq T_{\text{centered}} \text{ and } Y_{f,\text{R}}(r,k) \geq T_{\text{centered}} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5.18}$$

The location of every nonzero element is then stored for further processing as a potential switching candidate. The vector with candidate locations from the target is concatenated with that of the residue and duplicates are removed, leaving a location vector $p_{\text{candidates}}$ of length P with the associated amplitude vectors $q_{x,\text{candidates}}$ and $q_{y,\text{candidates}}$ containing the candidates' amplitude of the target and the residue respectively.

For every switching candidate $p_{\text{candidates}}(l)$, third-octave neighborhood regions $M_{x,l}(r',k')$ and $M_{y,l}(r',k')$ are constructed in the target and residue respectively, reaching one sixth of an octave below and one sixth of an octave above the coefficient and including the precedent 8 frames of the spectrogram (see section 2.5.1 for the definition of third-octave bands). The lth coefficient from $p_{\text{candidates}}$ is therefore located on the last column and middle row of $M_{x,l}$ or $M_{y,l}$, depending on wether it was selected from the target or the residue respectively.

The averages $m_{x,l}(r')$ and $m_{y,l}(r')$ over the third-octave of both neighborhoods are then computed in every time frame, giving a cue about the temporal evolution of the amplitude around the selected coefficient.

$$\begin{cases} m_{x,l}(r') = \frac{\sum_{k'=1}^{K'} M_{x,l}(r',k')}{K'} \\ m_{y,l}(r') = \frac{\sum_{k'=1}^{K'} M_{y,l}(r',k')}{K'} \end{cases}, \tag{5.19}$$

where K' is the length of the frequency vector defining the third-octave neighborhood, *i.e.* the number of rows in $M_{x,l}(r',k')$.

The mean of $m_{x,l}(r')$ is then computed for the 8 frames preceding the location of the selected coefficient:

$$\begin{cases} m_{x,l} = \frac{\sum_{r'=1}^{R'-1} m_{x,l}(r')}{R'-1} \\ m_{y,l} = \frac{\sum_{r'=1}^{R'-1} m_{y,l}(r')}{R'-1} \end{cases}, \tag{5.20}$$

where $R' = 9$.

In the last column, containing the selected coefficient, the mean of all coefficients, excepting the selected coefficient is computed to give a cue about its instantaneous audibility $b_{x,l}$ and $b_{y,l}$:

$$\begin{cases} b_{x,l} = \frac{\sum_{k' \neq (K'+1)/2} M_{x,l}(R',k')}{(K'+1)/2} \\ b_{y,l} = \frac{\sum_{k' \neq (K'+1)/2} M_{y,l}(R',k')}{(K'+1)/2} \end{cases}. \tag{5.21}$$

$m_{x,l}, m_{y,l}, b_{x,l}$ and $b_{y,l}$ are then used as audibility criteria for the selected coefficients. A coefficient is kept if one of the two following sets of conditions is fulfilled:

1. $q_{x,\text{candidates}}(l) > 2b_{y,l}$ and $q_{x,\text{candidates}}(l) > m_{x,l}$

2. $q_{y,\text{candidates}}(l) > 2b_{x,l}$ and $q_{y,\text{candidates}}(l) > m_{y,l}$

This ensures that a given coefficient is audible in its frame when switched and also that it's absence is well noticeable in its original neighborhood. Coefficients that do not meet one of the criteria are rejected, resulting in the retained coefficients vector $p'_{\text{candidates}}$ of length P' and the associated amplitude vector $q'_{\text{candidates}}$.

Once this candidate selection process is done, the coefficients to be switched are selected. The total amount of coefficients per frame to be switched is determined by the amount argument $a_{\text{musical noise}}$: a total of $K \times a_{\text{musical noise}}$ coefficients are switched between target and residue in every frame. The coefficients to be switched are selected by a succession of rules:

1. If $K \times a_{\text{musical noise}} \leq P'$, select the $K \times a_{\text{musical noise}} T_{\text{energy}}$ coefficients from $p'_{\text{candidates}}$ with the highest amplitude and $K \times a_{\text{musical noise}}(1 - T_{\text{energy}})$ coefficients randomly from the remaining candidates in $p'_{\text{candidates}}$. T_{energy} designates the ratio of high-energy coefficients that are selected and was arbitrarily fixed to $T_{\text{energy}} = 0.1$.

2. If $P' < K \times a_{\text{musical noise}} \leq P$, select all coefficients in $p'_{\text{candidates}}$ and randomly select $K \times a_{\text{musical noise}} - P'$ coefficients from $p_{\text{candidates}}$ that are not selected yet.

3. If $P' < P < K \times a_{\text{musical noise}}$, select all coefficients in $p_{\text{candidates}}$ and randomly select $K \times a_{\text{musical noise}} - P$ coefficients that are not selected yet.

This drawing procedure ensures a fixed number of switched coefficients, while using the most audible ones first.
The spectrograms of the unfiltered signals are finally computed (keeping the phase)

$$\left\{\begin{array}{l} X(r,k) = \sum_{m=0}^{N-1} x[rR+m]\,w[m]\,e^{-j\frac{2\pi}{L}km} \\ Y(r,k) = \sum_{m=0}^{N-1} y[rR+m]\,w[m]\,e^{-j\frac{2\pi}{L}km} \\ S(r,k) = \sum_{m=0}^{N-1} (x[rR+m] + y[rR+m])\,w[m]\,e^{-j\frac{2\pi}{L}km} \end{array}\right. , \tag{5.22}$$

and for every selected coefficient location, the value of $S(r, k_{\text{selected}})$ is randomly assigned to either $X(r, k_{\text{selected}})$ or $Y(r, k_{\text{selected}})$, while the other one is set to 0. The simulated estimated signals $\tilde{x}$ and $\tilde{y}$ can then be computed by applying the inverse short-time Fourier transform to the resulting spectrograms.
Once again, this procedure was developed by trial and error, trying to perceptually match musical noise-type artifacts as they are produced by real-life ASS algorithms. In the presented studies, $a_{\text{musical noise}}$ varied between 0 and 1 percent.

5.3 Binaural pre-study

A first experiment was conducted as pre-study to evaluate the degradation synthesis algorithms. The research question being answered was: "Is there a linear relationship between the synthesis parameters and the perceived quality of degraded sound excerpts?" This was examined in various experimental conditions.

5.3.1 Experimental Design

This pre-study was split into 3 sessions, differing by the degradation family under test. Musical noise-type artifacts were rated in the first session, target distortion was evaluated during the second session and the third session contained the onset misallocation degradation family. The three sessions took place on three different days for each participant to prevent excessive fatigue.

13 participants (1 woman and 12 men) between ages 22 and 39 (M = 28.4 SD = 4.1), took part in the study. All participants did have experience with audio and did not report any listening impairment.

The experimental design in this case was a repeated measures design with 3 factors: the song (6 levels), the degradation level (6 levels) and the target-to-residue ratio (3 levels). The degradation family is not treated as a factor because of the design of the experimental task, which prevents comparisons between sessions.

5.3.2 Experimental Task

The subjective evaluation task was defined according to a modified Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) protocol [64] as used in state-of-the-art studies for ASS evaluation [42]. The participant had to rate the overall quality of a given number of stimuli as compared to a given reference. A hidden reference was included in the stimuli. A low anchor was however not included in the stimuli. The goal of the study was to evaluate if the proposed synthesis methods enabled the creation of controlled signals which might serve as anchors in oncoming studies. There is therefore no need of low anchors to get a low reference.

As required by the MUSHRA protocol, each experimental session was composed of a training phase and of an evaluation phase. During the training phase, the participants could get accustomed to the rating interface and listen to all the stimuli of the session to appreciate the range of degradation which is encountered during the evaluation phase. This allows participants to construct their perceptive scale by remembering the worst overall stimulus for the bottom end of the scale.
During the evaluation phase, on each test page of the interface, the participant was presented with 6 test sounds (including the reference) in random order. The original proposition of the MUSHRA protocol would recommend that participants rate every sound on a quality scale that is continuous, but with 5 regions labelled “Excellent”, “Good”, “Fair”, “Poor” and “Bad” [64]. In order to avoid possible bias due to perceptually nonlinear scales [63, 145], the labelling was omitted for this experiment. The scales were defined to range from 0 to 100 and no graduation other than the current rating in numerical form was given, except for the terms “best” and “worst” labeling the top and the bottom of the scale respectively to give a sense of the direction of the scale.

5.3.3 Stimuli

A set of 6 songs which are freely available in stem format (*i.e.* in the form of clean separated source tracks) were selected as basis for the stimuli. An excerpt was chosen out of each song (see table 5.1 for details). The lead voice and the residue were then mixed down as two stereo files for further processing. This corresponds to perfect separation of the lead voice in the ASS context.

For each family of signal degradations described in the section above, except for the interference degradation family, 6 stimuli were constructed per musical excerpt. They were based on 6 levels of degradations, corresponding to variations in their respective amount arguments $a'_{\text{musical noise}}$, $a'_{\text{target dist}}$ and a'_{onset}. The formulae for each amount level of each degradation family are given below. Note that amount arguments used in this pre-study are based on preliminary degradation synthesis algorithms and are therefore noted with an apostrophe $(\cdot)'$, because the algorithms may not correspond exactly to those presented in the first part of this chapter. The interference degradation family was excluded because its perceptual impact is estimated to be minimal in 3D audio contexts as compared to the other degradation families that directly concern the audio quality of the estimated target or contain dynamic elements. Furthermore, the interference degradation family is already well-known from PEASS [42] and no modifications was proposed in this thesis.

- Musical noise-type artifacts:

$$\begin{cases} a'_{\text{musical noise}} = 0 \\ a'_{\text{musical noise}} = \frac{2}{2^{(6-k)}} \qquad \text{with } 2 \le k \le 6 \end{cases} \tag{5.23}$$

- Target distortion:

$$a'_{\text{target dist}} = 500 \cdot 2^{\frac{6-k}{5} \log_2 \frac{22050}{500}} \qquad \text{with } 1 \le k \le 6 \tag{5.24}$$

- Onset misallocation:

$$a'_{\text{onset}} = \frac{1}{5}(k-1) \qquad \text{with } 1 \le k \le 6 \tag{5.25}$$

For each combination of song and degradation level, 3 levels of Target to Residue Ratio (TRR) were synthesised by multiplying the residue by a factor corresponding to -6, -12 or -18 dB, resulting in a corresponding positive TRR. Since binaural spatialisation was used with stimuli presented over headphones, they were then convolved with the Head Related Transfer Function (HRTF) corresponding to 30° and $-30°$ azimuth and 0° elevation and with the Headphone Transfer Function (HpTF) correction filter. The TRR is needed to make the degradations audible when both signals are at the same location. The nondestructive nature of the synthesis algorithms, combined with the matching locations of the target and the

Table 5.1 – Musical excerpts used for the stimuli in the pre-study

#	Artist	Title	Start	End	Duration
1	Beyoncé	End of Time	1:58.0	2:14.3	0:16.3
2	Kanye West	Love Lockdown	0:47.6	1:03:7	0:16.1
3	Phoenix	Lisztomania	1:42.1	1:54.5	0:12.4
4	Shannon Hurley	Sunrise	1:14.0	1:25.9	0:11.9
5	The Ultimate NZ Tour	Four Good Reasons	0:10.5	0:27.5	0:17.0
6	Shannon Hurley	We are in love	2:38.1	3:02.8	0:24.7

residue eliminate any degradation if a TRR of 0 dB is used. If distinct locations are used for the target and the residue, degradations will be audible even when the TRR is equal to 0 dB, due to different auditory phenomena, such as auditory localisation (parts of sources effectively switching place) and binaural unmasking.

5.3.4 Results

Post-Screening

The results of the participants were submitted to a post-screening procedure. For all scores given to hidden reference signals, the Euclidian distance to the sample mean of the corresponding session was computed. For each session, a critical distance equal to the average of the computed distances was then defined. Results of participants whose distances were above the critical distance on average were discarded. This criteria works well with few extreme outliers as it is the case in the presented study. For the first session, the results of two participants had to be discarded. The critical distance was $D_c = 31.9$ and discarded participants had distances of $D = 75.6$ and $D = 138.9$. The mean distance amongst remaining participants is $D = 18.2$. For the second session, the results of three participants had to be discarded. The critical distance was $D_c = 26.6$ and discarded participants had distances of $D = 46.7$, $D = 64.8$ and $D = 90.5$. The mean distance amongst remaining participants is $D = 14.3$. For the third session, the results of 3 participants had to be discarded. The critical distance was $D_c = 6.4$ and discarded participants had distances of $D = 7.9$, $D = 12.6$ and $D = 30.8$. The mean distance amongst remaining participants is $D = 3.2$.

Sample Means

For each degradation type / degradation level / TRR / song combination, the Mean Objective Score (MOS) has been calculated. Results are shown on figures 5.2, 5.3 and 5.4. The mean score over all songs is also represented (bold lines).

For the musical noise-type degradation, the mean scores decrease monotonically as a function of the ratio of swapped time/frequency bins. They also decrease monotonically with TRR. For a TRR of 6 dB, the mean score across songs decreases to 83.8 for the maximum degradation.

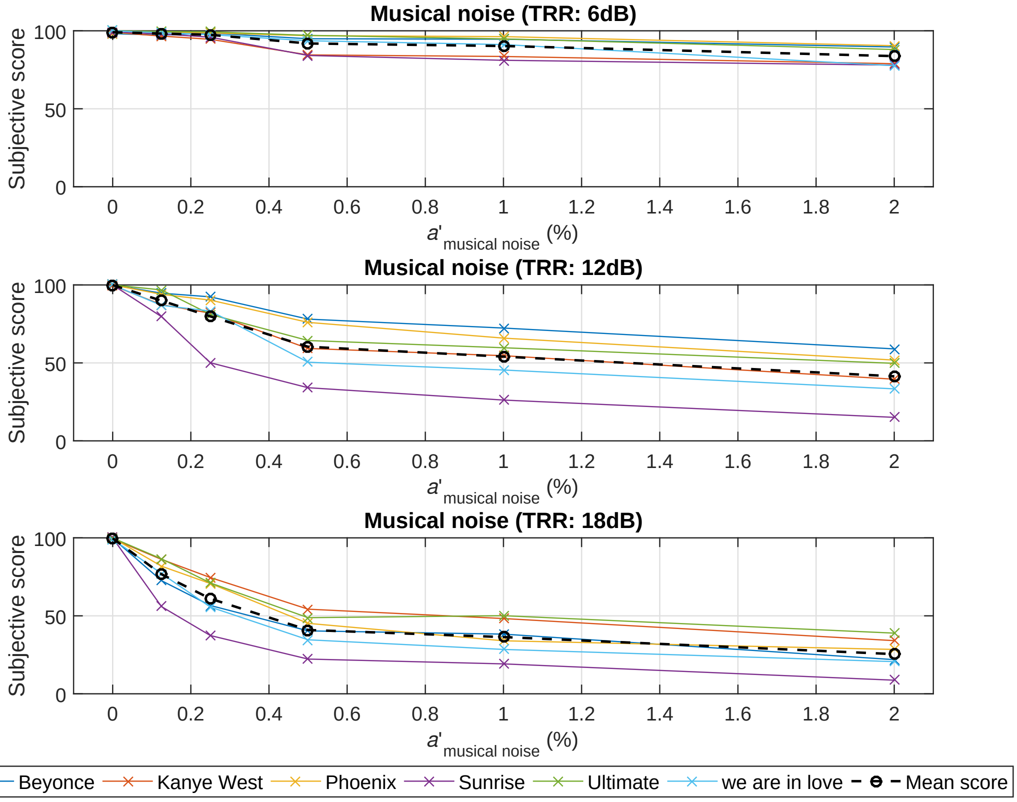


Figure 5.2 – Sample means for the musical noise degradation for all three TRR. Individual songs are plotted, as well as the mean score over all songs.

For a TRR of 12 dB, it decreases to 41.4 and for 18 dB, it reaches 25.5. Note that scores seem to decrease inversely proportional to the amount argument

In the case of onset misallocation degradation, the mean score also decreases almost monotonically with the amount argument, with an exception in the 6 dB TRR case, where the mean score slightly increases in the last step. This is essentially due to the Kanye West excerpt that got a much higher score for $a'_{\text{onset}} = 1$. The decrease with respect to TRR is also monotonic. For the 6 dB TRR, the subjective score decreases from 100 to 67.3; in the 12 dB case, it reaches 40.9 and in the TRR 18 dB case, the lowest score is 33.2. Once again, the subjective score seems to follow an inversely proportional law with respect to the amount argument.

The mean score also follows a monotonic decrease with respect to the cut-off frequency for the target distortion degradation. The same behavior is observed with respect to the TRR. For a 6 dB TRR, the mean score reaches 62.6, for 12 dB it is 31.4 and for 18 dB 18.5. Note that the direction of the horizontal axis on figure 5.4 is inverted, beginning at 22050 Hz and following a logarithmic scale.

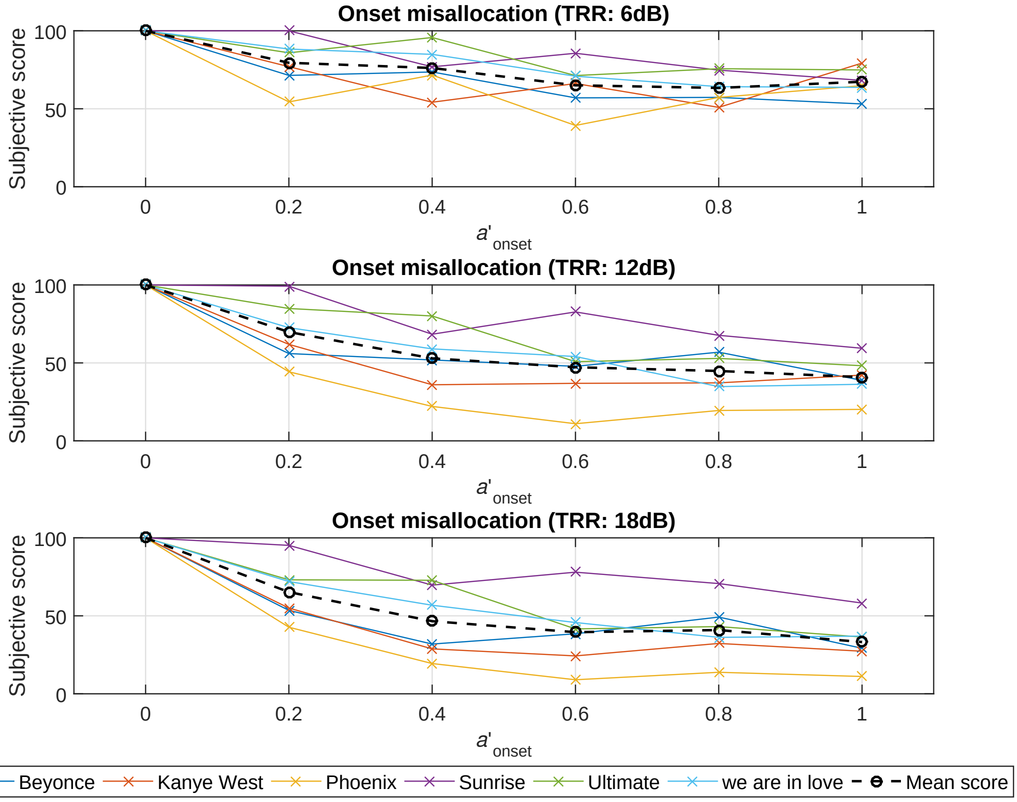


Figure 5.3 – Sample means for the onset misallocation degradation for all three TRR. Individual songs are plotted, as well as the mean score over all songs.

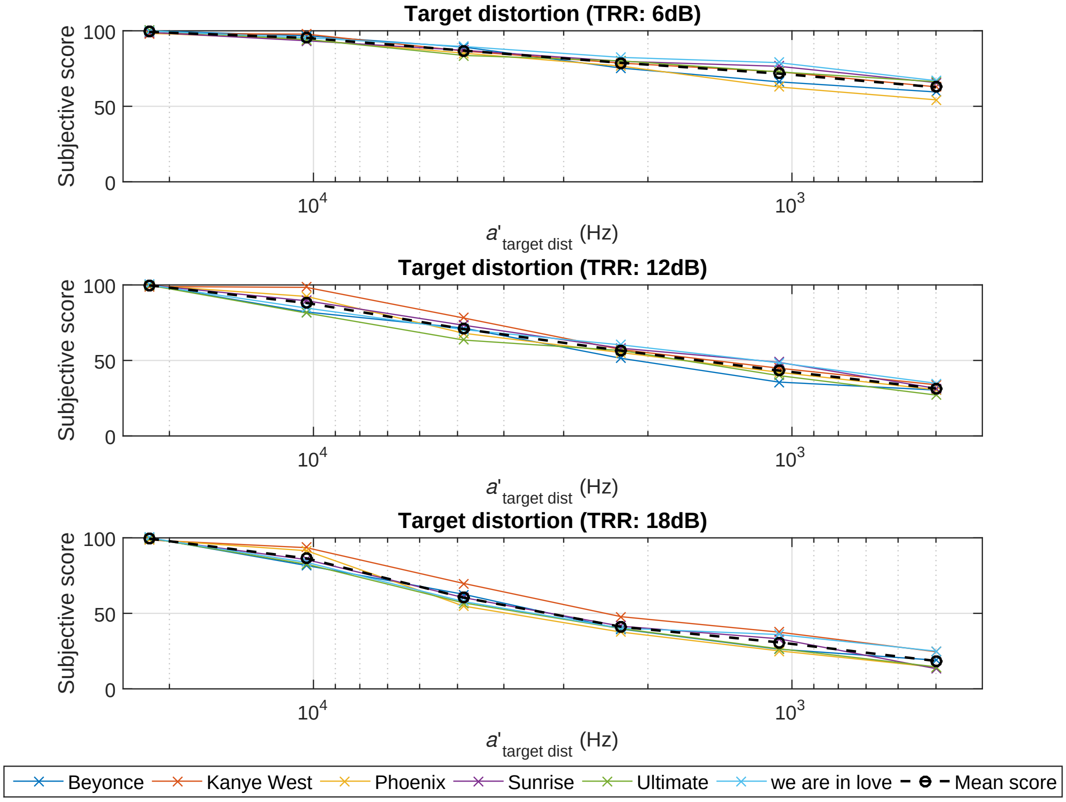


Figure 5.4 – Sample means for the target distortion degradation for all three TRR. Individual songs are plotted, as well as the mean score over all songs.

Feedback of the participants

Oral feedback of the participants was collected and analysed. The following points were common to several participants:

- The musical excerpts should be shorter. A length of 12-25 s per stimulus is too long considering the number of stimuli.
- The onset misallocation degradation could be essentially evaluated by counting the misallocations and strongly depends on the part of the musical excerpt which is considered (only occurring at the end of a stimuli for example).
- The experiment as a whole is rather long.

5.3.5 Discussion

As expected, scores for the hidden reference signals are all very close to 100, whereas they decrease as degradations increase. This holds true for all degradation types and TRRs. This validates the artificial degradation synthesis algorithms in the sense that the amount arguments seem to have a reliable effect on the perceived quality of the degraded sounds. The quality does not decrease linearly with respect to the amount arguments, which seems to infer that smaller degradations might be more detrimental regarding perceived quality than bigger degradations, *i.e.* adding more degradations does not “hurt” as much as adding a little bit of degradations.
Note that the magnitude of the decrease cannot be compared across degradation types, since the respective scales might be different. The experimental task specified that the worst sound over the whole session had to be rated lowest, so it is possible that scales changed between sessions.

The feedback of the participants provided valuable information about the degradation synthesis algorithms and about the experimental protocol for the studies presented in chapters 6 and 7. The first and last points mentioned above were taken into account for the binaural and Wave Field Synthesis (WFS) studies. The second point was taken into account in the final version of the degradation synthesis algorithms presented at the beginning of this chapter. In fact, the onset detection criteria was relaxed a bit, to allow enough onsets to be detected, so that participants are not able to count them anymore.

5.4 Proposed modelling approach

As mentioned in the introduction of this chapter, the (extended) PEASS model presented in chapter 3 is quite intensive in terms of computational power needed and does not integrate psychoacoustic knowledge about spatial hearing. A lighter model is therefore proposed, based on the Perceptual Model of Audio Quality (PEMO-Q) [58], that is also part of the PEASS model.

In fact, PEMO-Q is designed for quality evaluation tasks, since it compares a signal under test to a reference signal. This very much resembles the experimental task for the pre-study presented above or for the subjective study that was presented with the PEASS model [42]. The proposes modelling approach takes advantage of this feature and uses it in a binaural listening context.

5.4.1 Binaural models

The PEMO-Q model already integrates a lot of psychoacoustic knowledge about monaural processing. However, spatial listening involves additional phenomena, such as auditory localisation and binaural unmasking, as presented in section 4.2. A way of integrating such features into a PEMO-Q-based model is therefore searched.

Localisation modelling

Basic localisation models work mainly with the Interaural Level Difference (ILD) and Interaural Time Difference (ITD) cues presented in section 4.2.1. Most often, a monaural processing stage not unlike those of the models presented in chapter 2 precedes the cue extraction to take into account variations with frequency of the perceived cues. The ITD can then be extracted from the Interaural Cross-Correlation (IACC). The location of the maximum of the IACC in fact corresponds to the perceived ITD. The set of possible source locations is then determined based on the ITD (as explained by Blauert *e.g.* [13]). The ILD is given by the level difference between the left and right auditory channels. Blanco-Martín *et al.* exploit these properties to estimate the perceived source location based on a HRTF database [12]. Additionally, they use the high-frequency cues of monaural signals to localize sources in the median plane.

Karjalainen presented a binaural auditory model with a binaural processor that extracts source location estimates from the running IACC and the ILD of estimated neural firing rates of the right and left auditory pathways [72]. The neural firing rates are very similar to the internal representations from other auditory models such as PEMO-Q. While there is no reported data validating Karjalainen's model in the cited paper, he states that the estimated source location estimates may be used together with the computed neural firing rate patterns in a variety of application scenarios ranging from quality evaluation to source localisation.

The Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener (QESTRAL) model presented by Rumsey *et al.* [26, 40, 67, 109] aims at predicting the spatial quality of reproduced sound using an artificial listener. It is not dissimilar to the PEASS model in its structure (but not in its objective). The extracted binaural features are based mostly on the computation of the IACC, the ITD and of the Interaural Intensity Difference (IID).

Other models integrate more advanced mechanisms involving the modelling of higher stages of the neural processing chain. Goodman and Brette for example use neuronal synchronicity patterns to estimate source location [48]. For a review of binaural localisation models, see for example Merimaa [85].

Binaural unmasking modelling

While sound source localisation modelling or spatial reproduction quality is not the primary interest of this thesis, binaural unmasking involves the repositioning of a sound source and may therefore very well be sensitive to the ILD, ITD and IACC cues used for source localisation. In fact, binaural unmasking can also be modelled based on the interaural cues. Equalisation-cancellation theory predicts that the auditory system is able to cancel an interferer if it has a different ITD than the target [41]. This is implemented for example by Lavandier & Culling [81] to predict binaural speech intelligibility in rooms. They used the following formula to compute the Binaural Masking Level Difference (BMLD), based on psychoacoustic experiments conducted by Culling *et al.* [32, 33]:

$$\text{BMLD} = 10\log\left[\frac{k - cos(\phi_s - \phi_m)}{k - \rho}\right], \tag{5.26}$$

where $(\phi_s - \phi_m)$ is the phase difference between the signal and the masker at the maximum of the IACC, ρ is the interaural coherence (*i.e.* the maximum value of the IACC) and $k = (1+\sigma_\epsilon^2)e^{\omega^2\sigma_\delta^2}$, is a function of the angular frequency ω and the two constants σ_ϵ^2 and σ_δ^2 that were proposed by Durlach [41]. Note that this computation is done after filtering both signals with a gammatone filter bank with 2 filters per Equivalent Rectangular Bandwidth (ERB).

The Interaural Phase Difference (IPD)-based computation above does not take into account better-ear listening that is a consequence of ILD. Lavandier & Culling's model therefore also computes Signal to Noise Ratios (SNRs) in all frequency bands based on excitation patterns and keeps the maximum value of both ears to take it into account. The SNR is then integrated across frequency to compute the target-to-interferer ratio. The BMLD is also integrated across frequency to compute the global binaural advantage which is then added to the target to interferer ratio to get the "effective'" target-to-interferer ratio from which the estimated intelligibility is then derived. The schematic representation of this model is given on figure 5.5. Results show a high correlation between predicted and subjective values, even for reverberant conditions. While the model was revised in a later paper by Jelfs *et al.* [68], the proposed improvements were essentially of computational nature and did not affect the results. Instead of using signals that were convolved with Binaural Room Impulse Responses (BRIRs), the BRIRs were directly taken as input to the model and instead of computing excitation patterns, the outputs of the gammatone filter bank were directly analysed.

5.4.2 PEMO-Q-based binaural model

The different binaural modelling approaches summarised above show that IACC and ILD seem to play a major role in the perception of source location and in binaural unmasking. Different source locations can be expressed as variations in the IACC and ILD. In terms of perceived quality modelling such as it is implemented in the PEMO-Q model, this translates to the ILD and IACC being different in a sound scene under test as compared to a reference sound scene, leading to a perceived quality degradation. The same is true for binaural unmasking. It is

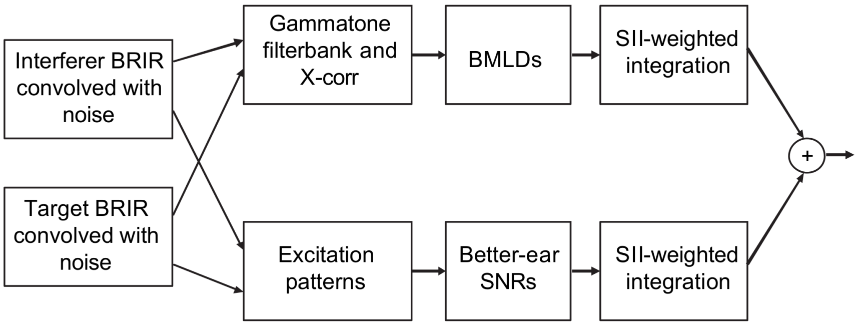

Interferer BRIR convolved with noise
Target BRIR convolved with noise
Gammatone filterbank and X-corr
BMLDs
SII-weighted integration
Excitation patterns
Better-ear SNRs
SII-weighted integration
+

Figure 5.5 – Block scheme of Lavandier & Culling's model [81] as cited by Jelfs *et al.* [68]. Source: Jelfs *et al.* [68].

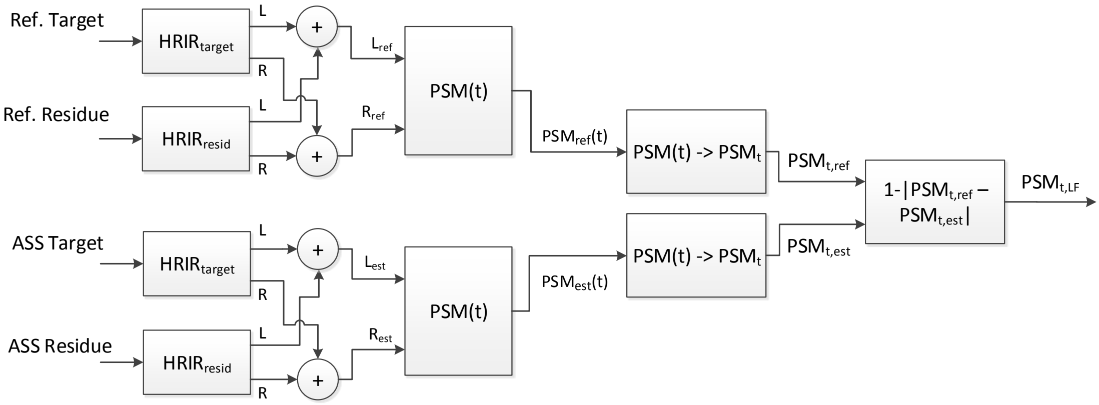


Figure 5.6 – Structure of the low-frequency part of the proposed binaural model. The HRIR blocks model the acoustic pathway between the source and the listener.

therefore suggested to use the PEMO-Q model in a binaural context, especially since it already relies on internal correlation between the signal under test and the reference signal.

The proposed model consists of two parts similar to Lavandier & Culling's model [81] presented above: A low-frequency part relying on some sort of IACC and a high-frequency part, relying on some sort of ILD and therefore better-ear hearing. The proposed structure is illustrated in figures 5.6 and 5.7. Binaural signals are assumed at the input of the model, hence the Head-Related Impulse Response (HRIR) blocks on the two figures. This can be achieved either by simulation of a given spatialisation system or by recording with a manikin or in-ear microphones for example. The resulting reference signals denoted L_{ref} and R_{ref} on figures 5.6 and 5.7 and the estimated signals L_{est} and R_{est} therefore contain all spatial information that is inherent to HRIR-filtered signals.

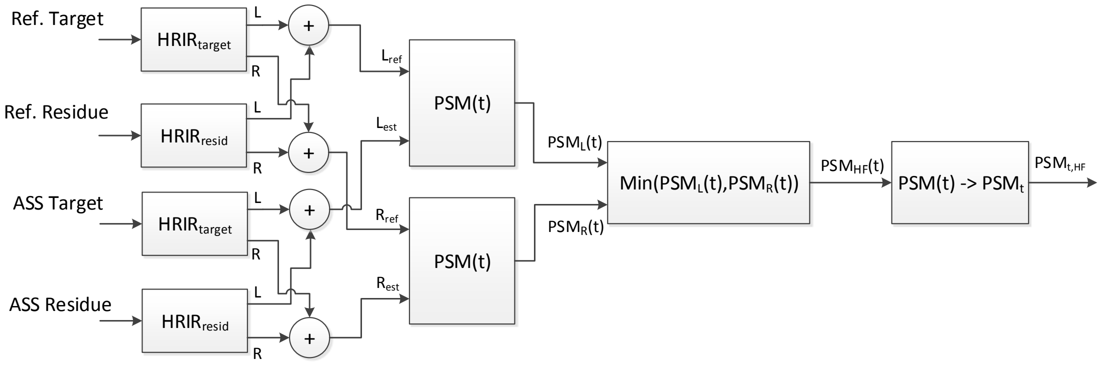


Figure 5.7 – Structure of the high-frequency part of the proposed binaural model. The HRIR blocks model the acoustic pathway between the source and the listener.

Low-frequency part

For the low-frequency part of the proposed model, the Perceptual Similarity Measure (PSM) is computed binaurally for both the reference and the estimated signals. The resulting PSM basically contains information about the similarity of the left and right channels for both cases through interaural cross-correlation. The low-frequency $\mathrm{PSM}_{t,\mathrm{LF}}$ is then computed as

$$\mathrm{PSM}_{t,\mathrm{LF}} = 1 - |\mathrm{PSM}_{t,\mathrm{ref}} - \mathrm{PSM}_{t,\mathrm{est}}|. \tag{5.27}$$

This measure is sensitive to differences between the binaural representations of the reference signals and the estimated signals. Note that the partial assimilation between the two signals as implemented in PEMO-Q does not make sense here. It is based on the assumption that missing components are less disturbing in the signal under test than additive components. This makes sense when comparing signals in the same channel. In the low-frequency part of the proposed binaural model, the left and the right channels of either the sound scene under test or the reference sound scene are compared. The “reference” and the “signal under test” (*i.e.* the two signals that are processed by one instance of the PEMO-Q block) are therefore from different channels and it does not make sense to prefer missing components in one channel as compared to the other.

High-frequency part

For the high-frequency part of the model, the PSM(t) is calculated in a more conventional way between the reference and estimated signals for both the left and right channels. To implement better-ear hearing, the minimal value between the left and right ear’s PSM(t) is kept in every temporal frame and filter band. In fact, better-ear hearing stipulates that if the reference and the estimate are at the same spatial location, spatial release from masking will be null, translating to a perfect correlation between the two signals at both ears (if spatial distortion is the only degradation occurring). If the estimate is moved away due to spatial distortions, spatial unmasking will occur [116] and the correlation between the reference and the estimate will therefore degrade in both ears. It will be the lowest in the ear that perceives the most spatial distortion. A low correlation between the reference and the estimate translates to high detectability in terms of better-ear hearing, hence the minimal value selection. The global $\mathrm{PSM}_{t,\mathrm{LF}}$ measure is then computed based on the resulting worst-case signal.

General remarks

The two extracted features (high- and low-frequency PSM_t: $\mathrm{PSM}_{t,\mathrm{HF}}$ and $\mathrm{PSM}_{t,\mathrm{LF}}$) can then be mapped to a subjective scale by means of linear regression analysis or Artificial Neural Network (ANN) training, depending on the relationship between the features and the subjective scores to be matched. Note that the boundary between the low-frequency and the high-frequency model parts f_b also is a parameter that can be optimised in combination with the mapping. The literature would suggest a limit between 1.5 and 2 kHz, since better-ear hearing mainly occurs above 2 kHz [116] and ILD provides information above 1.5 kHz [12]. In the presented model, this boundary was implemented by computing the low-frequency PSM based on the filter bands of the internal representations that were centered on frequencies below the given boundary, whereas the high-frequency PSM were computed with the filter bands of the internal representations with center frequencies above the boundary. All internal parameters of the PEMO-Q model were left at their default value, *i.e.* the optimised values reported by Vincent [131] for PEASS.

On a sidenote, Karjalainen stated that the precedence effect (the predominance of the first incoming wavefront in terms of source localisation) is partly modelled by the adaptation modelling [72] in his binaural model. If a first onset is barely compressed, subsequent onsets will be subject to a higher compression rate, lowering their perceptual importance. This is also expected to happen with any PEMO-Q-based model, since adaptation loops are part of this auditory model.

In terms of required computational power, this model implies that the auditory model of PEMO-Q is run on 4 signals. The high- and low-frequency features can then be computed based on the internal representations. The PEASS model requires 4 complete PEMO-Q model runs (5 for the extended version proposed in this thesis), *i.e.* the processing of 8 signals (10 for the extended version). Additionally, the preceding decomposition of the error term as presented in chapter 1 is required. The proposed model was implemented in MATLAB® based on the PEMO-Q version that was released as part of PEASS v.2 [1] and all values reported hereafter refer to this implementation.

[1] http://bass-db.gforge.inria.fr/peass/

5.4.3 Performance with subjective data from pre-study

The proposed binaural model was tested against the subjective scores collected during the pre-study. The boundary frequency between the high- and low-frequency parts of the model f_b was set to 1600 Hz and the mapping from the objective scale to the subjective scale was determined based on visual inspection. In fact, the relationship between the subjective scale and the objective scale seems to be quite linear even though the magnitude does not match. The objective scores of the high-frequency part of the model seem to decrease at roughly half the rate of the subjective scores. A mapping of $100(1-2(1-\text{PSM}_{t,\text{HF}}))$ was therefore chosen. For consistency, the same mapping $100(1-2(1-\text{PSM}_{t,\text{LF}}))$ was chosen for the low-frequency part. The mapping of the combined model therefore was $100(1-2(1-\text{PSM}_{t,\text{HF}})-2(1-\text{PSM}_{t,\text{LF}}))$. Figures 5.8, 5.9 and 5.10 report the matching between subjective and objective scores for all 3 degradation families in all 3 TRR levels for the low-frequency and the high-frequency parts of the model as well as the combined score. Note that this is based on the MOS across all 6 songs and all subjects for the subjective scores and on the mean objective score across all songs.

The relationship between the objective and the subjective scores seems to be linear and correlation coefficients between the subjective and objective scores for the different model parts are therefore calculated. For the low-frequency part of the model, a correlation of $\rho_{\text{LF}} = 0.56$ is measured. For the high-frequency part, $\rho_{\text{HF}} = 0.90$ and the combined correlation coefficient is $\rho_{\text{comb}} = 0.92$. The high-frequency part therefore seems to be more relevant in predicting perceived quality of the presented stimuli. Note however that the combined model benefits from the low-frequency part as well, since the correlation coefficient further increases.

Based on these excellent results, it is expected that the model also performs very well in other ASS evaluation situations. The stimuli of the pre-study being presented over headphones with the use of HRIRs, the binaural modelling stage seems to be a good strategy. However, the stimuli so far were spatialised only in front of the listener ($+30°, -30°$ azimuth) whereas the experiments presented in the following 2 chapters also involve other spatial configurations.

5.5 Conclusion

The artificial degradation synthesis algorithms presented in the first part of this chapter allow the computation of stimuli containing a controlled amount of different degradations as encountered in real-life ASS signals. This provides a tool for further investigation of the perceptual impact of the different degradation types, since controlled conditions are a must. The good results of a psychoacoustic pre-study about ASS evaluation with the artificial degradations legitimate the proposed algorithms.

Furthermore, a binaural modelling approach, based on an existing auditory model is proposed. Composed of two parts distinguished by the predominant perceptual cues at high and low frequencies, the model proposes a relatively simple but efficient way of modelling binaural phenomena such as binaural unmasking. The model is tested against the subjective scores of

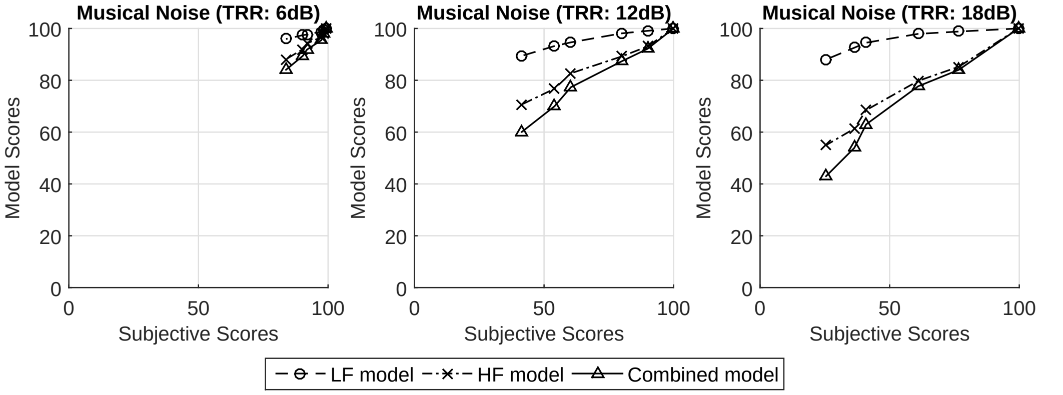


Figure 5.8 – Subjective scores plotted against objective scores for the two separate model parts as well as for the combined model for the musical noise-type artifact degradation family.

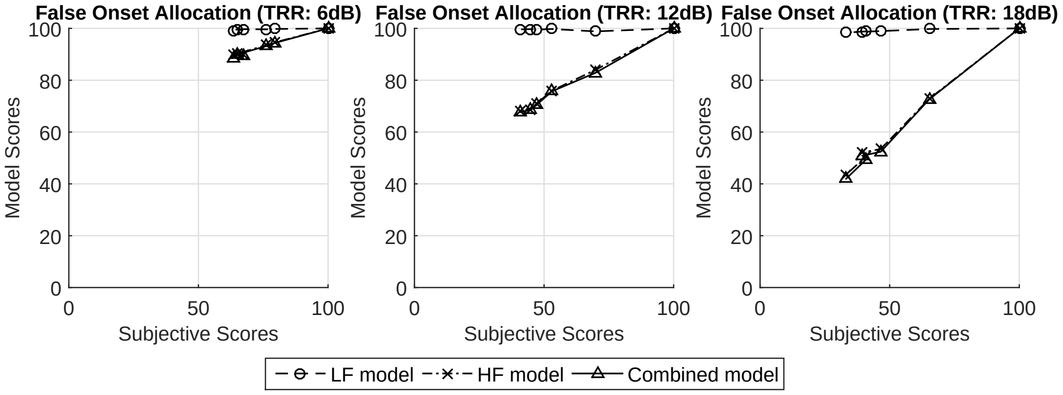


Figure 5.9 – Subjective scores plotted against objective scores for the two separate model parts as well as for the combined model for the onset misallocation degradation family.

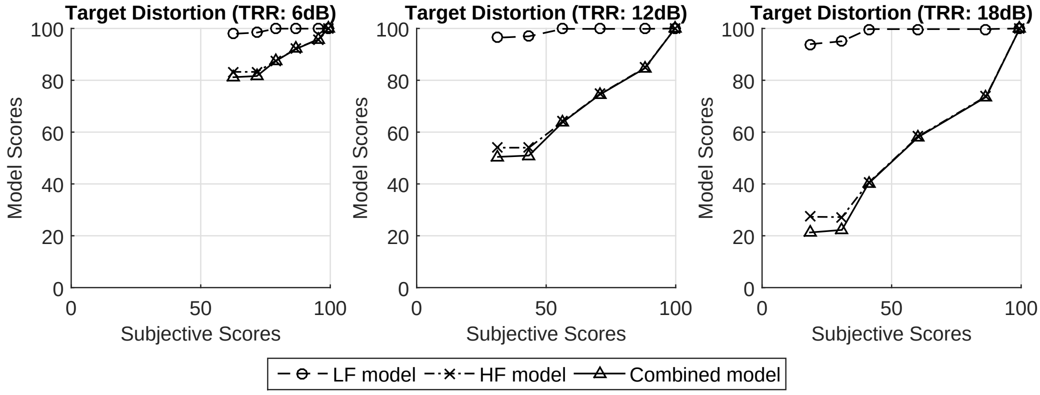


Figure 5.10 – Subjective scores plotted against objective scores for the two separate model parts as well as for the combined model for the target distortion degradation family.

the pre-study and a correlation coefficient of 0.92 is found between subjective and objective scores even with a manual objective to subjective scale mapping. This seems to validate the modelling approach.

The following two chapters present two studies in the context of 3D audio that were conducted to further assess the performance of the proposed model with the presented degradation synthesis algorithms. A more rigorous mapping is then employed, since the data becomes too complex to allow a mapping based on visual inspection.

6 ASS evaluation in a binaural rendering context

6.1 Introduction

Following the user feedback and the conclusions of the pre-study, a subjective study was conducted with 2 objectives: assess the relative importance of the degradation families and collect subjective data for the proposed modelling approaches. For this first study, binaural spatialisation is used with Interaural Time Difference (ITD) synthesis only, in order to have a very controlled spatialisation setting. The binaural synthesis approach was chosen in order to minimise material requirements while gaining first insights into subjective and objective Audio Source Separation (ASS) evaluation in the context of spatial audio.

This chapter first reports the design and method of this binaural study, followed by the detailed analysis and discussion of the subjective results.
The last part of this chapter then reports the scores of the objective ASS evaluation modelling approaches presented in chapters 3 and 5 against the subjective results of the psychoacoustic study, together with a brief discussion.

6.2 Experimental design

The design of this study is similar to what was done in the pre-study with a few modifications. Since the relative importance of the proposed degradation families is to be assessed, a common scale for the degradations is needed and therefore stimuli with several degradation types simultaneously. Additionally, the location of the target should be variable in order to assess the importance of binaural listening phenomena. For this second experiment, there were therefore 7 factors being considered: Target to Residue Ratio (TRR), spatial configuration, musical excerpt, onset misallocation level, target distortion level, musical noise artifacts level and last but not least, the number of the participant. Note that the static interference degradation is not considered (as in the pre-study). The 3 other degradations are reckoned to influence perceived quality more in a spatial audio context and it would further increase the number of stimuli. To reduce the number of possible combinations, only 2 TRR levels, 2 spatial configurations, 5 musical excerpts and 4 degradation levels of each type were considered. Even with these simplifications though, the number of combinations per participant is prohibitive if a full factorial design was considered ($2 \cdot 2 \cdot 5 \cdot 4^3 = 1280$ stimuli, not counting references and anchors). A D-optimal design was therefore constructed, reducing the number of measurements per participant while maximising the determinant of the information matrix of the experiment (see Eriksson *et al.* [43] *e.g.* for an introduction). The design was computed with the MATLAB® software and its coordinate exchange algorithm contained in the `cordexch` function. Constraints were a linear model with interactions, 20 participants and 3200 runs in total, aiming for 160 stimuli per participant. The levels of all factors were predetermined.

Since the experimental protocol was the same as in the pre-study, *i.e.* a modified Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) protocol after Emiya *et al.* [42], the rows of the resulting design matrix (runs) were grouped by experimental condition regarding TRR, spatial configuration, musical excerpt and participant number. Because of the graphical test interface allowing for only a limited number of stimuli being presented together on the same interface page, every group was then segmented into pages of 6 runs or less (see the next section for a description of the graphical user interface).

Since a D-optimal design does not use the same runs for every participant and therefore not necessarily the same number of runs for every participant, there were notable differences in number of runs across participants. In order to exclude an effect due to different fatigue states across participants, the number of pages and runs was equalised across participants. Constraining a combinatory algorithm to use pages from the precedent participant to complete the design matrix for every participant (in a cyclic fashion, *i.e.* using pages from the last participant for completing the first one), the smallest achievable number of pages was determined to be 42. The number of runs per participant was 191 in that case. A reference and an anchor was then added to every page, adding 84 runs for every participant, summing up to a total of 275 runs per participant.

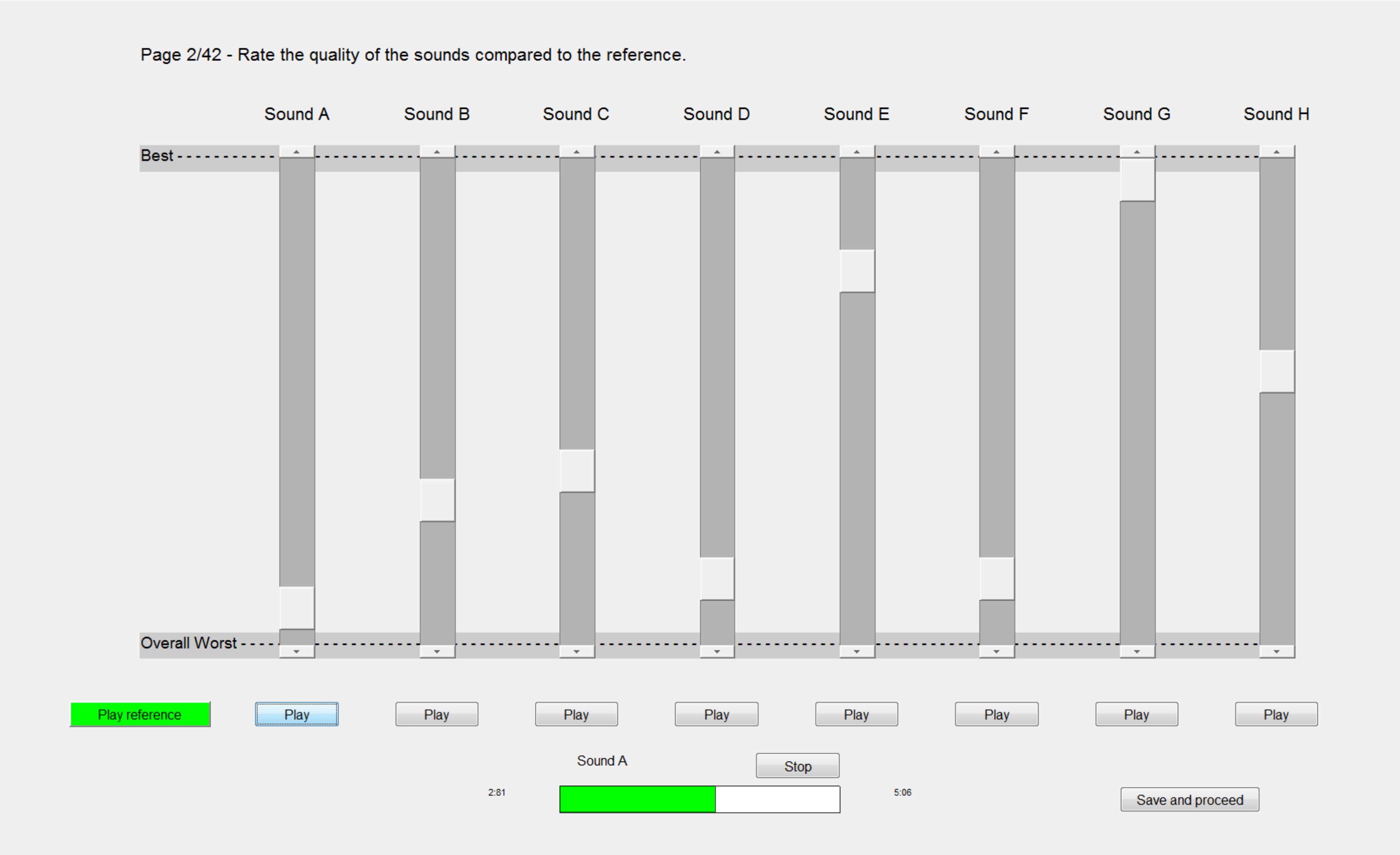
A total of 20 participants took part in this study (2 females, 18 males). They were not experts in audio quality evaluation, but all had experience with audio and critical listening in general. All participants reported normal hearing (no known impairments) although no audiological measurements were made. They were not remunerated for their participation.

6.2.1 Experimental task

As for the pre-study, the binaural study was based on a modified MUSHRA protocol. A training phase preceded the evaluation phase as in the pre-study. To shorten the training phase and to take into account the observed behavior of the pre-study, only the reference and the most degraded stimulus for each test page were presented during the first step of the training phase, where participants could listen to the stimuli to get a sense of the involved scale of quality. Participants could then get accustomed to the evaluation interface with one evaluation example (*i.e.* one page).

The stimuli selected for each participant as a result of the D-optimal design were grouped such that a single evaluation test page showed only stimuli of the same song, TRR and spatial configuration. As opposed to the pre-study, the evaluation test pages contained a low anchor in this second experiment which was synthesised as described in the next section. As in the pre-study, participants were asked to rate the quality of all sound excerpts on a page (including a low anchor and a hidden reference) as compared to a labelled reference using a graphical interface on a computer. The sliders used to rate the quality of each stimuli did no show the current rating any more, only displaying “best” at the top of the scale and “overall worst” at the bottom. This is due to the fact that some participants were rounding their ratings towards the next multiple of 10 during the pre-study. The rest of the evaluation interface was kept the same with some minor improvements in the audio playback handling. A screenshot of the graphical user interface is shown in figure 6.1. Participants could listen to all stimuli and the reference as often as they wanted. In fact, Thiede *et al.* stated “Subjective listening tests that evaluate audio quality depend primarily on both the short-term and the echoic memory stores in order to detect small differences between signals. This limitation is accommodated in the ITU-R recommendation (ed. note: ITU-R BS.1116 [63]) for the subjective listening test procedure by allowing test subjects to loop through short subsequences and to switch at will between reference and test signals in the listening test”. This later was also recommended in ITU-R recommendation BS.1534 defining the MUSHRA protocol [64]. On average, participants took about 1 hour to complete the experiment.

In order to try to keep the perceptual scale as relevant as possible, the training page containing references and low anchors of all pages was shown to the listener after every 5 pages during the evaluation phase. Since a large number of stimuli were used for every participant, this should limit the influence of participants changing the low end of the subjective scale due to forgetting what the worst overall stimulus sounds like.



Page 2/42 - Rate the quality of the sounds compared to the reference.

Figure 6.1 – Evaluation mask of the graphical user interface used for the subjective studies.

6.2.2 Stimuli

The musical excerpts from the pre-study were further shortened to respond to the feedback of the participants and to further reduce the length of the experiment. The excerpt of Kanye West was completely eliminated, since it was found to present artistic effects which were partially similar to the degradation types under investigation. In terms of audibility, this represents a case were ASS degradations may be masked by the content of the original material. The new excerpt data are given in table 6.1. The 5 excerpts were equalised in loudness before the degradation synthesis, in order to minimize the risk of degradations being more easily detectable for certain excerpts due to increased loudness.

The amount arguments for the degradation synthesis algorithms were chosen such that the resulting degradations were clearly audible at their maximum setting in all cases. Their values were the following:

- Musical noise artifacts:

$$a_{\text{musical noise}}[k] = \begin{cases} 0 & \text{if } k = 0 \\ \frac{1}{4^{(3-k)}} & \text{if } 1 \leq k \leq 3 \end{cases} \tag{6.1}$$

- Target distortion:

$$a_{\text{target dist}}[k] = 500 \cdot 2^{\frac{3-k}{3} \log_2 \frac{22050}{500}} \qquad \text{with } 0 \leq k \leq 3 \tag{6.2}$$

- Onset misallocation:

$$a_{\text{onset}}[k] = \begin{cases} 0 & \text{if } k = 0 \\ 0.2k & \text{if } 1 \leq k \leq 2 \\ 1 & \text{if } k = 3 \end{cases} \tag{6.3}$$

The low anchor presented on each test page is defined by the same spatial configuration, musical excerpt and TRR as the stimuli under test, but contains all degradation types at their maximal level (*i.e.* the maximal level used in the stimuli). The degradation synthesis algorithms were applied sequentially, beginning with the target distortion, continuing with musical noise artifacts and finishing with onset misallocation. This order was chosen based on trials with different orders. If the musical noise artifacts or the onset misallocation degradations were synthesised first, the target distortion might filter them out again. Also, if the musical noise artifacts synthesis algorithm came after the onset misallocation synthesis, the time-frequency coefficient selection algorithm would be biased by the onsets.

This experiment was also conducted to test the degradation audibility modelling approach that is presented in section 5.4.2. For that reason, a very precise control over the spatialisation parameters was needed. This was achieved by not employing Head Related Transfer Functions

Table 6.1 – Musical excerpts used for the stimuli in the binaural and WFS studies

#	Artist	Title	Start	End	Duration
1	Beyoncé	End of Time	2:06.0	2:11.3	0:5.3
2	Phoenix	Lisztomania	1:42.1	1:47.2	0:5.1
3	Shannon Hurley	Sunrise	1:14.0	1:21.3	0:7.3
4	The Ultimate NZ Tour	Four Good Reasons	0:14.2	0:18.7	0:4.5
5	Shannon Hurley	We are in love	2:45.4	2:54.4	0:9.0

(HRTFs) as in the pre-study, but by simulating spatial positions through ITD modelling. ITDs were based on the model by Larcher & Jot [80] and two spatial configurations were used: In the first configuration, both the residue and the target were placed in a standard stereo configuration (virtual loudspeakers at −30° and 30° azimuth) like for the pre-study. In the second configuration, the target was moved to be centered at 270° azimuth (90° to the right of the listener), virtual loudspeakers therefore being located at 300° and 240° azimuth.
TRR levels were chosen to be 6 dB and 12 dB. The 18 dB TRR level from the pre-study was too drastic and caused too much loudness variation as compared to the 6 dB TRR condition and was therefore not retained.

6.2.3 Experimental setup

Due to practical constraints, the experiment was conducted in different locations with controlled acoustic conditions equivalent to listening booths. This may not be optimal in terms of repeatability, but since it is a binaural study conducted using headphones, the results should not be affected too much. The employed headphones were the same as in the pre-study (AKG K240) with the associated Headphone Transfer Function (HpTF).

To enable the comparison of the results between different sessions and participants, the output level was calibrated to attain 79.8 dB(A) in an artificial ear when playing back a sinusoidal reference sound at -6 dB(FS). This resulted in stimuli being played back at about 53 dB(A) which seemed comfortable for prolonged listening sessions without making degradations inaudible. For subjects who required it, the playback level could be slightly adjusted, as recommended by the International Telecommunication Union (ITU) recommendation ITU-R BS 1116 [63]. This approach is also taken in ITU-R BS 1534 [64] where the MUSHRA protocol is defined.

6.3 Results

6.3.1 Post-screening

The scores of the hidden reference stimuli included in the rating pages were used to detect participants who did not properly understand the task. Since the task was to rate the quality

of test sounds as compared to a given reference on each page, it is expected that the hidden reference that did not undergo any degradation processing yields perfect quality (perfect score) when comparing to the reference.
The grand mean μ across participants and pages of the subjective scores of the hidden reference signals was computed. The Euclidian distance of all participants to μ was then computed. The mean distance across participants is $D_\mu = 47.2$. An arbitrary critical distance $D_c = 100$ was then defined based on visual inspection of the resulting distribution. Results of participants whose distances were above the critical distance were discarded. Based on this criterion, the results of 3 participants had to be discarded. Discarded participants had distances of $D_4 = 114.8$, $D_9 = 112.3$ and $D_{19} = 140.2$. The mean distance amongst remaining participants is $D = 33.9$.

The low anchor scores were not used in the post-screening. Even though they should correspond to the signals with the worst scores on any page, it is not expected to have a consensus among participants concerning the range of the score.

6.3.2 Statistical analysis

The results of the study are analysed using Linear Mixed Models (LMMs). A short introduction to LMMs is given in appendix A. For a more complete introduction to LMMs, see Bates [6] or West *et al.* [140]. The notation of West *et al.* [140] is adopted here.

Statistical model building

In the presented binaural experiment, there were 6 covariates related to the stimuli: the Target to Residue Ratio “TRR” (6 or 12 dB), the spatial position “POS” (0° or 270°), the musical excerpt “SONG” (1 out of 5 possible), the amount of target distortion degradation (*i.e.* the low-pass cut-off frequency) “DEGDIST”, the amount of musical noise-type artifacts “DEGARTIF” and the amount of onset degradation “DEGONSET”. The participant number “PART” is another covariate. Moreover, the stimuli were presented grouped by pages, which introduces the last covariate, the page number “PAGE”.

Before the analysis, the DEGDIST, DEGARTIF, DEGONSET, TRR and POS covariates were scaled and centered. This allows for easier convergence of the optimisation algorithms, since the covariates are on very different scales. Additionally, the 3 degradation amount arguments were transposed to a logarithmic scale before scaling and centering to account for the behavior observed during the pre-study (perceived quality degrading inversely proportional to the amount arguments). For degradations including a zero-valued amount argument, 0.1 was added to all amount argument levels before the change of scales to avoid singularity.

In terms of data structure, this results in a three-level clustered data model. Stimuli, which are the unit of analysis, are clustered into pages, which are clustered into participants. Note that the PAGE factor is nested within the PART factor. The degradation covariates DEGDIST,

Table 6.2 – Overview of the nested structure of the data in the binaural study as well as the WFS study

Level of data		**Effects**
Level 3	Cluster of clusters (random factor)	Participant
Level 2	Cluster of units (random factor)	Page
	Covariates	TRR, spatial configuration, song
Level 1	Unit of analysis	Stimulus
	Dependent variable	Subjective score (quality)
	Covariates	Degradation amounts

DEGARTIF and DEGONSET are level 1 covariates, varying within pages. The SONG, TRR and POS covariates are level 2 covariates, varying across pages. This structure is summarised in table 6.2.

Two approaches can be taken for linear mixed models: the step-up strategy, adding effects step by step from a simple model towards a more complex one and testing hypotheses along the way; or the top-down strategy, starting with a full model specification and then simplifying the structure as seems fit [140]. Since the data that is to be fitted here is rather complex, the step-up approach is taken, following the guidelines given by West *et al.* [140]. All models are fitted using the R software, using the `lme4` package (see book draft by Bates [6] for a complete reference). The `lmerTest` package was used to obtain test statistics for fixed covariates.
The analysis steps can be summarised as follows:

1. Fit the initial "unconditional" (variance components) model (Model 1)

2. Build the level 1 model by adding level 1 covariates (Model 2)

3. Build the level 2 model by adding level 2 covariates (Model 3)

Variants of the three models can be used to test specific hypotheses about included effects. A general model for the data, including interactions and random coefficients can be written as

$$
\begin{aligned}
&\text{RESPONSE}_{ijkl} = \\
&\beta_0 + \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\
&+ \beta_4 \times \text{TRR}_{jkl} + \beta_5 \times \text{POS}_{jkl} + \\
&+ \beta_6 \times \text{DEGDIST}_{ijkl} \times \text{DEGARTIF}_{ijkl} + \beta_7 \times \text{DEGDIST}_{ijkl} \times \text{DEGONSET}_{ijkl} + \\
&+ \beta_8 \times \text{DEGARTIF}_{ijkl} \times \text{DEGONSET}_{ijkl} + \beta_9 \times \text{DEGDIST}_{ijkl} \times \text{TRR}_{jkl} + \\
&+ \beta_{10} \times \text{DEGARTIF}_{ijkl} \times \text{TRR}_{jkl} + \beta_{11} \times \text{DEGONSET}_{ijkl} \times \text{TRR}_{jkl} + \\
&+ \beta_{12} \times \text{DEGDIST}_{ijkl} \times \text{POS}_{jkl} + \beta_{13} \times \text{DEGARTIF}_{ijkl} \times \text{POS}_{jkl} + \\
&+ \beta_{14} \times \text{DEGONSET}_{ijkl} \times \text{POS}_{jkl} + \beta_{15} \times \text{TRR}_{jkl} \times \text{POS}_{jkl} + \\
&+ \beta_{16} \times \text{DEGDIST}^2_{ijkl} + \beta_{17} \times \text{DEGARTIF}^2_{ijkl} + \beta_{18} \times \text{DEGONSET}^2_{ijkl} + \\
&+ u_{0l} + u_{1l} \times \text{DEGDIST}_{ijkl} + u_{2l} \times \text{DEGARTIF}_{ijkl} + u_{3l} \times \text{DEGONSET}_{ijkl} + \\
&+ u_{4l} \times \text{TRR}_{jkl} + u_{5l} \times \text{POS}_{jkl} + \\
&+ u_{0k|l} + u_{1k|l} \times \text{DEGDIST}_{ijkl} + u_{2k|l} \times \text{DEGARTIF}_{ijkl} + u_{3k|l} \times \text{DEGONSET}_{ijkl} + \\
&\qquad\qquad + u_{j|l} + \epsilon_{ijkl}. \qquad (6.4)
\end{aligned}
$$

In this model specification, RESPONSE_{ijkl} represents the score assigned to stimulus i on page j for song k by participant l and β_0 - β_{15} represent the fixed intercept and the fixed effects associated with the covariates (DEGDIST, DEGARTIF, DEGONSET, TRR, POS and their interactions). u_{0l} represents the random intercept associated with participant l. u_{1l} through u_{5l} represent the random slopes associated with the different covariates DEGDIST, DEGARTIF, DEGONSET, TRR and POS for participant l. $u_{0k|l}$ represents the random effect with the random intercept for song k within participant l and $u_{1k|l}$ - $u_{3k|l}$ the random slopes associated with the covariates DEGDIST, DEGARTIF and DEGONSET for song k within participant l. Finally, $u_{j|l}$ represents the random intercept for page j within participant l and ϵ_{ijkl} represents the residual.

The distribution of the random effects associated with the participants in this model is supposed to be multivariate normal and can be written as

$$
u_l = \begin{pmatrix} u_{0l} \\ u_{1l} \\ u_{2l} \\ u_{3l} \\ u_{4l} \\ u_{5l} \end{pmatrix} \sim \mathcal{N}(0, \mathbf{D}_{\text{participant}}), \qquad (6.5)
$$

where $\mathbf{D}_{\text{participant}}$ is the matrix of variances and covariances between the different random effects with variances on the diagonal and covariances off the diagonal.
In a similar fashion, the distribution of the random effects associated with the songs given a

participant is also supposed to be multivariate normal and can be written as

$$u_k|l = \begin{pmatrix} u_{0k|l} \\ u_{1k|l} \\ u_{2k|l} \\ u_{3k|l} \end{pmatrix} \sim \mathscr{N}(0, \mathbf{D}_{\text{song|participant}}), \tag{6.6}$$

where $\mathbf{D}_{\text{song|participant}}$ is the matrix of variances and covariances between the different random effects.
Since there is only a random intercept associated with pages given a participant, a normal distribution is assumed:

$$u_{j|l} \sim \mathscr{N}(0, \sigma^2_{\text{page|participant}}), \tag{6.7}$$

Finally, the residual is assumed to follow a normal distribution with variance σ^2:

$$\epsilon_{ijkl} \sim \mathscr{N}(0, \sigma^2). \tag{6.8}$$

The summary of the 3 models that are considered in this analysis is given by table 6.3.

To avoid a bias of the statistical model towards higher scores because of all the scores corresponding to hidden references, they were excluded from the data for the statistical analysis. Moreover, since the Quantile-Quantile (Q-Q) plot of the residuals showed large evidence for a heavy-tailed distribution in a first fit of the model, all scores equal to 0 and 100 were excluded from the reported statistical models. For a discussion, see section 6.3.5.

Model 1

The first model is a fixed intercept model with random effects associated with the intercepts for pages (level 2) and participants (level 3). Since the musical excerpt is also modelled as a random effect, random intercepts for songs (level 2) are also included in this first model. This can be written as

$$\text{RESPONSE}_{ijkl} = b_{0ijkl} + b_{0jk|l} + b_{0l} + \beta_0 + u_{0l} + u_{0k|l} + u_{0j|l} + \epsilon_{ijkl}, \tag{6.9}$$

where b_{0l} represents unobserved random slopes at the participant level, $b_{0jk|l}$ represents unobserved fixed and random intercepts and slopes at the page level and b_{0ijkl} represents the yet unspecified interactions at the stimulus level.

This model allows to test two hypotheses regarding the variance of the page-specific and the song-specific intercepts. They assess if the two random effects associated with the page and the song are useful in the model specification. In the case of the page-specific intercept, the

Table 6.3 – Summary of the 3 models considered during the statistical analysis and the step-up process

		Term / Variable	Notation	Model		
				1	2	3
Fixed effects		Intercept	β_0	✓	✓	✓
		DEGDIST	β_1		✓	✓
		DEGARTIF	β_2		✓	✓
		DEGONSET	β_3		✓	✓
		TRR	β_4			✓
		POS	β_5			✓
		DEGDIST×DEGARTIF	β_6			✓
		DEGDIST×DEGONSET	β_7			✓
		DEGARTIF×DEGONSET	β_8			✓
		DEGDIST×TRR	β_9			✓
		DEGARTIF×TRR	β_{10}			✓
		DEGONSET×TRR	β_{11}			✓
		DEGDIST×POS	β_{12}			✓
		DEGARTIF×POS	β_{13}			✓
		DEGONSET×POS	β_{14}			✓
		TRR×POS	β_{15}			✓
		DEGDIST^2	β_{16}			✓
		DEGARTIF^2	β_{17}			✓
		DEGONSET^2	β_{18}			✓
Random effects	Page (j)	Intercept	$u_{0j\|l}$	✓	✓	✓
	Song (k)	Intercept	$u_{0k\|l}$	✓	✓	✓
		DEGDIST	$u_{1k\|l}$		✓	✓
		DEGARTIF	$u_{2k\|l}$		✓	✓
		DEGONSET	$u_{3k\|l}$		✓	✓
	Participant (l)	Intercept	u_{0l}	✓	✓	✓
		DEGDIST	u_{1l}		✓	✓
		DEGARTIF	u_{2l}		✓	✓
		DEGONSET	u_{3l}		✓	✓
		TRR	u_{4l}			
		POS	u_{5l}			
Residuals	Stimulus (i)		ϵ_{ijkl}	✓	✓	✓

null hypothesis H_0 and the alternative hypothesis H_A are:

$$H_0 : \sigma^2_{\text{int:PAGE}} = 0$$
$$H_A : \sigma^2_{\text{int:PAGE}} > 0$$

To test this hypothesis, a nested model 1A is constructed, not including the random intercept for pages. A likelihood ratio test is then conducted, subtracting the -2 Maximum Likelihood (ML) log-likelihood for model 1 from the corresponding value of model 1A based on ML fits. To obtain a p-value for this test statistic, it is referred to a mixture of χ^2 distributions with 0 and 1 degrees of freedom and equal weights 0.5 (following the procedure in West *et al.* [140]). Since the result is significant ($\chi^2(0:1) = 199.73, p < .001$), the alternative hypothesis is confirmed. The random intercept associated with the page is therefore kept for all the following models. In the case of the song-specific intercept, the same procedure is followed, with the null and alternative hypotheses being:

$$H_0 : \sigma^2_{\text{int:song}} = 0$$
$$H_A : \sigma^2_{\text{int:song}} > 0$$

To test this hypothesis, another nested model 1B is fitted, not including the random intercept for songs. The likelihood ratio test is once again significant ($\chi^2(0:1) = 67.43, p < 0.001$) and the random intercept associated with the song is also kept in the following models.

Model 2

For the second model, the level 1 fixed effects (DEGDIST, DEGARTIF, DEGONSET) are first added to model 1. This translates to a model written as

$$\begin{aligned}
&\text{RESPONSE}_{ijkl} = b_{0jk|l} + b_{0ijkl} + \beta_0 + \\
&+ \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\
&\qquad\qquad\qquad\qquad + u_{0l} + u_{0k|l} + u_{0j|l} + \epsilon_{ijkl},
\end{aligned} \tag{6.10}$$

where $b_{0jk|l}$ represents unobserved fixed and random intercepts at the page level and b_{0ijkl} represents the yet unspecified interactions at the stimulus level. Once again, a likelihood ratio test is conducted, based on ML fits of models 1 and 2 to test for the hypothesis that those effects are null. The null and alternative hypotheses therefore are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta 4 = 0$$
H_A : At least one level 1 fixed effect is not equal to zero.

To test for significance of the alternative hypothesis, the -2 ML log-likelihood of model 2 is subtracted from its counterpart in model 1. Under the null hypothesis, this test statistic is asymptotically following a χ^2 distribution with 3 degrees of freedom. Since the test is significant ($\chi^2(3) = 1354.9, p < .001$), the alternative hypothesis is preferred and the level 1 covariates are added.

The level 1 covariates are then added as level 3 random effect coefficients, allowing for random slopes for the degradation amounts in every participant. This translates to different participants having different sensitivities for the different degradation families. The null and alternative hypotheses to test therefore are:

$H_0 : \sigma^2_{\text{DEGDIST:PART}} = \sigma^2_{\text{DEGARTIF:PART}} = \sigma^2_{\text{DEGONSET:PART}} = 0$
H_A : At least one of the variances is not zero.

The test statistic follows a χ^2 distribution with 9 degrees of freedom (one for every additional variance / covariance parameter to be estimated when comparing to a simple intercept). Since the likelihood ratio test is significant ($\chi^2(9) = 155.52, p < .001$), the alternative hypothesis is preferred and the random coefficients associated with the degradation levels at the participant level are therefore kept in the model.

As a second improvement to this second model, the level 1 covariates are also added as random coefficients to the SONG random effect. This translates to model 2 being written as

$$\begin{aligned}
&\text{RESPONSE}_{ijkl} = b'_{0jk|l} + b'_{0ijkl} + b'_{0l} + \beta_0 + \\
&+ \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\
&+ u_{0l} + u_{1l} \times \text{DEGDIST}_{ijkl} + u_{2l} \times \text{DEGARTIF}_{ijkl} + u_{3l} \times \text{DEGONSET}_{ijkl} + \\
&+ u_{0k|l} + u_{1k|l} \times \text{DEGDIST}_{ijkl} + u_{2k|l} \times \text{DEGARTIF}_{ijkl} + u_{3k|l} \times \text{DEGONSET}_{ijkl} + \\
&\hfill + u_{0j|l} + \epsilon_{ijkl}, \qquad (6.11)
\end{aligned}$$

where b'_{0l} represents unobserved random slopes at the participant level, $b'_{0jk|l}$ represents unobserved fixed intercepts at the page level and b'_{0ijkl} represents the yet unspecified interactions at the stimulus level. The null and alternative hypotheses thus are:

$H_0 : \sigma^2_{\text{DEGDIST:SONG}} = \sigma^2_{\text{DEGARTIF:SONG}} = \sigma^2_{\text{DEGONSET:SONG}} = 0$
H_A : At least one of the variances is not zero.

Once again, the statistic for the according likelihood ratio test follows a χ^2 distribution with 9 degrees of freedom. Since the test is significant ($\chi^2(9) = 222.21, p < .001$) the random coefficients are kept. This accounts for different degradation families having different effects for different songs. The degradation covariates are not added to the page random effect, because having different effects for degradation on different pages does not make sense in this analysis.

Model 3

For the third model, the level 2 fixed effects (POS, TRR) are added as fixed effects. To see if adding these effects is useful in this model, the following null versus alternative hypotheses are tested:

$H_0 : \beta_4 = \beta_5 = 0$
H_A : At least one fixed effect is not zero.

The resulting test statistic follows a χ^2 distribution with 2 degrees of freedom and the test is significant ($\chi^2(2) = 438.52, p < .001$), which means that at least one of the effects can be assumed to be different from zero.

The level 2 covariates are then added as random coefficients at level 3, allowing for different effects of the TRR and the spatial configuration for every participant. A likelihood ratio test is conducted to see if at least one of the added variances is not equal to zero. The resulting test statistic follows a χ^2 distribution with 11 degrees of freedom. The test is not significant ($\chi^2(11) = 18.509, p = 0.07$), and the null hypothesis of all variances being zero can therefore not be rejected. The level 2 covariates POS and TRR are therefore not kept as random coefficients at level 3 (effectively making $u_{4l} = 0$ and $u_{5l} = 0$).

The second-order degradation effects are then added (DEGDIST^2, DEGARTIF^2, DEGONSET^2), since it is expected that the degradation amounts may affect the audibility of the corresponding degradations in a quadratic way despite the logarithmic scaling of the input values. A likelihood ratio test for the statistical significance of at least one of the associated intercepts reveals that the second-order terms should be kept in the model ($\chi^2(3) = 55.78, p < .001$).

Finally, since it is expected that interactions between the different degradation families may occur, or that certain degradation may be more harmful in terms of audio quality depending on spatial configuration or TRR, interactions of the fixed effects are added to the model. The null hypothesis of all interactions being zero is rejected ($\chi^2(10) = 183.12, p < .001$). The full model as expressed by equation (6.4) is therefore reached, except for the random POS and TRR slopes at the participant level.

6.3.3 Fixed-effect parameter estimates

The literature references do not agree on how the test statistic of the fixed-effect parameters is distributed. This is why the precedent section used only -2 ML log-likelihood ratio tests. However, approximate tests can be conducted following different assumptions. Here, the `lmerTest` package of the R software was used to compute p-values based on t-statistics using the Satterthwaite approximations to degrees of freedom. The results of the corresponding tests, as well as estimates for the fixed-effect parameters are reported in table 6.4.

6.3.4 Covariance parameter estimates and Intraclass Correlation Coefficients

Table 6.5 reports the estimated variance components for the random effects of the final model. Corresponding standard deviations are also reported.

Table 6.4 – Estimates of the fixed-effect parameters in model 3. Note that all independent variables reported here had been standardised (scaled and centered) before fitting the model.

Fixed-effect parameter	Estimate	Std. Error	df	t-value	p-value
β_0 (Intercept)	44.81	4.48	12.3	10.01	<.001
β_1 (DEGDIST)	12.07	2.86	5.1	4.23	<.01
β_2 (DEGARTIF)	-3.26	0.80	10.7	-4.06	<.01
β_3 (DEGONSET)	-4.74	1.29	6.9	-3.69	<.01
β_4 (TRR)	-9.49	0.42	665.9	-22.67	<.001
β_5 (POS)	-4.64	0.43	710.7	-10.83	<.001
β_6 (DEGDIST×DEGARTIF)	0.47	0.36	2877.6	1.30	0.19
β_7 (DEGDIST×DEGONSET)	-3.90	0.37	2861.1	-10.48	<.001
β_8 (DEGARTIF×DEGONSET)	0.58	0.36	2929.2	1.62	0.11
β_9 (DEGDIST×TRR)	0.12	0.32	2866.3	0.363	0.72
β_{10} (DEGARTIF×TRR)	-0.56	0.32	2863.9	-1.75	0.08
β_{11} (DEGONSET×TRR)	-0.54	0.31	2891.9	-1.71	0.09
β_{12} (DEGDIST×POS)	0.20	0.35	2862.2	0.56	0.58
β_{13} (DEGARTIF×POS)	0.74	0.35	2853.5	2.10	<.05
β_{14} (DEGONSET×POS)	-0.85	0.34	2801.7	-2.53	<.05
β_{15} (TRR×POS)	2.40	0.42	716.7	5.66	<.001
β_{16} (DEGDIST^2)	-0.81	0.42	2856.6	-1.94	0.05
β_{17} (DEGARTIF^2)	-1.93	0.43	2849.1	-4.47	< .001
β_{18} (DEGONSET^2)	2.06	0.38	2869.1	5.48	<.001

Table 6.5 – Estimates of the variance components in model 3

Random effect parameter	Variance	Std. dev.
$u_{j\|l}$ (Intercept for PAGE)	68.86	8.30
$u_{0k\|l}$ (Intercept for SONG)	52.04	7.21
$u_{1k\|l}$ (DEGDIST slope for SONG)	35.34	5.95
$u_{2k\|l}$ (DEGARTIF slope for SONG)	1.37	1.17
$u_{3k\|l}$ (DEGONSET slope for SONG)	5.62	2.37
u_{0l} (Intercept for PART)	154.47	12.43
u_{1l} (DEGDIST slope for PART)	16.13	4.02
u_{2l} (DEGARTIF slope for PART)	4.25	2.06
u_{3l} (DEGONSET slope for PART)	6.40	2.53

While the estimated variance components of the full model may be of interest for the ICC reported below, the evolution of the variance components throughout the model building process may also be of interest. In this case, the addition of the degradation amounts as fixed effects (first step from model 1 to model 2) reduces the estimated residual variance by 42.4% (estimated residual variance of 493.3 in model 1 vs. 283.99 after adding DEGDIST - DEGONSET). The variance at the page level however is increased by 42.8% and the variance at the participant level is increased by 30.4%. The fixed effects DEGDIST - DEGONSET therefore explain a lot of the variation at the stimulus level, but they don't explain variation among participants or pages / songs which is even highlighted in their presence , *i.e* some variance is attributed to the random effects rather than the added fixed effects.
Introducing random slopes for the degradation amounts at the participant level further reduces the residual variance by 8.0%. The variance increases by 1.3% at the page level and by 16.3% at the participant level. Including these random slopes therefore explains some of the random variance at the stimulus level without explaining the random variation between participants or pages / songs. Including the same random slopes for the SONG random effect diminishes the residual variance by another 8.9% but increases the variance at the page level by 19.7%. The variance at the participant level also raises by 2.0%. Once again, this explains some of the variance at the stimulus level but attributes it to higher level components.
When adding the covariates POS and TRR at the page level, the variance at that level is diminished by 42.9%. The residual variance is unaffected whereas the participant-level variance is augmented by 7.3%. A lot of the variance at the page level is therefore explained by the TRR and POS effects.
Adding the squared terms for DEGDIST to DEGONSET diminishes the residual variance by 2.0% and leaves the page-level and participant-level variances unaffected. Finally, adding the different interactions between the first-order fixed effects diminishes the residual variance by 5.6% but adds 0.5% to the participant-level variance and 2.8% to the participant-level variance. Overall, the different components added during the model building process affect positively (*i.e.* diminish) the random variance at the expected levels (level 1 covariates affect residual variance, level 2 covariates affect page-level variance).

For the final model, ICCs are used to measure the homogeneity of the responses within a given cluster. For the presented statistical model, two ICCs are of interest: one for the participant level, and one for the page level.
The participant level ICC can be written as

$$\mathrm{ICC}_{\mathrm{part}} = \frac{\sigma^2_{\mathrm{part}}}{\sigma^2_{\mathrm{part}} + \sigma^2_{\mathrm{page}} + \sigma^2_{\mathrm{song}} + \sigma^2}, \tag{6.12}$$

where σ^2_{part} is the sum of the variance components associated with the participant, σ^2_{page} is the variance of the random effect associated with the page nested within participants, σ^2_{song} is the sum of the variance components associated with the song and σ^2 is the residual variance. This ICC is high if the total random variation is dominated by the variance of the random participant effect, *i.e.* if the results from a participant are relatively homogeneous, but they

vary widely from participant to participant. For the model specified above, the value of this ICC is $\text{ICC}_{\text{part}} = 32.0\%$.
The second ICC that can be defined is associated with the page. It is defined as

$$\text{ICC}_{\text{page}} = \frac{\sigma^2_{\text{part}} + \sigma^2_{\text{page}} + \sigma^2_{\text{song}}}{\sigma^2_{\text{part}} + \sigma^2_{\text{page}} + \sigma^2_{\text{song}} + \sigma^2}. \tag{6.13}$$

This indicates whether the responses within a page are homogeneous or not. In this case, $\text{ICC}_{\text{page}} = 60.9\%$. Scores from a single participant are therefore modestly correlated, while scores from a participant on a single page are pretty highly correlated.
While the measures presented above give a sense of the distribution of the random variance within the model, the goodness of fit of the model can be measured by the squared correlation coefficient R^2 of the model. While this is clearly defined for linear regression with only fixed effects, it is less obvious for linear mixed models. Here, the proposition of Nakagawa and Schielzeth is adopted [90], allowing to take into account the variance explained by the fixed and random effects. For the presented statistical model, the marginal R^2 value for LMMs, giving the ratio of explained variance by fixed effects only is $R^2_{\text{GLMM(m)}} = 0.53$. The conditional R^2 value for LMMs, giving the total ratio of explained variance (random plus fixed effects) is $R^2_{\text{GLMM(c)}} = 0.61$.

6.3.5 Residual diagnostics

One assumption that was made to define the model presented above is the normality of the residuals. Figure 6.2 shows the normal Q-Q plot of the residuals from model 3. Even though all data points at 0 and 100 were excluded, the plot shows evidence for a fat-tailed (or heavy-tailed) distribution of the residuals (especially for lower residuals). A preliminary fit of the model including the excluded values showed even more deviation from the normal distribution. For the sake of generalisation and validity of the test statistics, those values were excluded in the reported analysis. The results of test statistics for the random and fixed effects may therefore be taken with a grain of salt here. In a possible future analysis of the data, this may be taken into account by employing statistical methods that account for such behavior, but no R implementation of a robust method (*i.e.* accounting for heavy-tailed distributions) allowing for nested models with fixed and random effects could be found.

The variance across levels of the covariates is also assumed to be constant. This seems to be a reasonable assumption for this data as depicted by figures 6.3 through 6.7. There is no evidence for systematic variation of the distribution across levels of a given input variable.
Finally, figure 6.8 shows the residuals from Model 3 plotted against the fitted values. While the diagonal lines are simply due to participants grouping their responses (in this case towards multiples of 10) as explained by Searle [114], the apparent correlation between residuals and fitted value indicates unexplained effects in the model. This is confirmed by figure 6.9 where the correlation is once again apparent between the residuals and the subjective data. This is however expected with the explained variance attaining 61%. A general sense for the fit by

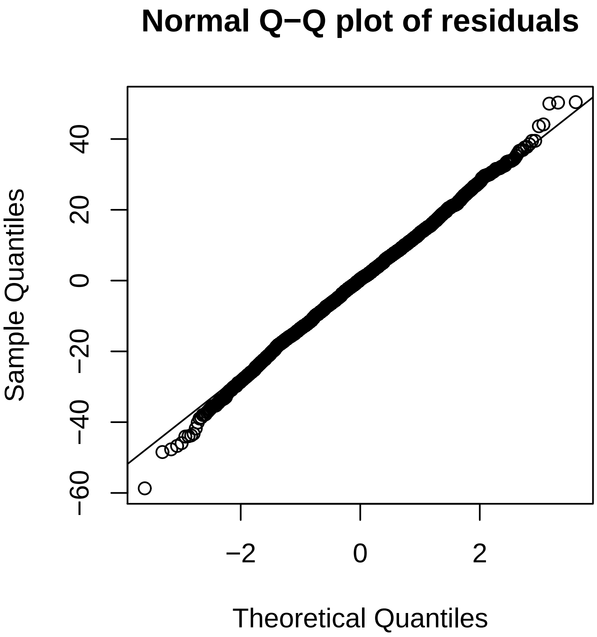


Figure 6.2 – Normal Q-Q plot of the residuals from model 3.

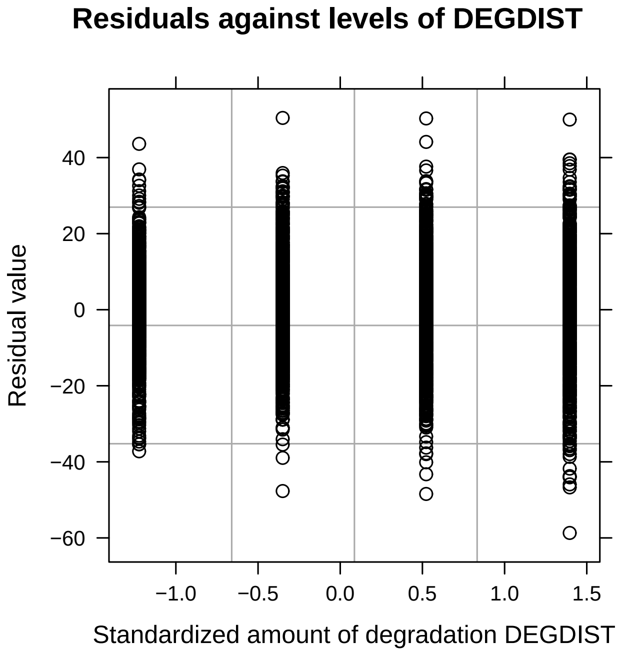


Figure 6.3 – Residuals of model 3 plotted against levels of degradation covariate DEGDIST.

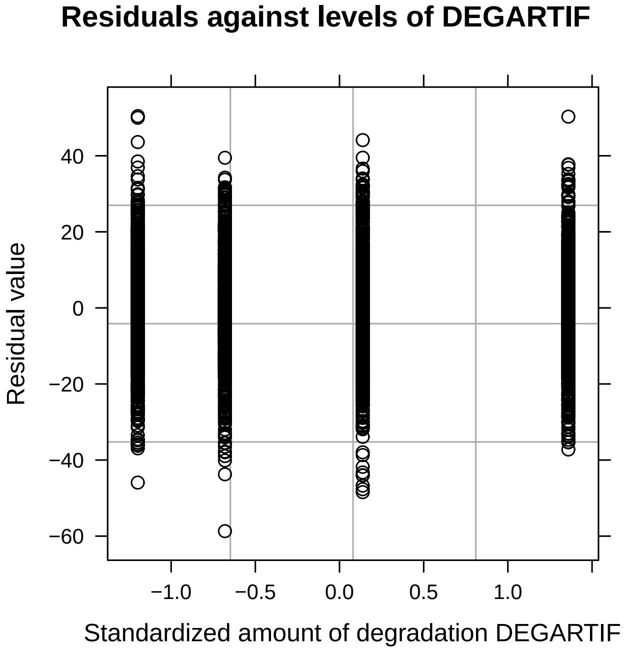


Figure 6.4 – Residuals of model 3 plotted against levels of degradation covariate DEGARTIF.

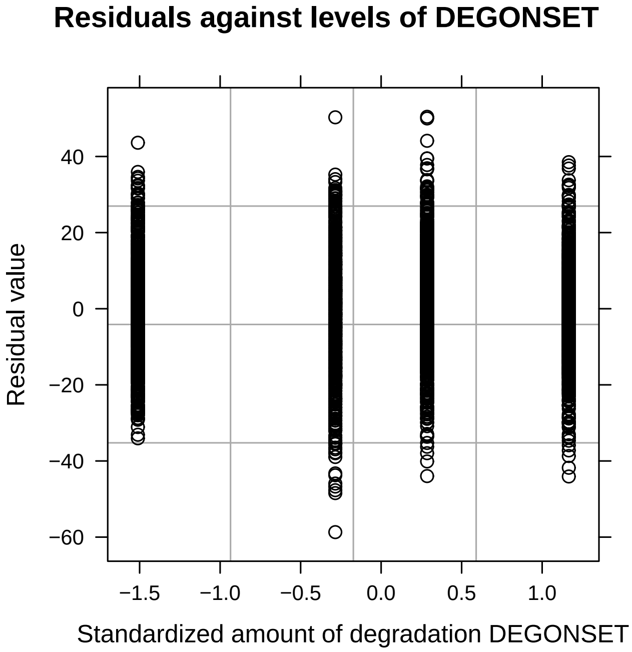


Figure 6.5 – Residuals of model 3 plotted against levels of degradation covariate DEGONSET.

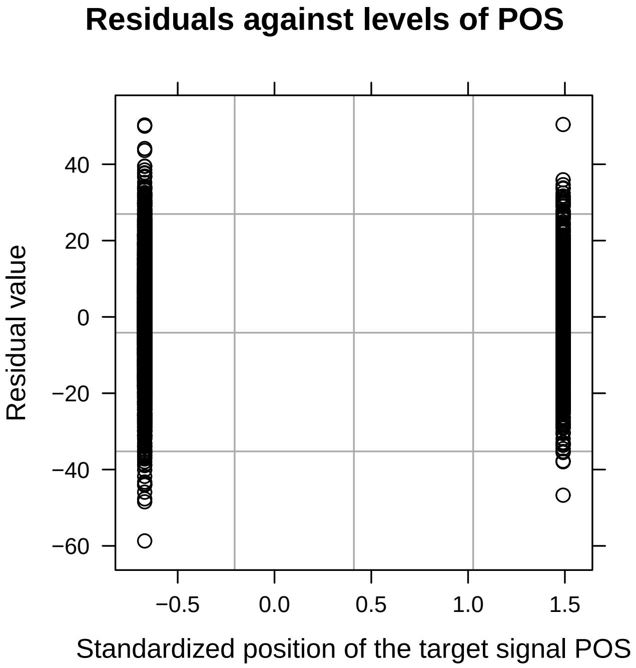


Figure 6.6 – Residuals of model 3 plotted against levels of the spatial configuration covariate POS.

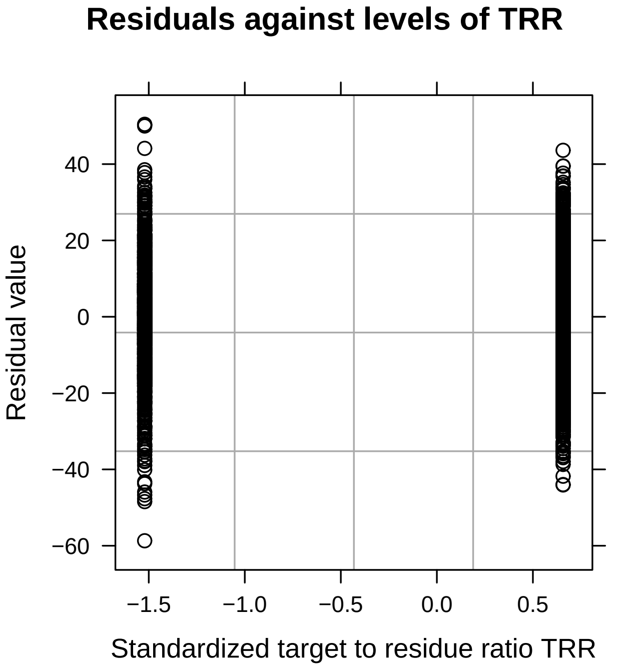


Figure 6.7 – Residuals of model 3 plotted against levels of the target to residue ratio covariate TRR.

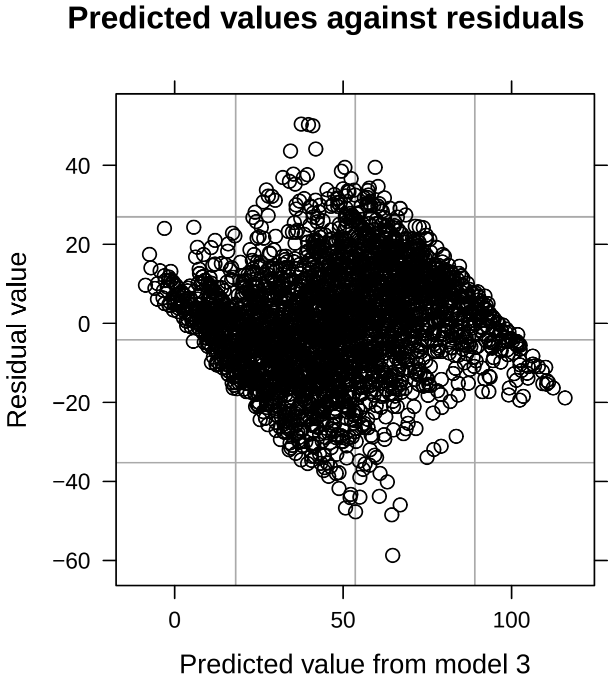


Figure 6.8 – Residuals from model 3 plotted against the predicted values.

model 3 can be gained from figure 6.10, illustrating the correlation between the predicted data by model 3 and the actual responses from participants.

6.4 Discussion

6.4.1 Detailed discussion

The fixed-effect parameter estimates reported in table 6.4 are based on scaled and centered variables. While the scaling and centering may seem counterintuitive in terms of interpretability, this allows for the direct comparison between the effects as pointed out by Schielzeth [112]. The scale of the scaled and centered variables are not in their original units any more but rather in units of their standard deviation. The reported fixed-effect parameters are therefore the estimated variation in perceived quality for an increment of 1 standard deviation in the corresponding variable. Note that the sign of the variation still depends on the units of the original variable.

While the reported intercept is not of much interest, since it does not have much meaning with centered and scaled variables, other fixed effects may be more interesting. The results report an increment of 12.07 in the expected score for an increment of 1 standard deviation in the degradation amount for degradation 1, *i.e.* in the low-pass filter frequency for the target distortion degradation family. The higher the low-pass filter frequency, the higher the expected

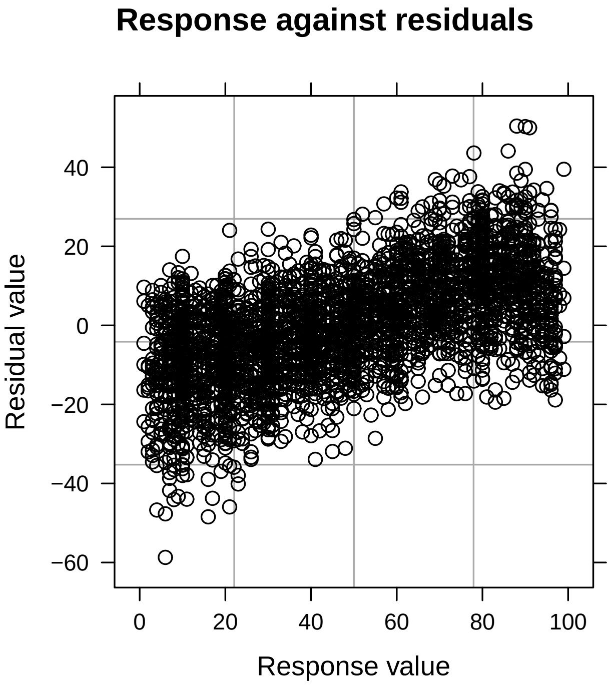


Figure 6.9 – Residuals from model 3 plotted against the data gained from participants.

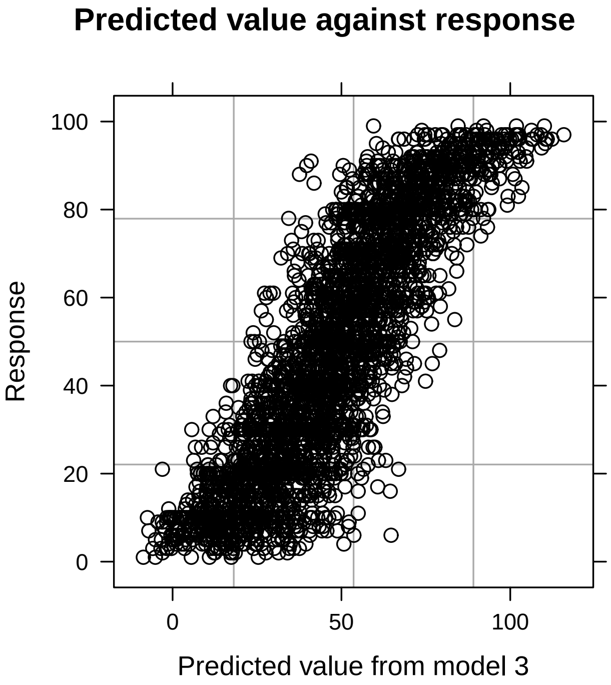


Figure 6.10 – Responses from participants plotted against the predicted values from model 3

response. The squared DEGDIST term is not significant and can therefore be ignored. This corresponds to expected behavior.
For the musical noise-type artifacts degradation, for an increase of 1 standard deviation in the amount argument, *i.e.* in the rate of exchanged time/frequency bins, a decrease of 3.26 of the response is expected. The corresponding quadratic term is significant and adds to the decrease with 1.93 points per standard deviation. Once again, this seems to confirm the perceptual impact of the degradation algorithms, since the decrease is monotonic with respect to the amount argument.
For the onset misallocation degradation, the expected response decreases by 4.74 for an increase of 1 standard deviation in the third amount argument, *i.e.* of the ratio of exchanged onsets. However, the quadratic term for DEGONSET is significant and positive with an increase of 2.06 points expected for an increase of 1 standard deviation. This is quite challenging to interpret. A closer look at the standardised values of the amount argument for the onset misallocation degradation however sheds some light on this situation. Standardised variables generally take values between -3.0 and 3.0 [112]. A plot indicating the combined value of the first-order and quadratic term over that range is therefore given in figure 6.11. Note that the standardised values taken by the DEGONSET input variable are $\{-1.51, -0.28, 0.28, 1.17\}$. Predictions below the lowest value of -1.51 do not make sense, since this value corresponds to a $a_{\text{onset}} = 0$. The four values taken by the input variable are highlighted by the markers. The parabolic curve shows that the values taken by DEGONSET in this experiment all fall on its descending side and that the quadratic term therefore does not change the sign of the derivative of the overall effect, but rather modifies its magnitude as a function of the abscissa. This translates to an expected negative variation of the response as a function of DEGONSET for the values that were used in this experiment. Generalisation towards higher values might however be difficult, but the proposed values already cover a wide range of quality degradation in the stimuli. Note that the curve shown on figure 6.11 does not take into account interactions, which may be interpreted separately, since the input variables were standardised [112]. Overall, the onset misallocation degradation seems to behave as expected. Since the degradation amount arguments were chosen such as to encompass realistic ranges of degradations encountered in real life, the collected data seems to show evidence for the predominance of target distortion and onset misallocation over musical noise-type artifacts in terms of perceived quality degradation.

A quite interesting finding is the effect of the position of the target. Playing back the target away from the residue seems to have a global impact on perceived quality. This might be due to the listeners expecting to have the voice spatialised at the same location than the music. Having them in different locations is a quite unusual situation even for the experienced listener and despite the experimental task specifying that the quality had to be rated with respect to a reference that presented the same spatial configuration, the position of the target seems to have an impact on the perceived quality. A similar reasoning can be made for the TRR, where targets which are significantly louder than the residue seem to have a negative impact on perceived quality.

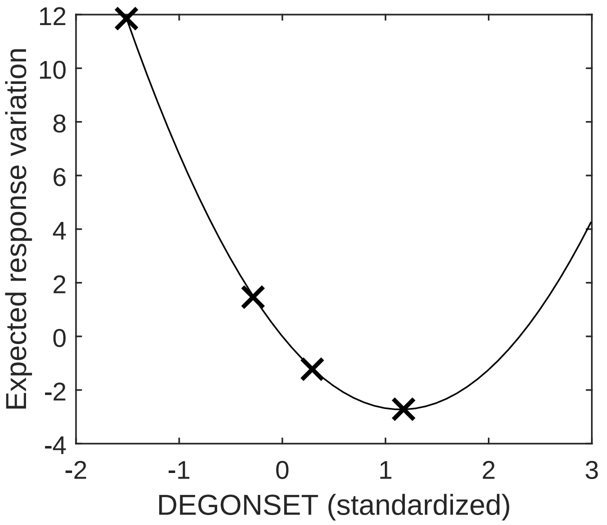


Figure 6.11 – Expected variation of the response due to linear and quadratic effects of DEGONSET. Actual values taken by the input variable are highlighted by crossed markers.

Most of the interaction effects were evaluated as being nonsignificant. The two interactions between DEGARTIF and POS and between DEGONSET and POS are reported as significant with $p < .05$ in table 6.4. However, the degrees of freedom are estimated by an approximation without a clear consensus in the literature as to how they should be estimated and the distribution of the residuals seems to be slightly heavy-tailed. An interpretation of these two interactions seems therefore inappropriate.

A highly significant interaction is however reported between DEGDIST and DEGONSET with an estimated negative value of -3.9. Having only one degradation type between target distortion or onset misallocation therefore results in a penalty in the estimated expected response while having large amounts of both or none at all results in a positive term added to the estimated expected response. Note however that the main effect of DEGDIST is dominating as seen on figure 6.12, where the combined effects of the main, the quadratic and the interaction terms is plotted against the values of DEGDIST and DEGONSET used in this experiment. A highly significant positive interaction is also reported between TRR and POS. Since only 2 values are used in TRR and POS, this can be interpreted as follows: If the TRR is 6 dB and the target is spatialised in front of the listener, the response is slightly better on average than expected only from the main effects. The same is true when the TRR is 12 dB and the target is spatialised to the right of the listener. For the other two combinations, the response is slightly worse than expected. Note that this interaction is also dominated by the two main effects and therefore only slightly alters the expected response.

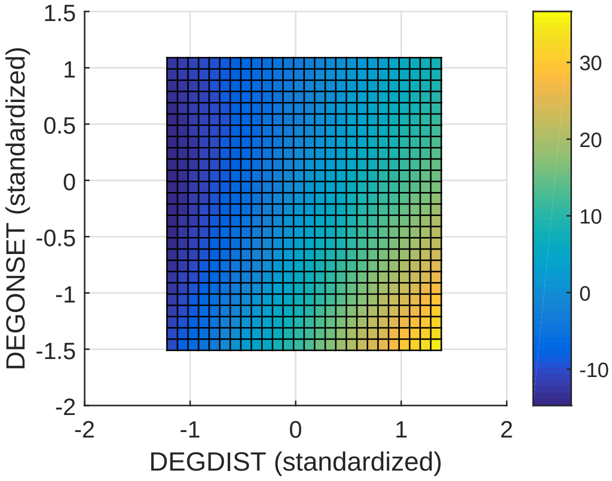


Figure 6.12 – Expected variation of the response due to the main, quadratic and interaction effects of DEGDIST and DEGONSET over the span of values covered in this experiment.

6.4.2 Summary

Overall, the analysis indicates that the results behave as expected. The main effects of the degradations indicate a monotonic decrease of the perceived quality as a function of the amount arguments of the degradation synthesis algorithms. Moreover, since the input variables were standardised, the magnitudes of the main and quadratic terms of DEGDIST, DEGARTIF and DEGONSET seem to indicate that target distortion is the most important degradation type with regards to perceived quality, followed by onset misallocation and then musical noise-type artifacts.

6.5 Objective scores for this data set

6.5.1 PEASS-based approach with decomposition

The state-of-the-art Perceptive Evaluation Methods for Audio Source Separation (PEASS) model as well as the extended model that was proposed in chapters 1 and 3 was used to predict the scores of this experiment. The binaural signals that were used for the subjective experiment were used as inputs for the model, *i.e.* both the degraded target and the reference target and interfering sources were binaural signals. The model therefore operated on two channels with a degraded target as signal under test and the clean target as reference. The residue was specified as potentially interfering source.

The internal parameters of the model were kept at their optimised defaults (see section 3.3) with the filter length for the decomposition at 500 ms. The mapping however was optimised for the new data set, selecting the optimal number of neurons for the mapping and the optimal subset of features to be used. The Artificial Neural Network (ANN) training was kept similar to what was presented in section 3.3 with an 85-fold cross-validation (17 subjects × 5 mixtures).

Multiple references were accounted for by using the corresponding Mean Objective Score (MOS). Note however that this MOS did not always have the same number of measurements, because of the D-optimal experimental design matrix. For that same reason, the training features were not based on the MOS but individual scores were used.

For the state-of-the-art model, a correlation of 0.53 was achieved between the subjective and the objective scores when using $n_{\text{neur}} = 3$ hidden neurons and feature vector $\mathbf{q}_j = [q_j^{\text{overall}}, q_j^{\text{target}}, q_j^{\text{artif}}]$. This might seem low when comparing to the PEASS database. However, since the statistical model was able to explain only 53% of the variance by fixed effects only (*i.e.* knowing all degradation levels, TRRs and spatial configurations), this seems rather good. Note that the R^2 measure reported for linear models corresponds to the coefficient of determination which is the square of the correlation coefficient. A coefficient of determination of 53% would therefore correspond to a hypothetical correlation of 0.73. This can't be applied directly to linear mixed models, but it gives an idea of the performance of the objective model for the given situation. In this particular case, with all (significant and nonsignificant) parameters included, model 3 gives a correlation coefficient of 0.87 between the predicted and measured responses. Note however that this is achieved by excluding all references, all responses with value 100 and all responses with value 0 due to constraints explained above. The objective models were trained including all stimuli, but using MOS for the hidden references and are therefore expected to be less accurate (but more generalizable).

The extended version of PEASS achieved a correlation coefficient of 0.57, outperforming the state of the art by 0.05 points, which is small but not negligible. The number of hidden neurons was $n_{\text{neur}} = 7$ and the optimal feature vector was found to be $\mathbf{q}_j = [q_j^{\text{overall}}, q_j^{\text{artif}}, q_j^{\text{onset}}]$. The modelling of the additional onset component therefore seems to improve the accuracy of the model for 3D audio, whereas it reported equivalent performance for the case without spatialisation.

While these scores might seem quite low performance, one might refer to the first version of PEASS that was published by Emiya *et al.* [42] where the performance of state-of-the-art models were 0.61 at most. Since the proposed models are first extensions of existing ASS evaluation models (or auditory models in general) into the context of ASS evaluation with spatial audio, the performance measured here seems to be rather satisfactory. Additionally, the reader is reminded that the basic version of PEASS and the extended version that were used here were also tested against the PEASS subjective scores and a correlation coefficient of 0.90 was found. The decreased performance in the presented study may therefore be attributed more to the use of MOS for the test against the PEASS subjective database vs. individual scores here and to the more complex experimental task since spatial components are present.

6.5.2 PEMO-Q-based approach without decomposition

The PEMO-Q-based approach proposed in chapter 5 was also tested against the subjective scores from the binaural experiment. As for the PEASS model, binaural signals were used as

inputs. Here however, the sum of the clean target and residue served as reference and the sum of the degraded target and the degraded residue was used as signal under test.
The internal parameters of the PEMO-Q model were kept at their optimised defaults. Since the pre-study seemed to indicate a linear relationship between the predicted scores and the subjective scores, the mapping was based on a linear regression. While no cross-validation setting is used in this case, the weights for the high- and low-frequency features as well as an intercept are estimated by linear regression, using the subjective scores as responses and the objective scores as predictors.

The PEMO-Q-based model achieved a correlation coefficient of 0.60 for the data of the binaural study, which is slightly more accurate than the extended PEASS model. The computation was however much faster as the regression analysis is much faster than the ANN training.

6.6 Conclusion

A subjective experiment about ASS evaluation in binaural listening conditions was conducted and results presented in this chapter. While spatialisation was reduced to a minimum by using ITD synthesis only, the study gives first insights into subjective ASS evaluation in the context of spatial audio. The results of the study suggest that the degradation synthesis algorithms presented in chapter 5 work as intended. Additionally, the fitted statistical model (even though the fit is not perfect) suggests that target distortion is the most harmful degradation in terms of perceived quality, followed by onset misallocation and then musical noise-type artifacts.

The subjective scores collected in the presented study were used to test the accuracy of the two modelling approaches presented in chapters 3 (basic and extended PEASS model) and 5. The basic PEASS model was found to have an accuracy of 0.53, while the extended version achieved an accuracy of 0.57. The PEMO-Q-based binaural model achieved a correlation of 0.60. These results indicate that the extended version of PEASS might perform better in a spatial context, while the PEMO-Q-based binaural model seems to outperform both versions at a reduced computational cost (greatly reduced for training, slightly reduced for prediction). These results are judged satisfactory for a first extension of ASS evaluation models into the spatial audio context.

The next chapter presents a very similar study conducted in WFS rendering conditions. While only ITD synthesis has been used in the presented study, WFS rendering should provide a full set of binaural cues to the participants and therefore be more representative of a “natural” binaural hearing situation. The results should provide insights about the generalisation of the conclusions drawn here, relativising them where necessary. Additionally, the differences between the binaural model and the basic and extended PEASS version should be accentuated, because of the binaural modelling in the former.

7 ASS evaluation in a WFS rendering context

7.1 Introduction

The same subjective study that was presented in a binaural context in chapter 6 was also conducted in a Wave Field Synthesis (WFS) context. In that context, there are no approximations made about spatial perception, as opposed to the binaural study, where the only spatial cue available to the listener is the Interaural Time Difference (ITD). However, the WFS system may introduce some spatial degradations, but they are expected to be kept at a minimum, since the spatialisation method presented in section 4.4 is used. This choice was made even though only horizontal spatialisation is used, since the same system would likely be used for further "full 3D" studies.

This chapter presents the study in the WFS context and its results are discussed and compared to the results of the binaural study. For the sake of clarity and to spare the reader from having to switch back and forth between chapter 6 and this chapter, the design and results sections are fully reported here, even though certain parts are repeated word for word.
Results of the two proposed modelling approaches using the new subjective data are also reported, together with a short discussion of the results.

7.2 Experimental design

For this third experiment, the same 7 factors as for the binaural study were considered: Target to Residue Ratio (TRR), spatial configuration, musical excerpt, onset misallocation level, target distortion level, musical noise-type artifacts level and last but not least, the number of the participant. The same experimental design approach was used as for the binaural experiment. To reduce the number of possible combinations, only 2 TRR levels, 2 spatial configurations, 5 musical excerpts and 4 degradation levels of each type were considered. Even with these simplifications though, the number of combinations per participant is prohibitive if a full factorial design were considered ($2 \cdot 2 \cdot 5 \cdot 4^3 = 1280$ stimuli, not counting references and anchors). A D-optimal design was therefore computed to reduce the number of stimuli per participant while maximising the determinant of the information matrix of the experiment (see Eriksson *et al.* [43] *e.g.* for an introduction). The design was computed with the MATLAB® software and its coordinate exchange algorithm contained in the `cordexch` function. Note that the design matrix used in this third WFS experiment is different from that in the binaural experiment. Constraints were a linear model with interactions, 20 participants and 3200 runs in total, aiming for 160 stimuli per participant. The levels of all factors were predetermined. Since the experimental protocol was a modified Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) protocol, the rows of the resulting design matrix (runs) were grouped by experimental condition regarding TRR, spatial configuration, musical excerpt and participant. Every group was then segmented into pages of 6 runs or less.

Since a D-optimal design does not use the same runs for every participant and therefore not necessarily the same number of runs for every participant, there were notable differences in number of runs across participants. In order to exclude an effect due to different fatigue sates across participants, the number of pages and runs was equalised across participants. Constraining a combinatory algorithm to use pages from the precedent participant to complete the design matrix for every participant (in a cyclic fashion, *i.e.* using pages from the last participant for completing the first one), the smallest achievable number of pages was determined to be 42. The number of runs per participant was 191 in that case. A reference and an anchor was then added to every page, adding 84 runs for every participant summing up to a total of 275 runs per participant.

A total of 21 participants (1 female, 20 males) between ages 18 and 31 (M = 23.6, SD = 3.4) took part in this study. Mainly students, doctoral students and post-doc researchers from EPFL, they were not experts in audio quality evaluation, but all had experience with audio and critical listening in general. All participants reported normal hearing (no known impairments) although no audiological measurements were made. All participants were remunerated for their participation in this experiment.

7.2.1 Experimental task

Like the pre-study and the binaural study, this WFS study was based on a modified MUSHRA protocol. A training phase preceded the evaluation phase as for the other experiments. To shorten the training phase, only the reference and the most degraded stimuli for each test page were presented during the first step of the training phase, where participants could listen to the stimuli to get a sense of the involved scale of quality. Participants could then get accustomed to the evaluation interface with one evaluation example (*i.e.* one page).

The stimuli selected for each participant as a result of the D-optimal design were grouped such that a single evaluation test page showed only stimuli of the same song, TRR and spatial configuration. As for the binaural study, the evaluation test pages contained a low anchor which was synthesised as described in the next section. Participants were asked to rate the quality of all sound excerpts on a page (including a low anchor and a hidden reference) as compared to a labelled reference using a graphical interface on a computer. The sliders used to rate the quality of each stimuli only displayed “best” at the top of the scale and “overall worst” at the bottom to mitigate the construction of personal subscales by the participants. The graphical user interface was the same as for the binaural study (a screenshot is presented in figure 6.1). Participants could listen to all stimuli and the reference as often as they wanted. On average, participants took about 1 hour to complete the experiment.

In order to try to keep the perceptual scale as relevant as possible, the training page containing references and low anchors of all pages was shown to the listener after every 5 pages during the evaluation phase. Since a large number of stimuli were used for every participant, this should limit the influence of participants changing the low end of the subjective scale due to forgetting what the worst overall stimulus sounds like.

7.2.2 Stimuli

The musical excerpts used in this third experiment were the same as for the binaural study. Details about the excerpts are given in table 6.1. The 5 excerpts were equalised in loudness before the degradation synthesis, in order to minimize the risk of degradations being more easily detectable for certain excerpts due to increased loudness.
The amount arguments for the degradation synthesis algorithms were chosen for degradations to vary between inaudible and clearly audible for all 3 degradation families. Their values are given by equations (6.1)-(6.3).
The low anchor presented on each test page is defined by the same spatial configuration, musical excerpt and TRR as the stimuli under test, but contains all degradation types at their maximal level (*i.e.* the maximal level used in the stimuli). The degradation synthesis algorithms were applied sequentially, beginning with the target distortion, continuing with musical noise-type artifacts and finishing with onset misallocation. This order was chosen based on trials with different orders. If the musical noise artifacts or the onset misallocation degradations were synthesised first, the target distortion might filter them out again. Also, if the musical noise artifacts synthesis algorithm came after the onset misallocation synthesis, the time-frequency coefficient selection algorithm would be biased by the onsets.

This experiment was also conducted to test the degradation audibility modelling approach that is presented in section 5.4.2. For that reason, a very precise control over the spatialisation parameters was needed. This was achieved through the placement of virtual sources in the WFS system. Two spatial configurations were used: In the first configuration, both the residue and the target were placed in a standard stereo configuration (virtual loudspeakers at $-30°$ and $30°$ azimuth). In the second configuration, the target was moved to be centered at $270°$ azimuth ($90°$ to the right of the listener), virtual loudspeakers therefore being located at $300°$ and $240°$ azimuth. All virtual sources were positioned in the horizontal plane at 5.6 m distance from the listener.
TRR levels were chosen to be 6 dB and 12 dB.

7.2.3 Experimental setup

A WFS setup was installed in the same room as for the vertical localisation study in 3D WFS conducted in section 4.5. The room was a listening room of 6.70 x 6.80 x 2.60 m. The mean reverberation time of the room was measured to be about 0.25 s and flat below 5.3 kHz and decaying for higher frequencies to reach 0.18 s at 16 kHz, which is similar to studio conditions. The background noise level of the room was measured to be approximately 23 dB(A) (1 second integration period, averaged over 2x 10 minutes of measurement). 3 of the walls are coated with absorbing materials (mineral wool covered with tissue), the floor is entirely covered with carpet and the ceiling is acoustically treated. The fourth wall contains windows. Even though no measurements for early reflections were made, the configuration of the room and the covering of the windows with heavy curtains should minimize the influence of the room.

The WFS rendering system was however different for this study and composed of 60 ELAC 301.2 loudspeakers and 4 Velodyne SPL 800i subwoofers, which were distributed as illustrated on figure 7.1: two outer horizontal rings of 27 and 20 loudspeakers at heights 0 m and 1.20 m respectively relative to the position of a listener's head (blue and green rows) and a ceiling over which the remaining 13 loudspeakers were distributed (yellow rows). The subwoofers were placed in the four bottom corners of the system. The loudspeaker setup therefore covered the upper hemisphere with respect to a listener's head.
The 3D WFS algorithm was implemented on a Sonic Wave 1 3D sound processor[1], which delivered the loudspeaker driving signals for the ELAC loudspeakers to 8 sonic emotion M3S amplifiers through a RME ADI-648 MADI to ADAT converter. Since the subwoofers contain their own amplifiers, the driving signal from the WFS system was delivered to them through RME ADI-2 AD/DA converters. All software components, commands and stimuli were generated with MATLAB® on a PC connected to a DirectOut EXBOX.UMA soundcard.
The set of possible virtual source locations was given by a hemisphere with a radius of 5.4 m and the portion of the horizontal place for greater radii (see figure 7.2).

To enable the comparison of the results between different sessions and participants, the output level was set to +4.5 dB in the WFS software. Since the soundcard and the amplifiers both have fixed gains, this resulted in stimuli being played back at about 65 dB(A) which seemed comfortable for prolonged listening sessions. For subjects who required it, the playback level could be slightly adjusted, as recommended by International Telecommunication Union (ITU) recommendation ITU-R BS.1116 [63]. This approach is also taken in ITU-R BS.1534 [64] where the MUSHRA protocol is defined.

[1] http://www.sonicemotion.com/professional

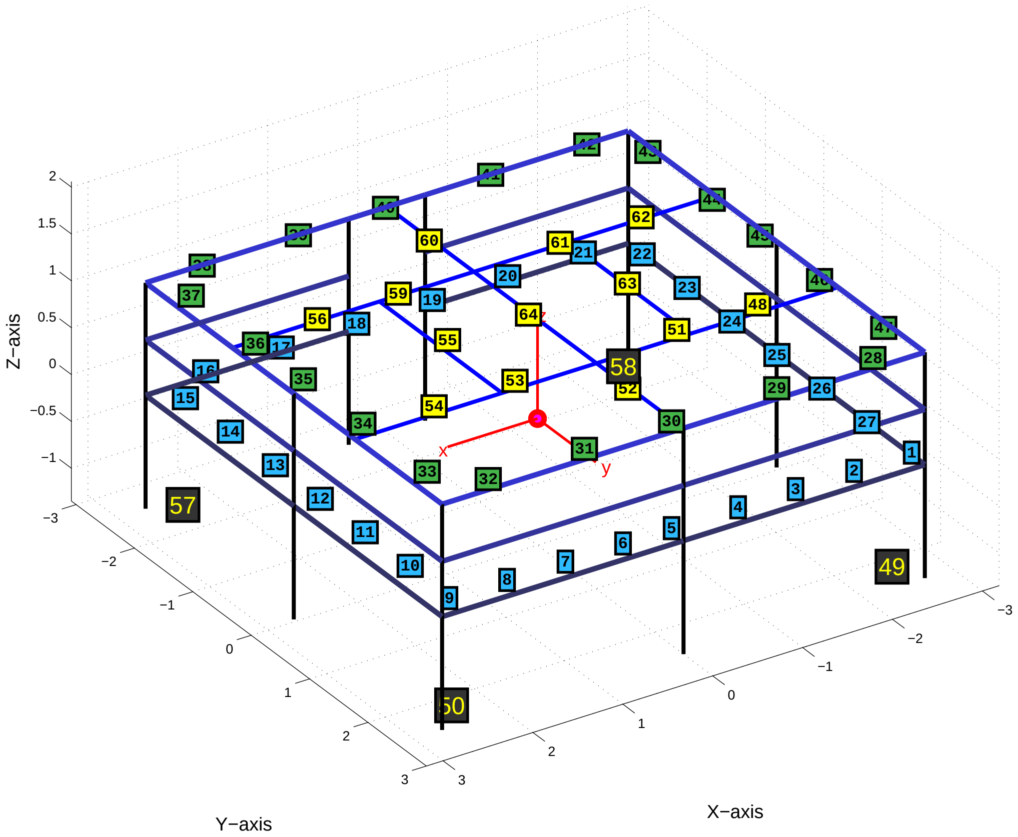


Figure 7.1 – 64-channel loudspeaker setup for the psychoacoustic study with WFS spatialisation. Blue squares correspond to loudspeakers at ear-level of a sitting listener at the listening position in pink. The green squares correspond to loudspeakers on the outer edge of the structure at 1.20 m above ear level. Yellow squares correspond to loudspeakers distributed over the ceiling (also about 1.20 m above ear level). Dark grey squares correspond to subwoofers on ground level.

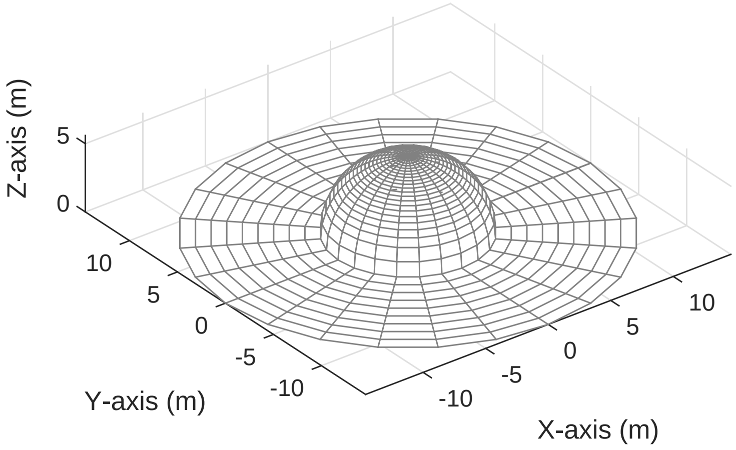


Figure 7.2 – The set of possible source locations for the implemented WFS system. The center hemisphere has a radius of 5.4 m.

7.3 Results

7.3.1 Post-screening

Just like in the binaural experiment, the scores on the hidden reference stimuli included in the rating pages were used to detect participants who did not properly understand the task. Since the task was to rate the quality of test sounds as compared to a given reference on each page, it is expected that the hidden reference that did not undergo any degradation processing yields perfect quality (perfect score) when comparing to the reference.

A slightly different approach that is judged more consistent than that for the binaural study was used for this WFS experiment. For all scores given to hidden reference signals, the Euclidian distance D to the perfect score of all 100 was computed for every participant. The mean distance across participants is $D_{\mu} = 95.6$. A critical distance $D_c = D_{\mu} + \sigma_D$ was then defined where σ_D is the standard deviation of D. Results of participants whose distances were above the critical distance were discarded. Based on this criterion, the results of 4 participants had to be discarded. Discarded participants had distances of $D_{13} = 156.7$, $D_{17} = 182.3$, $D_{20} = 206.1$ and $D_{21} = 196.3$. The mean distance amongst remaining participants is $D = 74.5$.

As previously, the low anchor scores were not used in the post-screening. Even though they should be the signals with the worst scores on any page, it is not expected to have a consensus among participants concerning the range of the score.

7.3.2 Statistical analysis

Statistical model building

In the presented binaural experiment, there were 6 covariates related to the stimuli: the Target to Residue Ratio “TRR” (6 or 12 dB), the spatial position “POS” (0° or 270°), the musical excerpt “SONG” (1 out of 5 possible), the amount of target distortion degradation (*i.e.* the low-pass cut-off frequency) “DEGDIST”, the amount of musical noise-type artifacts “DEGARTIF” and the amount of onset misallocation degradation “DEGONSET”. The participant number “PART” is an another covariate. Moreover, the stimuli were presented grouped by pages, which introduces the last covariate, the page number “PAGE”.

Before the analysis, the DEGDIST, DEGARTIF, DEGONSET, TRR and POS covariates were scaled and centered. This allows for easier convergence of the optimisation algorithms, since the covariates are on very different scales. Additionally, the 3 degradation amount arguments were transposed to a logarithmic scale before scaling and centering to account for the behavior observed during the pre-study (perceived quality degrading inversely proportional to the amount arguments). For degradations including a zero-valued amount argument, 0.1 was added to all amount argument levels before the change of scales to avoid singularity.

In terms of data structure, this results in a three-level clustered data model. Stimuli, which are the unit of analysis, are clustered into pages, which are clustered into participants. Note that the PAGE factor is nested within the PART factor. The degradation covariates DEGDIST, DEGARTIF and DEGONSET are level 1 covariates, varying within pages. The SONG, TRR and POS covariates are level 2 covariates, varying across pages. This structure is the same as for the binaural study and summarised in table 6.2.

Once again, since the data that is to be fitted here is rather complex, a step-up modelling approach is taken, following the guidelines given by West *et al.* [140]. All models are fit using the R software, using the `lme4` package (see book draft by Bates [6] for a complete reference). The `lmerTest` package was used to obtain test statistics for fixed covariates.
The analysis steps can be summarised as follows:

1. Fit the initial “unconditional” (variance components) model (Model 1)
2. Build the level 1 model by adding level 1 covariates (Model 2)
3. Build the level 2 model by adding level 2 covariates (Model 3)

Variants of the three models can be used to test specific hypotheses about included effects. A general model for the data, including interactions, square terms and random coefficients can be written as

$$\begin{aligned}
&\text{RESPONSE}_{ijkl} = \\
&\beta_0 + \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\
&+ \beta_4 \times \text{TRR}_{jkl} + \beta_5 \times \text{POS}_{jkl} + \\
&+ \beta_6 \times \text{DEGDIST}_{ijkl} \times \text{DEGARTIF}_{ijkl} + \beta_7 \times \text{DEGDIST}_{ijkl} \times \text{DEGONSET}_{ijkl} + \\
&+ \beta_8 \times \text{DEGARTIF}_{ijkl} \times \text{DEGONSET}_{ijkl} + \beta_9 \times \text{DEGDIST}_{ijkl} \times \text{TRR}_{jkl} + \\
&+ \beta_{10} \times \text{DEGARTIF}_{ijkl} \times \text{TRR}_{jkl} + \beta_{11} \times \text{DEGONSET}_{ijkl} \times \text{TRR}_{jkl} + \\
&+ \beta_{12} \times \text{DEGDIST}_{ijkl} \times \text{POS}_{jkl} + \beta_{13} \times \text{DEGARTIF}_{ijkl} \times \text{POS}_{jkl} + \\
&+ \beta_{14} \times \text{DEGONSET}_{ijkl} \times \text{POS}_{jkl} + \beta_{15} \times \text{TRR}_{jkl} \times \text{POS}_{jkl} + \\
&+ \beta_{16} \times \text{DEGDIST}^2_{ijkl} + \beta_{17} \times \text{DEGARTIF}^2_{ijkl} + \beta_{18} \times \text{DEGONSET}^2_{ijkl} + \\
&+ u_{0l} + u_{1l} \times \text{DEGDIST}_{ijkl} + u_{2l} \times \text{DEGARTIF}_{ijkl} + u_{3l} \times \text{DEGONSET}_{ijkl} + \\
&+ u_{4l} \times \text{TRR}_{jkl} + u_{5l} \times \text{POS}_{jkl} + \\
&+ u_{0k|l} + u_{1k|l} \times \text{DEGDIST}_{ijkl} + u_{2k|l} \times \text{DEGARTIF}_{ijkl} + u_{3k|l} \times \text{DEGONSET}_{ijkl} + \\
&\qquad\qquad + u_{j|l} + \epsilon_{ijkl}. \qquad (7.1)
\end{aligned}$$

In this model specification, RESPONSE_{ijkl} represents the score assigned to stimulus i on page j for song k by participant l and β_0 - β_{15} represent the fixed intercept and the fixed effects associated with the covariates (DEGDIST, DEGARTIF, DEGONSET, TRR, POS and their interactions). u_{0l} represents the random intercept associated with participant l. u_{1l} - u_{5l} represent the random slopes associated with the different covariates DEGDIST, DEGARTIF, DEGONSET, TRR and POS for participant l. $u_{0k|l}$ represents the random effect with the random intercept for song k within participant l and $u_{1k|l}$ - $u_{3k|l}$ represent the random slopes associated with the covariates DEGDIST, DEGARTIF and DEGONSET for song k within participant l. Finally, $u_{j|l}$ represents the random intercept for page j within participant l and ϵ_{ijkl} represents the residual.

The distribution of the random effects associated with the participants in this model is supposed to be multivariate normal and can be written as

$$u_l = \begin{pmatrix} u_{0l} \\ u_{1l} \\ u_{2l} \\ u_{3l} \\ u_{4l} \\ u_{5l} \end{pmatrix} \sim \mathcal{N}(0, \mathbf{D}_{\text{participant}}), \qquad (7.2)$$

where $\mathbf{D}_{\text{participant}}$ is the matrix of variances and covariances between the different random effects with variances on the diagonal and covariances off the diagonal.
In a similar fashion, the distribution of the random effects associated with the songs given a

participant is also supposed to be multivariate normal and can be written as

$$u_k|l = \begin{pmatrix} u_{0k|l} \\ u_{1k|l} \\ u_{2k|l} \\ u_{3k|l} \end{pmatrix} \sim \mathcal{N}(0, \mathbf{D}_{\text{song|participant}}), \tag{7.3}$$

where $\mathbf{D}_{\text{song|participant}}$ is the matrix of variances and covariances between the different random effects.
Since there is only a random intercept associated with pages given a participant, a normal distribution is assumed:

$$u_{j|l} \sim \mathcal{N}(0, \sigma^2_{\text{page|participant}}), \tag{7.4}$$

Finally, the residual is assumed to follow a normal distribution with variance σ^2:

$$\epsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2). \tag{7.5}$$

The summary of the 3 models that are considered in this analysis is given by table 6.3.
To avoid a bias of the statistical model towards higher scores because of all the scores corresponding to hidden references, they were excluded from the data for the statistical analysis. Moreover, since the Quantile-Quantile (Q-Q) plot of the residuals showed large evidence for a heavy-tailed distribution in a first fit of the model, all scores equal to 0 and 100 were excluded from the reported statistical models. For a discussion, see section 7.3.5.

Model 1

This first model is a fixed intercept model with random effects associated with the intercepts for pages (level 2) and participants (level 3). Since the musical excerpt is also modelled as a random effect, random intercepts for songs (level 2) are also included in this first model. This can be written as

$$\text{RESPONSE}_{ijkl} = b_{0ijkl} + b_{0jk|l} + b_{0l} + \beta_0 + u_{0l} + u_{0k|l} + u_{0j|l} + \epsilon_{ijkl}, \tag{7.6}$$

where b_{0l} represents unobserved random slopes at the participant level, $b_{0jk|l}$ represents unobserved fixed and random intercepts and slopes at the page level and b_{0ijkl} represents the yet unspecified interactions at the stimulus level.

This model allows for the test of two hypotheses regarding the variance of the page-specific and the song-specific intercepts. This allows to assess if the two random effects associated with the page and the song are useful in the model specification. In the case of the page-specific intercept, the null and alternative hypotheses are:

$$H_0 : \sigma^2_{\text{int:PAGE}} = 0$$
$$H_A : \sigma^2_{\text{int:PAGE}} > 0$$

To test this hypothesis, a nested model 1A is constructed, not including the random intercept for pages. A likelihood ratio test is then conducted, subtracting the -2 Maximum Likelihood (ML) log-likelihood for model 1 from the corresponding value of model 1A based on ML fits. To obtain a p-value for this test statistic, it is referred to a mixture of χ^2 distributions with 0 and 1 degrees of freedom and equal weights 0.5 (following the procedure in West *et al.* [140]). Since the result is significant ($\chi^2(0:1) = 625.66, p < .001$), the alternative hypothesis is confirmed. The random intercept associated with the page is therefore kept for all the following models. In the case of the song-specific intercept, the same procedure is followed, with the null and alternative hypotheses being:

$$H_0 : \sigma^2_{\text{int:song}} = 0$$
$$H_A : \sigma^2_{\text{int:song}} > 0$$

To test this hypothesis, another nested model 1B is fitted, not including the random intercept for songs. The likelihood ratio test is once again significant ($\chi^2(0:1) = 52.53, p < 0.001$) and the random intercept associated with the song is also kept in the following models.

Model 2

For the second model, the level 1 fixed effects (DEGDIST, DEGARTIF, DEGONSET) are added to model 1. This translates to a model written as

$$\begin{aligned}&\text{RESPONSE}_{ijkl} = b_{0jk|l} + b_{0ijkl} + \beta_0 + \\ &+ \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\ &\qquad\qquad + u_{0l} + u_{0k|l} + u_{0j|l} + \epsilon_{ijkl},\end{aligned} \tag{7.7}$$

where $b_{0jk|l}$ represents unobserved fixed and random intercepts at the page level and b_{0ijkl} represents the yet unspecified interactions at the stimulus level. Once again, a likelihood ratio test is conducted based on ML fits of models 1 and 2 to test for the hypothesis that those effects are null. The null and alternative hypotheses therefore are:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta 4 = 0$
H_A : At least one of the variances is not zero.

To test for significance of the alternative hypothesis, the -2 ML log-likelihood of model 2 is subtracted from its counterpart in model. Under the null hypothesis, this test statistic is asymptotically following a χ^2 distribution with 3 degrees of freedom. Since the test is significant ($\chi^2(3) = 1439.2, p < .001$), the alternative hypothesis is preferred and the level 1 covariates are added.

The level 1 covariates are then added as level 3 random effect coefficients, allowing for random intercepts for every degradation level in every participant. This translates to different participants having different sensitivities for the different degradation families. The null and alternative hypotheses to test therefore are:

$H_0 : \sigma^2_{\text{DEGDIST:PART}} = \sigma^2_{\text{DEGARTIF:PART}} = \sigma^2_{\text{DEGONSET:PART}} = 0$
H_A : At least one fixed effect is not equal to zero.

The test statistic follows a χ^2 distribution with 9 degrees of freedom (one for every additional variance / covariance parameter to be estimated when comparing to a simple intercept). Since the likelihood ratio test is significant ($\chi^2(9) = 113.36, p < .001$), the alternative hypothesis is preferred and the random coefficients associated with the degradation levels at the participant level are therefore kept in the model.

As a second alternative to the second model, the level 1 covariates are also added as random coefficients to the SONG random effect. This translates to model 2 being written as

$$\begin{aligned}&\text{RESPONSE}_{ijkl} = b'_{0jk|l} + b'_{0ijkl} + b'_{0l} + \beta_0 + \\ &+ \beta_1 \times \text{DEGDIST}_{ijkl} + \beta_2 \times \text{DEGARTIF}_{ijkl} + \beta_3 \times \text{DEGONSET}_{ijkl} + \\ &+ u_{0l} + u_{1l} \times \text{DEGDIST}_{ijkl} + u_{2l} \times \text{DEGARTIF}_{ijkl} + u_{3l} \times \text{DEGONSET}_{ijkl} + \\ &+ u_{0k|l} + u_{1k|l} \times \text{DEGDIST}_{ijkl} + u_{2k|l} \times \text{DEGARTIF}_{ijkl} + u_{3k|l} \times \text{DEGONSET}_{ijkl} + \\ &\qquad + u_{0j|l} + \epsilon_{ijkl}, \end{aligned} \tag{7.8}$$

where b'_{0l} represents unobserved random slopes at the participant level, $b'_{0jk|l}$ represents unobserved fixed intercepts at the page level and b'_{0ijkl} represents the yet unspecified interactions at the stimulus level. The null and alternative hypotheses thus are:

$H_0 : \sigma^2_{\text{DEGDIST:SONG}} = \sigma^2_{\text{DEGARTIF:SONG}} = \sigma^2_{\text{DEGONSET:SONG}} = 0$
H_A : At least one fixed effect is not zero.

Once again, the statistic for the corresponding likelihood ratio test follows a χ^2 distribution with 9 degrees of freedom. Since the test is significant ($\chi^2(9) = 117.49, p < .001$) the random coefficients are kept. This accounts for different degradation families having different effects for different songs. The degradation covariates are not added to the page random effect, because having different effects for degradation on different pages does not make sense.

Model 3

For the third model, the level 2 fixed effects (POS, TRR) are added as fixed effects. To see if adding these effects is useful in this model, the following null versus alternative hypotheses are tested:

$H_0 : \beta_4 = \beta_5 = 0$
H_A : At least one of the variances is not zero.

The resulting test statistic follows a χ^2 distribution with 2 degrees of freedom and the test is significant ($\chi^2(2) = 636.77 p < .001$), which means that at least one of the effects is not zero.

The level 2 covariates are then added as random coefficients at level 3, allowing for different effects of the TRR and the spatial configuration for every participant. A likelihood ratio test is conducted to see if at least one of the added variances is not equal to zero. The resulting test statistic follows a χ^2 distribution with 11 degrees of freedom. The test is significant ($\chi^2(11) = 21.20, p < .05$), but since it is only slightly significant and for the sake of better comparison with the binaural study presented in chapter 6, the random slopes associated with TRR and target position are not kept in the model, effectively making $u_{4l} = 0$ and $u_{5l} = 0$.

The second-order degradation effects are then added (DEGDIST^2, DEGARTIF^2, DEGONSET^2), since it is expected that the degradation amounts may affect their audibility in a quadratic way, despite the logarithmic mapping of the degradation covariates. A likelihood ratio test for the statistical significance of at least one of the associated intercepts reveals that the second-order terms should be kept in the model ($\chi^2(3) = 124.42, p < .001$).

Finally, since it is expected that interactions between the different degradation families may occur, or that certain degradation may be more harmful in terms of audio quality depending on spatial configuration or TRR, interactions of the fixed effects are added to the model. The null hypothesis of all interactions being zero is rejected ($\chi^2(10) = 346.61, p < .001$). The full model as expressed by equation (7.1) is therefore reached except for the random POS and TRR coefficients at the participant level.

7.3.3 Fixed-effect parameter estimates

The literature references do not agree on how the test statistic of the fixed-effect parameters is distributed. This is why the precedent section used only -2 ML log-likelihood ratio tests. However, approximate tests can be conducted following different assumptions. Here, the `lmerTest` package of the R software was used to compute p-values based on t-statistics using Satterthwaite approximations to degrees of freedom. The results of the corresponding tests, as well as estimates for the fixed-effect parameters are reported in table 7.1.

7.3.4 Covariance parameter estimates and Intraclass Correlation Coefficients

Table 7.2 reports the estimated variance components for the final model. While the estimated variance components of the full model may be of interest for the ICC reported below, the evolution of the variance components throughout the model building process may also be of

Table 7.1 – Estimates of the fixed-effect parameters in model 3. Note that all independent variables reported here had been standardised (scaled and centered) before fitting the model.

Fixed-effect parameter	Estimate	Std. Error	df	t-value	p-value
β_0 (Intercept)	38.85	3.56	6.0	10.92	<.001
β_1 (DEGDIST)	9.51	2.26	6.0	4.21	<.01
β_2 (DEGARTIF)	-5.85	0.93	6.0	-6.30	<.001
β_3 (DEGONSET)	-2.34	0.79	7.0	-2.98	<.05
β_4 (TRR)	-12.53	0.42	675.0	-29.982	<.001
β_5 (POS)	-5.80	0.43	733.0	-13.54	<.001
β_6 (DEGDIST×DEGARTIF)	0.79	0.34	3200.0	2.37	<.05
β_7 (DEGDIST×DEGONSET)	-4.05	0.33	3181.0	-12.42	<.001
β_8 (DEGARTIF×DEGONSET)	0.78	0.33	3215.0	2.33	<.05
β_9 (DEGDIST×TRR)	1.06	0.29	3156.0	3.68	<.001
β_{10} (DEGARTIF×TRR)	-1.12	0.29	3117.0	-3.83	<.001
β_{11} (DEGONSET×TRR)	-0.86	0.29	3147.0	-3.02	<.01
β_{12} (DEGDIST×POS)	-1.23	0.32	3153.0	-3.82	<.001
β_{13} (DEGARTIF×POS)	-1.61	0.33	3066.0	-4.90	<.001
β_{14} (DEGONSET×POS)	-1.10	0.31	3034.0	-3.56	<.001
β_{15} (TRR×POS)	2.50	0.42	726.0	5.90	<.001
β_{16} (DEGDIST^2)	1.12	0.38	3145	2.95	<.01
β_{17} (DEGARTIF^2)	-1.17	0.40	3158	-2.89	<.01
β_{18} (DEGONSET^2)	2.62	0.35	3197	7.57	<.001

Table 7.2 – Estimates of the variance components in model 3.

Random effect parameter	Variance	Std. dev.
$u_{j\|l}$ (Intercept for PAGE)	79.30	8.91
$u_{0k\|l}$ (Intercept for SONG)	50.86	7.13
$u_{1k\|l}$ (DEGDIST slope for SONG)	21.06	4.59
$u_{2k\|l}$ (DEGARTIF slope for SONG)	3.14	1.77
$u_{3k\|l}$ (DEGONSET slope for SONG)	1.79	1.34
u_{0l} (Intercept for PART)	33.67	5.80
u_{1l} (DEGDIST slope for PART)	12.94	3.60
u_{2l} (DEGARTIF slope for PART)	2.12	1.46
u_{3l} (DEGONSET slope for PART)	2.15	1.47

interest. In this case, the addition of the degradation amounts as fixed effects (first step from model 1 to model 2) reduces the estimated residual variance by 40.1% (estimated residual variance of 442.35 in model 1 vs. 264.76 after adding DEGDIST - DEGONSET) whereas the participant-level variance increases by 14.0% and the page-level variance increases by 23.4%. The fixed effects DEGDIST - DEGONSET therefore explain a lot of the variation at the stimulus level, but they don't explain variation among participants or pages / songs which is even highlighted in their presence , *i.e* some variance is attributed to the random effects rather than the added fixed effects.

Introducing random slopes for the degradation amounts at the participant level further reduces the residual variance by 5.6%. The variance increases by 59.4% at the participant level however and decreases by 0.5% at the page level however. Including these random slopes therefore explains some of the random variance at the stimulus level without explaining the random variation between participants or pages / songs. Including the same random slopes for the SONG random effect diminishes the residual variance by another 4.4% but increases the variance at the page level by 6.4% and the variance at the participant level by 1.0%. Random slopes for DEGDIST - DEGONSET therefore do not explain a lot of the random variance in the data at any level.

When adding the covariates POS and TRR at the page level, the variance at that level is diminished by 57.5%. The residual variance is unaffected whereas the participant-level variance is augmented by 1.4%. These two covariates therefore explain a lot of the variance at the page level.

Adding the squared terms for DEGDIST to DEGONSET diminishes the residual variance by 3.9%, and leaves the other levels of variance nearly unchanged. Finally, adding the different interactions between the first-order fixed effects diminishes the residual variance by 11.1% and the participant-level variance by 4.2%, while leaving the page-level variance nearly unchanged.

Overall, the different components added during the model building process affect positively (*i.e.* diminish) the random variance at the expected levels (level 1 covariates affect residual variance, level 2 covariates affect page-level variance).

For the final model, ICCs are used to measure the homogeneity of the responses within a given cluster. For the presented statistical model, two ICCs are of interest: one for the participant level, and one for the page level.

The participant level ICC can be written as

$$\mathrm{ICC}_{\mathrm{part}} = \frac{\sigma^2_{\mathrm{part}}}{\sigma^2_{\mathrm{part}} + \sigma^2_{\mathrm{page}} + \sigma^2_{\mathrm{song}} + \sigma^2}, \tag{7.9}$$

where σ^2_{part} is the sum of the variance components associated with the participant, σ^2_{page} is the variance of the random effect associated with the page nested within participants, σ^2_{song} is the sum of the variance components associated with the song and σ^2 is the residual variance. This ICC is high if the total random variation is dominated by the variance of the random participant effect, *i.e.* if the results from a participant are relatively homogeneous, but they

vary widely from participant to participant. For the model specified above, the value of this ICC is $\mathrm{ICC}_{\mathrm{part}} = 12.3\%$.
The second ICC that can be defined is associated with the page. It is defined as

$$\mathrm{ICC}_{\mathrm{page}} = \frac{\sigma^2_{\mathrm{part}} + \sigma^2_{\mathrm{page}} + \sigma^2_{\mathrm{song}}}{\sigma^2_{\mathrm{part}} + \sigma^2_{\mathrm{page}} + \sigma^2_{\mathrm{song}} + \sigma^2}. \tag{7.10}$$

This indicates whether the responses within a page are homogeneous or not. In this case, $\mathrm{ICC}_{\mathrm{page}} = 49.9\%$. Scores from a single participant are therefore very modestly correlated, while scores from a participant on a single page show better correlation.
While the measures presented above give a sense of the distribution of the random variance within the model, the goodness of fit of the model can be measured by the squared correlation coefficient R^2 of the model. While this is clearly defined for linear regression with only fixed effects, it is less obvious for linear mixed models. Here, the proposition of Nakagawa and Schielzeth is adopted [90], allowing to take into account the variance explained by the fixed and random effects. For the presented statistical model, the marginal R^2 value for Linear Mixed Models (LMMs), giving the ratio of explained variance by fixed effects only is $R^2_{\mathrm{GLMM(m)}} = 0.60$. The conditional R^2 value for LMMs, giving the total ratio of explained variance (random plus fixed effects) is $R^2_{\mathrm{GLMM(c)}} = 0.66$. The variance is therefore explained a bit better than in the binaural study.

7.3.5 Residual diagnostics

One assumption that was made to define the model presented above is the normality of the residuals. Figure 7.3 shows the normal Q-Q plot of the residuals from model 3. Even though all data points at 0 and 100 were excluded, the plot shows evidence for a fat-tailed (or heavy-tailed) distribution of the residuals. A preliminary fit of the model including the excluded values showed even more deviation from the normal distribution. For the sake of generalisation and validity of the test statistics, those values were excluded in the reported analysis. The results of test statistics for the random and fixed effects may therefore be taken with a grain of salt here. In a possible future analysis of the data, this may be taken into account by employing statistical methods that account for such behavior, but no R implementation of a robust method (*i.e.* accounting for heavy-tailed distributions) allowing for nested models with fixed and random effects could be found.

The variance across levels of the covariates is also assumed to be constant. This seems to be a reasonable assumption for this data as depicted by figures 7.4 through 7.8.
Finally, figure 7.9 shows the residuals from model 3 plotted against the fitted values. While the diagonal lines are simply due to participants grouping their responses (in this case towards multiples of 10) as explained by Searle [114], the apparent correlation between residuals and fitted value indicates unexplained effects in the model. This is confirmed by figure 7.10 where the correlation is once again apparent between the residuals and the subjective data. This is however expected with the explained variance attaining 61%. A general sense for the fit by

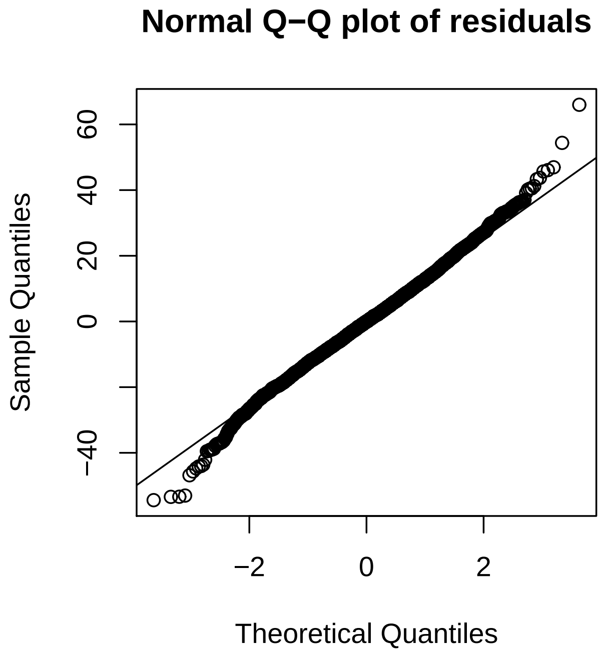


Figure 7.3 – Normal Q-Q plot of the residuals from model 3.

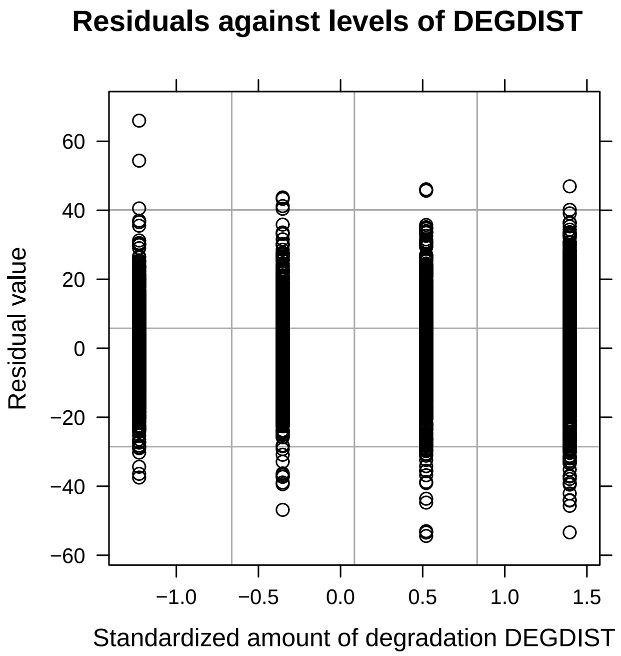


Figure 7.4 – Residuals of model 3 plotted against levels of degradation covariate DEGDIST.

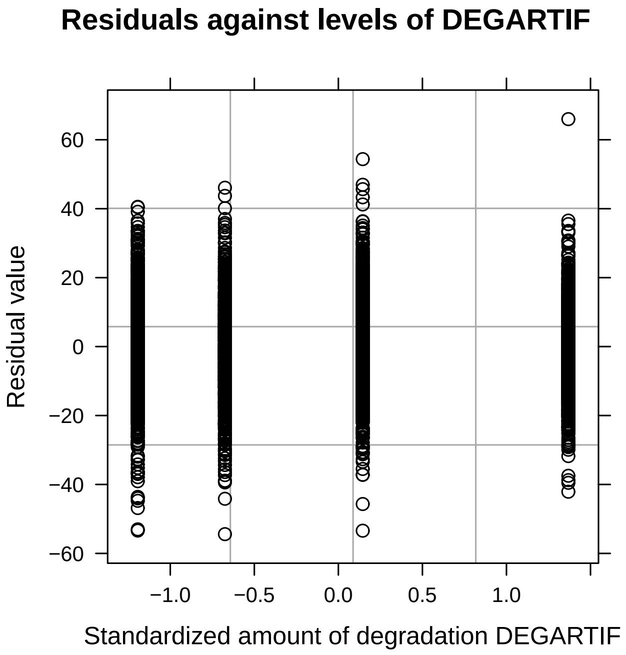


Figure 7.5 – Residuals of model 3 plotted against levels of degradation covariate DEGARTIF.

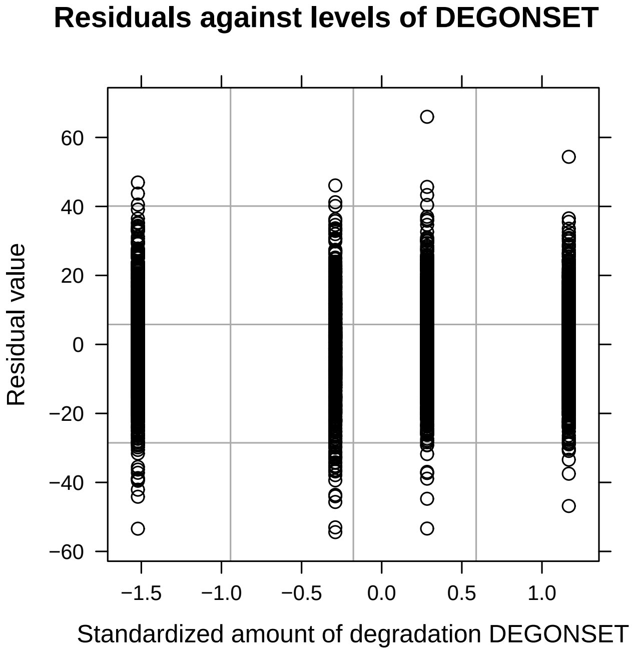


Figure 7.6 – Residuals of model 3 plotted against levels of degradation covariate DEGONSET.

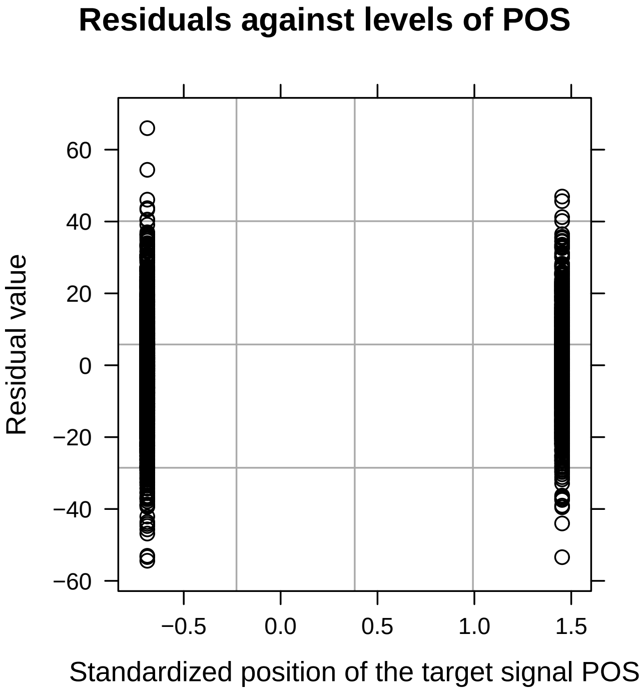


Figure 7.7 – Residuals of model 3 plotted against levels of the spatial configuration covariate POS.

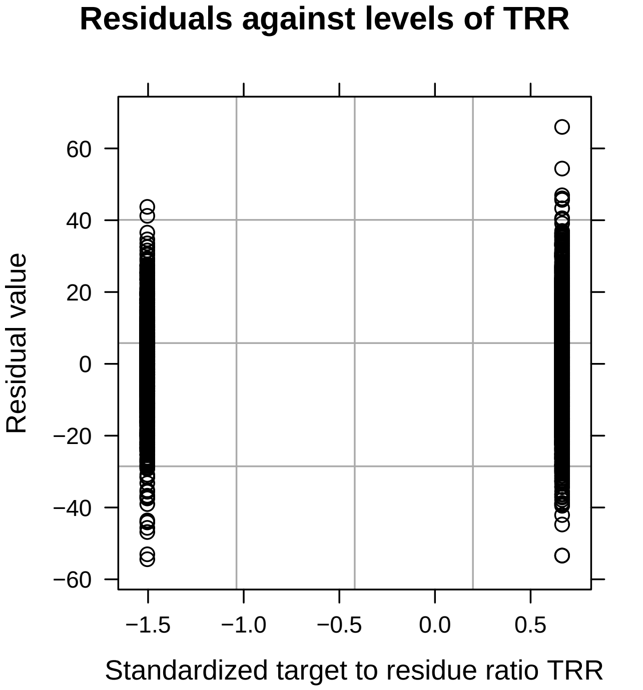


Figure 7.8 – Residuals of model 3 plotted against levels of the target to residue ratio covariate TRR.

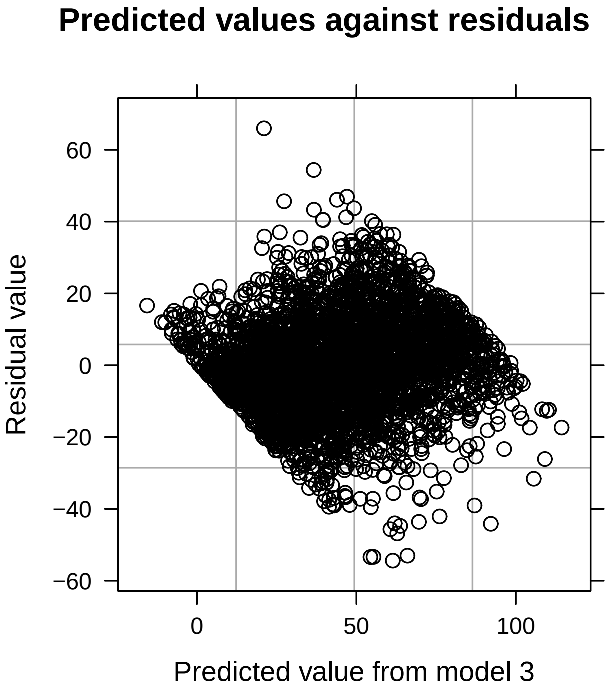


Figure 7.9 – Residuals from model 3 plotted against the predicted values.

model 3 can be gained from figure 7.11, illustrating the correlation between the predicted data by model 3 and the actual responses from participants.

7.4 Discussion

7.4.1 Detailed discussion

The fixed-effect parameter estimates reported in table 7.1 are based on scaled and centered variables. While the scaling and centering may seem counterintuitive in terms of interpretability, this allows for the direct comparison between the effects as pointed out by Schielzeth [112]. The scale of the scaled and centered variables are not in their original units any more but rather in units of their standard deviation. The reported fixed-effect parameters are therefore the estimated variation in perceived quality for an increment of 1 standard deviation in the corresponding variable. Note that the sign of the variation still depends on the units of the original variable.

While the reported intercept is not of much interest, since it does not have much meaning with centered and scaled variables, other fixed effects may be more interesting. The results report an increment of 9.51 in the expected score for an increment of 1 standard deviation in the degradation amount for degradation 1, *i.e.* in the low-pass filter frequency for the target distortion degradation family. The higher the cut-off frequency, the higher the expected

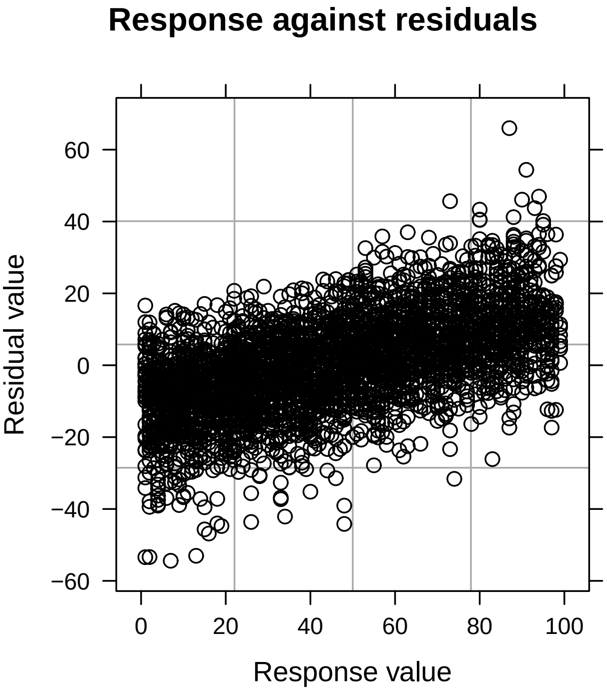


Figure 7.10 – Residuals from model 3 plotted against the data gained from participants.

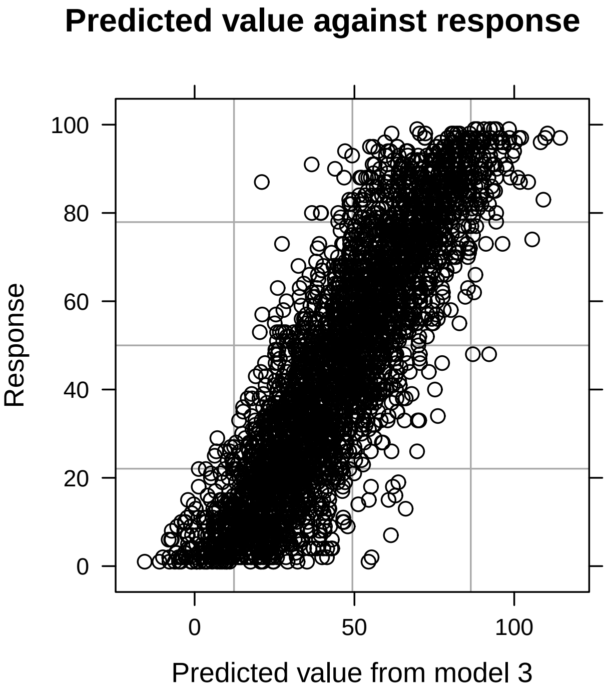


Figure 7.11 – Responses from participants plotted against the predicted values from model 3

response. The squared DEGDIST term is also significant with a value of 1.12 and adds to the positive effect. This corresponds to expected behavior.
For the musical noise-type artifacts degradation family, an estimated decrease in the expected response of -5.85 is reported. The more time/frequency bins exchanged, the lower the average perceived quality. Again, a significant quadratic term is reported (-1.17) that adds to the linear effect.
For the onset misallocation component, the main effect is reported as being significant with $p < .05$. Due to the restrictions stated in section 7.3.5, this could or could not be significant when accounting for the nonnormal residual distribution. No further comments will therefore be made here. The second-order term is however significant, but positive with a value of 2.62. This is again quite challenging to interpret. A closer look at the standardised values of the amount argument for the onset misallocation degradation allows for some comments. Standardised variables generally take values between -3.0 and 3.0 [112]. A plot indicating the combined value of the first-order and quadratic term over that range is therefore given in figure 7.12. Note that the standardised values taken by the DEGONSET input variable are $\{-1.51, -0.28, 0.28, 1.17\}$. Predictions below the lowest value of -1.51 do not make sense, since this value corresponds to a $a_{\text{onset}} = 0$. The four values taken by the input variable are highlighted by the crossed markers. The parabolic curve shows that the values taken by DEGONSET in this experiment mostly fall on its descending side. The last value however is on the ascending slope even though it is significantly lower than the first value. This means that the sign of the overall effect changes for this last amount argument level. It is suspected that this is an artifact due to the fitting of a quadratic term in the model in combination with the values of the amount argument a_{onset}. Even though a logarithmic scale is chosen, if the effect of an input variable is inversely proportional to it, a model with linear and quadratic terms will have problems fitting this behavior. Generalisation towards higher values might therefore be inaccurate and even for the chosen values, the model might show unexpected behavior. Note that the curve shown on 7.12 does not take into account interactions, which may be interpreted separately, since the input variables were standardised [112].

Since the degradation amount arguments were chosen such as to encompass realistic ranges of degradations encountered in real life, the collected data seems to show evidence for the predominance of target distortion over the other two types of degradations in terms of perceived quality degradation. The onset misallocation and musical noise-type artifact degradations are difficult to order by impact here, due to the second-order terms. While the main effect of DEGARTIF is bigger than that of DEGONSET, the inverse is the case for the quadratic effects.

The interactions between the degradation effects are all significant, but only the interaction between DEGDIST and DEGONSET is significant with $p < .01$ (in fact $p < .001$). While the other two interactions won't be commented due to the same reasons as the DEGONSET main effect above, this last interaction shows the same kind of behavior as for the binaural study. Figure 7.13 shows the combined effect of DEGDIST and DEGONSET. The effect of DEGDIST is dominating and the average response decreases with lower low-pass frequencies and higher amounts of onset degradation.

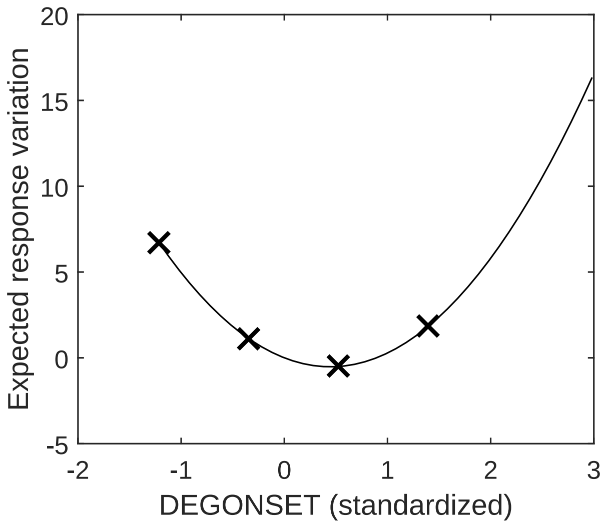


Figure 7.12 – Expected variation of the response due to linear and quadratic effects of DEGONSET. Actual values taken by the input variable are highlighted by crossed markers.

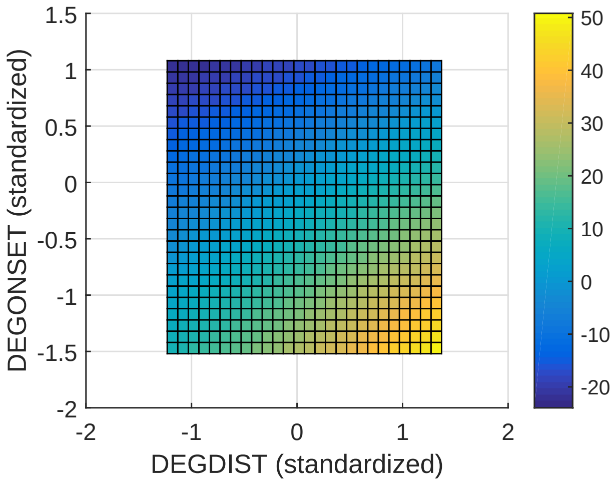


Figure 7.13 – Expected variation of the response due to the main, quadratic and interaction effects of DEGDIST and DEGONSET over the span of values covered in this experiment.

The TRR and POS effects are both significant and negative. This means that the response is lower when the target is spatialised to the right of the listener on average (including all degradation situations). It is also lower on average for a TRR of 12 dB than for 6 dB. The interaction between the TRR and the position of the target is also significant and positive. As for the binaural study, this can be interpreted as follows: If the TRR is 6 dB and the target is spatialised in front of the listener, the response is slightly better on average than expected only from the main effects. The same is true when the TRR is 12 dB and the target is spatialised to the right of the listener. For the other two combinations, the response is slightly worse than expected. Note that this interaction is also dominated by the two main effects and therefore only slightly alters the expected response.

All interactions between the degradations and the TRR are reported as being significant. While the magnitude of these interactions is quite small compared to the main effects, their sign indicates that for the 6 dB TRR, the main effects of the degradations are smaller whereas for the 12 dB TRR, they are bigger. This is expected, since the degradations are expected to affect quality more when the TRR is higher.

The interactions between the degradations and the position of the target are also all reported as being significant. Once again, their magnitude is quite small as compared to the main effects and their sign indicates a smaller effect of the degradations for the central position of the target than for the position to the right with the exception of the interaction with DEGDIST. Here the effect is the opposite, with the effect of the target distortion degradation being bigger for the central position. However, the magnitude of the main effects are substantially bigger than the interaction and its overall effect should therefore be minimal. These interactions (at least in the case of DEGARTIF and DEGONSET) may be interpreted as evidence for spatial unmasking of the degradations, with spatial separation between the residue and the target increasing the effect of the degradations.

7.4.2 Summary

Overall, the results in this WFS experiment seem to indicate that the results behave as expected. Higher degradation amounts result in lower perceived quality and spatial separation between the target and the residue or a higher TRR increase the impact of the degradations and therefore show evidence for binaural unmasking. The standardised variables allow to compare the impact of different degradations and a dominance of target distortion is reported here. The other two degradation types were found to be roughly equivalent in the WFS study, whereas in the binaural study, the onset misallocation degradation was found to be more important than the musical noise-type artifact degradation. The increased number of significant interactions in the WFS study can be taken as an indicator that the increased complexity of the spatialisation (complex WFS with psychoacoustic approximations reproduced within a given room as opposed to simple ITD spatialisation presented over headphones) seems to increase the complexity of the judgement of perceived quality as well.

7.5 Objective scores for this data set

7.5.1 PEASS-based approach with decomposition

The state-of-the-art Perceptive Evaluation Methods for Audio Source Separation (PEASS) model as well as the extension that was proposed in chapters 1 and 3 were used to predict the scores of this experiment. The test signals from the subjective study were recorded with a KEMAR manikin and the resulting binaural signals were used as inputs for the model, *i.e.* both the degraded target and the reference target and interfering sources were binaural signals. The model therefore operated on two channels with a degraded target as signal under test and the clean target as reference. The residue was specified as potentially interfering source.
The internal parameters of the model were kept at their optimised defaults (see section 3.3) with the filter length for the decomposition at 500 ms. The mapping however was optimised for the new data set, selecting the optimal number of neurons for the mapping and the optimal subset of features to be used. The Artificial Neural Network (ANN) training was kept similar to what was presented in section 3.3 with an 85-fold cross-validation (17 subjects × 5 mixtures). Multiple references were accounted for by using the corresponding Mean Objective Score (MOS). Note however that this MOS did not always have the same number of measurements, because of the D-optimal experimental design matrix. For that same reason, the training features were not based on the MOS but individual scores were used.

For the state-of-the-art model, a correlation of 0.57 was achieved between the subjective and the objective scores when using $n_{\text{neur}} = 5$ hidden neurons and feature vector $\mathbf{q}_j = [q_j^{\text{overall}}, q_j^{\text{target}}]$. This is slightly better than for the binaural study. Moreover, since the statistical model was able to explain only 60% of the variance by fixed effects only (*i.e.* knowing all degradation levels, TRRs and spatial configurations), this seems reasonable. Note that the R^2 measure reported for linear models corresponds to the coefficient of determination which is the square of the correlation coefficient. A coefficient of determination of 60% would therefore correspond to a correlation of 0.77. This can't be applied directly to linear mixed models, but it gives an idea of the performance of the objective model for the given situation. In this particular case, with all (significant and nonsignificant) parameters included, model 3 gives a correlation coefficient of 0.87 (the same as for the binaural study) between the predicted and measured responses. Note however that this is achieved by excluding all references, all responses with value 100 and all responses with value 0 due to constraints explained above. The objective models were trained including all stimuli, but using MOS for the hidden references and are therefore expected to be less accurate (but more generalizable).

The extended version of PEASS also achieved a correlation coefficient of 0.57, performing similarly to the state of the art. The number of hidden neurons was $n_{\text{neur}} = 6$ and the optimal feature vector was found to be $\mathbf{q}_j = [q_j^{\text{overall}}, q_j^{\text{artif}}]$ (identical to the state of the art). The modelling of the additional onset component therefore did not improve the prediction accuracy in this case. One possible explanation might be that the onset misallocation degradation component was less important in terms of perceived quality for this WFS study, as revealed by the statistical analysis above. The onset misallocation degradation main effect is barely significant. Note however that the quadratic term still is highly significant and further investigation of the reasons for this behavior may therefore be needed if the PEASS model is to be retained.

7.5.2 PEMO-Q-based approach without decomposition

The Perceptual Model of Audio Quality (PEMO-Q)-based approach proposed in chapter 5 was also tested against the subjective scores from the WFS experiment. As for the PEASS model, binaural recordings of the test signals from the WFS subjective study were used. The complete reference sound scene (target and residue) and the complete sound scene under test were recorded here as opposed to separate recordings for the target and the residue as used for the PEASS model.
The internal parameters of the PEMO-Q model were kept at their optimised defaults. Since the pre-study seemed to indicate a linear relationship between the predicted scores and the subjective scores, the mapping was based on a linear regression. While no cross-validation setting is used in this case, the weights for the high- and low-frequency features as well as a global intercept are estimated by linear regression, using the subjective scores as responses and the objective scores as predictors.

The PEMO-Q based model achieved a correlation coefficient of 0.67 for the data of the binaural study, which corresponds to an increase of 0.1 compared to the PEASS model. The computational requirements being lower as well, this is a major improvement over the state of the art.

7.6 Conclusion

This chapter presented a subjective study similar to what was presented in chapter 6 but conducted with WFS spatialisation. The main effects allow the same conclusions as for the binaural study: the amount arguments of the degradation synthesis algorithms are affecting perceived quality in a monotonic fashion and the target distortion degradation has the most impact in perceived quality. However, the onset misallocation degradation seems to be less important in the WFS study than in the binaural study. Interactions between TRR and degradations as well as position of the target and degradations suggest that both monaural unmasking and binaural unmasking play a role in perceived quality.

The basic PEASS model achieved a prediction accuracy of 0.57 for this new set of objective scores. The same performance was measured for the extended PEASS version. The more complex spatial configuration compared to the binaural experiment may explain the lack of performance increase between the two versions. Also, the fact that the statistical analysis only showed minor influence of the onset misallocation degradation component suggests that adding an onset degradation component in the objective model may not bring a performance increase.

The PEMO-Q based model outperformed the PEASS models by 0.1 points, achieving a correlation coefficient of 0.67 between objective and subjective scores. Once again, these results are judged satisfactory.

In the view of these results, the usefulness of performing a degradation decomposition is questioned. In fact, the optimal feature vectors of both the state-of-the-art version of PEASS as well as of the extension proposed in this thesis show that only a subset of the degradation components are used for the optimised mapping. While the different experimental tasks in the original release of PEASS did make use of different feature vectors, the overall quality measure only uses 2-3 components both for the subjective study reported by Emiya *et al.* [42] as well as in the experiments reported in this thesis.

8 Conclusion and perspectives

8.1 Context

The work presented in this thesis was conducted in the framework of the i3Dmusic project. The goal of this project is to enable the playback of existing audio content on spatial audio systems. Such systems require adapted mixing approaches with clean source tracks. In the case where those are not available, one solution is given by Audio Source Separation (ASS). Aiming at perfectly separating all sources from a given mixture, ASS is an active field of research. The modelling of sources to be extracted is a very complex mathematical and signal processing problem. As a consequence, the source estimation is not perfect and algorithms induce errors that result in more or less audible degradations in the extracted signals.

The audibility of those degradations has been investigated in the past with a resulting objective model known as Perceptive Evaluation Methods for Audio Source Separation (PEASS) [42]. Inspired from objective quality evaluation models for audio coding, it uses one of the more recent auditory models known as Perceptual Model of Audio Quality (PEMO-Q). The principle is to decompose the degradations from ASS into components associated with different typically encountered degradation types and then measure their perceptual impact. A nonlinear combination of the scores of the different components then yields a global quality measure for a given estimated source signal.

8.2 Results and original contributions

Extended PEASS decomposition

The approach taken by PEASS presented above does not take into account any binaural characteristics of the human auditory process. Dynamic interference degradations notably, which are postulated to be more important in terms of perceived quality when binaural hearing is involved, are not modelled in the state-of-the-art PEASS model. An extension to the PEASS model is therefore proposed, allowing for separate modelling of static and dynamic interference from other sources. After optimising the internal parameters of the model, the performance of the extended PEASS model is evaluated against the subjective data from the PEASS model and is found to be identical to the state of the art.

Artificial ASS degradation algorithms

Since one of the objectives of this thesis was to gain first subjective data for ASS evaluation in the context of 3D audio, controlled experimental conditions had to be achieved. While an experimental protocol has been proposed by the International Telecommunication Union (ITU) for audio quality evaluation purposes which has been used for subjective ASS evaluation for PEASS, controlled signals containing a given amount of certain degradation types were not available. Based on the low anchors from PEASS, a set of 4 degradation algorithms is proposed. Their validity is tested in a perceptual pre-study with stimuli containing only one type of degradation. The perceived quality is found to decrease monotonically with increasing degradation for all degradation types. Additionally, valuable user feedback is gained that is used to improve the degradation synthesis algorithms and the experimental protocol.

Binaural modelling approach based on PEMO-Q

Since the PEASS model does not integrate any psychoacoustic knowledge about binaural hearing, a binaural model based on the PEMO-Q monaural auditory model is also proposed. Better-ear hearing is modelled by computing the instantaneous perceptual similarity of the sound scene under test to the undegraded reference sound scene in every ear. The minimum similarity between both ears for every time frame is then used to compute a global similarity measure. The influence of the Interaural Cross-Correlation (IACC) in binaural hearing is taken into account by computing the perceptual similarity between the left and the right channel for the scene under test as well as for the reference sound scene. The resulting high- and low-frequency features can then be combined into a single objective score via a linear mapping. A first performance evaluation with the subjective data from the pre-study indicates a correlation of 0.92 between subjective data and objective scores from the proposed model.

Subjective ASS evaluation study in a binaural rendering context

A subjective study using binaural spatialisation (Interaural Time Difference (ITD) synthesis only) and 3 out of the 4 degradation synthesis algorithms is presented. Results indicate that the degradation synthesis algorithms impact perceived quality as expected. The target distortion degradation is found to be the most harmful for perceived quality, followed by onset misallocation (the dynamic interference) and musical noise-type artifacts. Also, spatially separating the estimated target source (voice) from the residue (music) seems to be perceived as a degradation in itself.

The basic and the extended versions of PEASS are tested against the subjective data gained from the binaural study. The extended version performs slightly better than the basic version in this case with a correlation coefficient of 0.57 and 0.53 respectively. The performance of the proposed binaural model is also evaluated. It is found to be even better than the extended PEASS version with a correlation coefficient of 0.60.

Vertical localisation study

In order to gain additional insight into spatial listening conditions, a 3D Wave Field Synthesis (WFS) system is considered. The WFS method that is used, is based on a novel practical formulation extending WFS to the third dimension without requiring an unrealistic number of loudspeakers. While the WFS algorithm development and implementation was not part of this thesis, its validity was evaluated by conducting a vertical localisation study. Results indicate that the WFS formulation is able to accurately render height, showing better performance than other spatialisation methods.

Subjective ASS evaluation study in a WFS rendering context

Finally, a second subjective ASS evaluation study is reported using this WFS implementation. The experimental design is the same as for the binaural study so that effects can be compared. Results indicate that the onset misallocation degradation is less important in that case, whereas target distortion still dominates the perceived quality. Additionally, the data shows evidence for binaural and monaural unmasking, which is not the case in the binaural spatialisation experiment.

The basic and extended PEASS models as well as the proposed binaural modelling approach are tested against the subjective data from the WFS study. The basic and extended versions of PEASS are found to show identical performance in this case, with correlation coefficients of 0.57. The binaural modelling approach reaches 0.67, showing clear improvement over the two PEASS versions.

8.3 Perspectives

The objective of this thesis was to give first insights into ASS evaluation in the context of spatial audio. To do so, subjective experiments were conducted, and first modelling approaches investigated. While the results of the subjective studies show evidence for the need of binaural models and while the proposed binaural modelling approach outperforms monaural models, these are just first steps into ASS evaluation in the context of 3D audio. Many improvements and additional investigations may be taken to continue this work. More formal subjective ASS studies are needed, isolating single effects and studying movement of sources fore example. The following paragraphs shall give an overview of potential issues to be addressed and/or future work that could be done in the field.

8.3.1 Subjective ASS evaluation

The subjective ASS evaluation studies presented here have been conducted with a single task addressing overall quality. While this scale gives a sense of the overall perceived quality, it also includes all other effects that may be subjectively perceived as quality degradations, such as target position or Target to Residue Ratio (TRR) in the presented experiments. The subjective experiment presented with PEASS used a set of 3 additional tasks addressing specifically each of the proposed degradation components. While this allows to construct scales that clearly relate to certain degradation types, it does exclude all degradations that may come from ASS but don't fall into one of the proposed categories. In fact, there is no study known to the author that formally classifies the degradations encountered in real-life ASS algorithms. This might therefore be a path to explore. With such a categorisation, degradations types could then be more formally established, together with the corresponding decomposition methods for objective ASS evaluation and the experimental tasks for subjective studies. This might also partly explain why only certain components of the degradation are used when mapping PEASS scores to subjective scores.

An additional point that could be addressed is that of the subjective grading protocol. In the studies conducted for this thesis (as well as in the PEASS model), a modified Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) protocol, as defined by the ITU, is used. This protocol presents the problem that the low end of the scale is more or less floating, since it is defined by the low anchors (ideally). Generalisation of the results may therefore be difficult. Additionally, for studies with a large number of stimuli as it is the case in the presented experiments, participants may have a hard time remembering what they subjectively defined to be the low end of the scale based on the training phase of the experiment. This was addressed in the presented studies by periodically showing the training mask with the highest and lowest quality stimuli again. This does however not resolve the problem with the lower end of the scale and a protocol mitigating this issue would therefore be helpful.

8.3.2 Objective ASS quality evaluation modelling

Extended PEASS model

While the proposed PEASS extension does not perform as well as the binaural model in the spatial audio context, it still achieves very high prediction accuracy in the more traditional multichannel approach without considering spatialisation. Even though the proposed extension shows performance that is equal to that of the basic version in that setting, one point could be explored to see if the prediction accuracy could be improved. In fact, the onset misallocation component in the error decomposition uses the same filter length as the target distortion component. The filter length was determined such as to maximize prediction accuracy, but it may be worth trying different filter lengths for the two components, since typical onset misallocations are much shorter than the 500 ms filter length. The target distortion component was shown to be more important in overall quality than the onset misallocation component during the subjective experiments and it is therefore suspected that using two filter lengths would improve the prediction accuracy. But this might also increase the computational requirements further.

Additionally, it is suspected that the proposed static versus dynamic interference decomposition algorithm might assign too much of the error component to the static interference. As soon as some interference is present in some of the frames, the projection for static interference will be nonzero, even though it might be entirely due to dynamic interference. Investigating a way of mitigating this might be beneficial to the prediction accuracy of the extended PEASS version.

PEMO-Q-based binaural modelling approach

Regarding the proposed binaural model, it is clear that the binaural model in its actual state is based on the fact that the components of the back-end of PEMO-Q can be used to roughly simulate cues that are used to model binaural unmasking and better-ear hearing. A more thorough approach towards the modelling of those phenomena by appropriately modifying the Perceptual Similarity Measure (PSM) for example would be desirable.

Harlander *et al.* [51] proposed to use the Computational Auditory Signal-processing and Perception model (CASP) as proposed by Jepsen *et al.* [69] as a front-end for their audio quality evaluation in combination with the back-end of PEMO-Q. The CASP is superior to the auditory stage of PEMO-Q in the sense that it also models the nonlinear behavior of the basilar membrane, adds an outer- and middle-ear transfer function and a square-law behavior of nerve firing rate as a function of the Sound Pressure Level (SPL) [51]. An absolute hearing threshold depending on frequency was also implemented. Since the resulting CASP-Q model (with no expansion stage though) seems to slightly outperform PEMO-Q in most settings, this would be worth investigating for the binaural model proposed in this thesis as well.

A Appendix: Short introduction to Linear Mixed Models

To give the noninitiated reader a better understanding of the statistical analysis of the presented subjective Audio Source Separation (ASS) evaluation studies, a brief introduction to statistical modelling of data using Linear Mixed Models (LMMs) is given here. For a more complete introduction to LMMs, see Bates [6] or West *et al.* [140]. The notation of West *et al.* [140] is adopted here.

The framework for statistical analysis is similar most of the time. Some variable (the dependent variable) is observed as a response to some other variables that are recorded along the way (the independent variables). For perceptual experiments as described here, the response is often elicited by some stimuli with varying characteristics under conditions that may be varying as well. These variations of the experimental environment and the stimuli are called covariates. The statistical analysis of all the recorded variables then aims at describing the relationship between the response and the covariates.
Covariates can be classified as quantitative, describing a variable which contains magnitude information based on a quantitative scale, or as categorical, describing a variable that can take a given set of values . The values of quantitative covariates have a meaning (e.g. weight, temperature, etc.) whereas the values of categorical covariates are simply labels used to designate the given level. Categorical covariates are often called factors.

Statistical analysis aims at describing the relationship between the response and the covariates using so-called effects, which are parameters associated with particular levels of the independent variables (or covariates in general when they are quantitative). When the set of possible levels of a covariate is fixed and reproducible, the covariate is modelled using fixed-effect parameters. If the values represent a random sample from all possible levels, the corresponding covariate is modelled using random effects.

When containing only fixed effects, a linear model can be expressed as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{A.1}$$

where $\mathbf{y}$ is the vector of observed responses (*i.e.* one realisation of the random variable $\mathscr{Y}$), $\mathbf{X}$ is the design matrix of the experiment, containing all the information about the covariates and β is the unknown vector of coefficients (the fixed effects) minimising the residue vector ϵ resulting in the property $E(\mathbf{y}) = \mathbf{X}\beta$ where $E(\cdot)$ denotes the expected value (resulting in the mean, if $\mathscr{Y}$ is normally distributed). A linear fixed-effects model therefore approximates the data of an experiment in terms of an overall mean (modelled in β_0) and slopes for every covariate.

Random effects are different from fixed effects in the sense that they represent variations that are uncorrelated to the levels of the corresponding independent variables, whereas fixed effects represent variations that are correlated to the levels of the corresponding covariates. They give an estimation of the variance of the underlying random variable $\mathscr{U}$ and not of its expected value.
A mixed-effects model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon \tag{A.2}$$

where $\mathbf{Z}$ is the design matrix of the random effects and $\mathbf{u}$ is the unknown vector of random effects with a given variance-covariance matrix var($\mathbf{u}$) = $\mathbf{G}$. $\mathbf{Z}$ presents a structure that is often assumed to be known, based on the design of the experiment. When establishing the model, $\mathbf{u}$ and ϵ are assumed to follow multidimensional normal distributions:

$$\mathbf{u} \sim \mathscr{N}(\mathbf{0}, \mathbf{G}) \tag{A.3}$$

$$\epsilon \sim \mathscr{N}(\mathbf{0}, \mathbf{R}) \tag{A.4}$$

where $\mathbf{R}$ is the covariance matrix of the residuals. Most of the time, $\mathbf{R}$ is supposed to have a diagonal structure in which case the residuals associated with observations on the same block of observations (generally one level of the random effects) are assumed to be uncorrelated and have equal variance σ^2. In that case, $\mathbf{R} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The $\mathbf{R}$ matrix can also have different assumed structures with more or less parameters to be estimated, depending on the experimental design. The parameters of $\mathbf{R}$ are more generally denoted θ_R. The variance-covariance matrix of the random effects $\mathbf{G}$ is generally a block-diagonal matrix with blocks on the diagonal defined by the $\mathbf{D}$ matrix. $\mathbf{D}$ is the variance-covariance matrix for the different levels of the random effects $\mathbf{u}_i$. Its structure can be specified in a similar way to that of the $\mathbf{R}$ matrix, resulting in a given number of parameters θ_D to be estimated.

Different algorithms can be used to estimate the parameters β, θ_D and θ_R by optimising a likelihood function based on assumptions about the distributions of the parameters. The description of the involved algorithms and computational methods is beyond the scope of this short introduction, as it is only supposed to allow the interpretation of the models described in chapters 6 and 7. Some of the advantages of LMMs are that they allow for unbalanced designs (not all blocks of the same size) and for missing data points if needed. Since the design that was chosen for the subjective ASS evaluation experiments presented in this thesis is D-optimal and therefore unbalanced, LMMs are a sensible choice.

Bibliography

[1] J. Ahrens and S. Spors. Applying the ambisonics approach on planar and linear arrays of loudspeakers. In *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, France, May 6-7 2010.

[2] J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, 2(1):3–29, January 1985.

[3] J. Antoni, F. Guillet, M. El Badaoui, and F. Bonnardot. Blind separation of convolved cyclostationary processes. *Signal Process.*, 85(1):51–66, 2005.

[4] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux. The 2011 signal separation evaluation campaign (SiSEC2011): Audio source separation. In *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 414–422, Tel Aviv, Israel, March 2012.

[5] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. Duong. The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation. In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 114–122, St. Malo, France, September 2010.

[6] D. M. Bates. *lme4: Mixed-effects modeling with R.* Book draft, 2010.

[7] J. G. Beerends and J. A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963–978, December 1992.

[8] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, September 2005.

[9] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[10] S. Bertet, J. Daniel, E. Parizet, L. Gros, and O. Warusfel. Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3D microphones. In *AES 30th International Conference*, Saariselkä,

Finland, March 2007.

[11] V. Best, E. Ozmeral, F. J. Gallun, K. Sen, and B. G. Shinn-Cunningham. Spatial unmasking in human listeners: Energetic and informational factors. *Journal of the Acoustical Society Of America*, 118(6):3766–3773, December 2005.

[12] E. Blanco-Martín, F. J. Casajús-Quirós, J. J. Gómez-Alfageme, and L. I. Ortiz-Berenguer. Objective measurement of sound event localization in horizontal and median planes. *Journal of the Audio Engineering Society*, 59(3):124–136, March 2011.

[13] J. Blauert. Sound localization in the median plane. *Acustica*, 22:205–213, 1969-70.

[14] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, Massachussetts, revised edition, 1999.

[15] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloğlu. *The Technology of Binaural Listening*, chapter Further Challenges and the Road Ahead, pages 477–501. Springer, ASA Press, Heidelberg, Germany, 2013.

[16] K. Brandenburg. Evaluation of quality for audio encoding at low bit rates. In *82nd AES Convention*, London, March 10-13 1987.

[17] A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, January/February 2000.

[18] F. Chen. The reaction time for subjects to localize 3D sounds via headphones. In *AES 22nd International Conference*, Aspoo, Finland, June 15-17 2002.

[19] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society Of America*, 25(5):975–979, September 1953.

[20] L. Chittka and A. Brockmann. Perception space - the final frontier. *PLoS Biology*, 3(4):564–568, April 2005.

[21] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee. Review blind source separation and independent component analysis: A review, 2004.

[22] H. Chung, S. B. Chon, J.-h. Yoo, and K.-M. Sung. Analysis of frontal localization in double layered loudspeaker array system. In *Proceedings of 20th International Congress on Acoustics*, Sydney, Australia, 23-27 August 2010.

[23] M. Cobos, J. J. Lopez, A. Gonzalez, and J. Escolano. Stereo to wave-field synthesis music up-mixing: An objective and subjective evaluation. In *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pages 1279–1284, Malta, March 12-14 2008.

[24] C. Colomes, M. Lever, J.-B. Rault, Y.-F. Dehery, and G. Faucon. A perceptual model applied to audio bit-rate reduction. *Journal of the Audio Engineering Society*, 43(4):233–240, April 1995.

[25] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Elsevier Academic Press, New York, NY, USA, 2010.

[26] R. Conetta, F. Rumsey, S. Zieliński, P. J. B. Jackson, M. Dewhirst, S. Bech, D. Meares, and S. George. QESTRAL (part 2): Calibrating the qestral model using listening test data. In *125th AES Convention*, San Francisco, USA, October, 2-5 2008.

[27] E. Corteel. Equalization in an extended area using multichannel inversion and wave field synthesis. *Journal of the Audio Engineering Society*, 54(12):1140–1161, December 2006.

[28] E. Corteel. Synthesis of directional sources using wave field synthesis, possibilities and limitations. *EURASIP Journal on Advances in Signal Processing*, 2007(1):188–205, January 2007.

[29] E. Corteel, R. Pellegrini, and C. Kuhn-Rahloff. Wave field synthesis with increased aliasing frequency. In *AES 124th Convention*, Amsterdam, The Netherlands, May 17-20 2008.

[30] E. Corteel, L. Rohr, X. Falourd, K.-V. Nguyen, and H. Lissek. A practical formulation of 3 dimensional sound reproduction using wave field synthesis. In *International Conference on Spatial Audio 2011*, Detmold, Germany, 2011.

[31] E. Corteel, L. Rohr, X. Falourd, K.-V. NGuyen, and H. Lissek. Practical 3 dimensional sound reproduction using wave field synthesis, theory and perceptual validation. In *Acoustics 2012*, Nantes, France, April 23-27 2012.

[32] J. F. Culling, M. L. Hawley, and R. Y. Litovsky. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *Journal of the Acoustical Society Of America*, 116(2):1057–1065, August 2004.

[33] J. F. Culling, M. L. Hawley, and R. Y. Litovsky. Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [j. acoust. soc. am.116, 1057 (2004)]. *Journal of the Acoustical Society Of America*, 118(1):552, July 2005.

[34] J. Daniel. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *AES 23rd International Conference*, Helsingøor, Denmark, May 23-25 2003.

[35] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the “effective” signal

processing in the auditory system. i. model structure. *Journal of the Acoustical Society Of America*, 99(6):3615–3622, June 1996.

[36] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the “effective” signal processing in the auditory system. ii. simulations and measurements. *Journal of the Acoustical Society Of America*, 99(6):3623–3631, June 1996.

[37] W. P. J. De Bruijn. *Application of Wave Field Synthesis in Videoconferencing*. PhD thesis, Delft University of Technology, Delft, The Netherlands, 2004.

[38] L. Deng and D. Yu. *Deep learning - Methods and Applications*, volume 7 of *Foundations and Trends in Signal Processing*. Now Publishers Inc, Hanover, MA, USA, 2014.

[39] Deutsches Institut für Normung (DIN). DIN 45631: Procedure for calculating loudness level and loudness, 1991.

[40] M. Dewhirst, R. Conetta, F. Rumsey, P. Jackson, S. Zieliński, S. George, S. Bech, and D. Meares. QESTRAL (part 4): Test signals, combining metrics and the prediction of overall spatial quality. In *125th AES Convention*, San Francisco, USA, October 2-5 2008.

[41] N. I. Durlach. *Foundations of Modern Auditory Theory*, volume II, chapter Binaural signal detection: Equalization and cancellation theory, pages 371–462. Academic Press, New York, NY, USA, 1972.

[42] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, September 2011.

[43] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold. *Design of Experiments - Principles and Applications*. Umetrics AB, Umea, Sweden, third revised and enlarged edition, 2008.

[44] R. Feldtkeller and E. Zwicker. *Das Ohr als Nachrichtenempfänger*, volume XIX of *Monographien der elektrischen Nachrichtentechnik*. S. Hirzel Verlag Stuttgart, Stuttgart, Germany, 1956.

[45] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12:47–66, 1940.

[46] M. A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, February 1973.

[47] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, August 1990.

[48] D. F. M. Goodman and R. Brette. Spike-timing-based computation in sound localization. *PLoS Computational Biology*, 6(11):e1000993, November 2010.

[49] P. Guillon. *Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*. PhD thesis, Université du Maine, Le Mans, France, 2009.

[50] M. Hansen and B. Kollmeier. Objective modelling of speech quality with a psychoacoustically validated auditory model. *Journal of the Audio Engineering Society*, 48(5):395–409, May 2000.

[51] N. Harlander, R. Huber, and S. D. Ewert. Sound quality assessment using auditory models. *Journal of the Audio Engineering Society*, 62(5):324–336, May 2014.

[52] J. Herre, E. Eberlein, H. Schott, and K. Brandenburg. Advanced audio measurement system using psychoacoustic properties. In *92nd AES Convention*, Vienna, March 24-27 1992.

[53] T. Herzke and V. Hohmann. Improved numerical methods for gammatone filterbank analysis and synthesis. *Acta Acustica united with Acustica*, 93(3):498–500, 2007.

[54] V. Hohmann. Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica*, 88(3):433–442, 2002.

[55] M. P. Hollier, D. R. Guard, and M. J. Hawksford. Objective perceptual analysis: Comparing the audible performance of data reduction schemes. In *96th AES Convention*, Amsterdam, The Netherlands, February 26 - March 01 1994.

[56] M. P. Hollier, M. J. Hawksford, and D. R. Guard. Characterization of communication systems using a speechlike test stimulus. *Journal of the Audio Engineering Society*, 41(2):1008–1021, December 1993.

[57] M. P. Hollier, M. O. Hawksford, and D. R. Guard. Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain. *IEE Proceedings Vision, Image & Signal Processing*, 141(3):203–208, June 1994.

[58] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902–1911, November 2006.

[59] M. Z. Ikram. Blind separation of delayed instantaneous mixtures: a cross-correlation based approach. *International Journal of Adaptive Control and Signal Processing*, 18(3):265–278, 2004.

[60] International Standardization Organisation (ISO). ISO 532:1975 acoustics - method for calculating loudness level, 1975.

[61] International Standardization Organisation (ISO). ISO 266:1997 acoustics - preferred frequencies, 1997.

[62] International Telecommunication Union (ITU). Recommendation ITU-R BS.468-4: Measurement of audio-frequency noise voltage level in sound broadcasting, July 1986.

[63] International Telecommunication Union (ITU). Recommendation ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, October 1997.

[64] International Telecommunication Union (ITU). Recommendation ITU-R BS.1534: Method for the subjective assessment of intermediate quality level of coding systems, 2001.

[65] International Telecommunication Union (ITU). Recommendation ITU-R BS.1387-1: Method for objective measurements of perceived audio quality, 2002.

[66] International Telecommunication Union (ITU). Recommendation ITU-T P.863: Perceptual objective listening quality assessment, September 2014.

[67] P. Jackson, M. Dewhirst, R. Conetta, F. Rumsey, D. Meares, S. Bech, and S. George. QESTRAL (part 3): system and metrics for spatial quality prediction. In *125th AES Convention*, San Francisco, USA, October 2-5 2008.

[68] S. Jelfs, J. F. Culling, and M. Lavandier. Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, 275:96–104, 2011.

[69] M. L. Jepsen, S. D. Ewert, and T. Dau. A computational model of human auditory signal processing and perception. *Journal of the Acoustical Society Of America*, 124(1):422–438, July 2008.

[70] R. Kapust. A human ear related objective measurement technique yields audible error and error margin. In *11th AES International Conference*, pages 191–202, 1992.

[71] M. Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '85*, volume 10, pages 608–611, April 1985.

[72] M. Karjalainen. A binaural auditory model for sound quality measurements and spatial hearing studies. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*, volume 2, pages 985–988, Atlanta, GA, USA, May 7-10 1996.

[73] J. Käsbach, S. Favrot, and J. Buchholz. Evaluation of a mixed-order planar and periphonic ambisonics playback implementation. In *Forum Acusticum 2011*, Aalborg, Denmark, 27 June - 1 July 2011.

[74] S. M. Kay. *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1993.

[75] R. B. King and S. R. Oldfield. The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays. *Human Factors*, 39(2):287–295, June 1997.

[76] E. I. Knudsen and M. S. Brainard. Creating a unified representation of visual and auditory space in the brain. *Annual Review of Neuroscience*, 18:19–43, 1995.

[77] M. Kolundžija, C. Faller, and M. Vetterli. Reproducing sound fields using MIMO acoustic channel inversion. *Journal of the Audio Engineering Society*, 59(10):721–734, October 2011.

[78] R. Lambert. Difficulty measures and figures of merit for source separation. In *Proceedings of the International Symposium on Independent Component Analysis and Blind Source Separation (ICA 99)*, pages 133–138, Aussois, France, January 1999.

[79] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60(2):115–142, July 1992.

[80] V. Larcher and J.-M. Jot. Techniques d'interpolation de filtres audio-numériques, application à la reproduction spatiale des sons sur écouteurs. In *Congrès Français d'Acoustique*, Marseille, France, April 1999.

[81] M. Lavandier and J. F. Culling. Prediction of binaural speech intelligibility against noise in rooms. *Journal of the Acoustical Society of America*, 127(1):387–399, January 2010.

[82] T. Letowski and S. Letowski. *Advances in Sound Localization*, chapter Localization Error: Accuracy and Precision of Auditory Localization, pages 55–78. InTech, 2011.

[83] S. Makino, T.-W. Lee, and H. Sawada, editors. *Blind Speech Separation*. Springer, New York, NY, USA, 2007.

[84] J. C. Makous and J. C. Middlebrooks. Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society Of America*, 87(5):2188–2200, May 1990.

[85] J. Merimaa. *Analysis, synthesis and perception of spatial sound - binaural localization modeling and multichannel loudspeaker reproduction*. PhD thesis, Helsinki University of technology, Helsinki, Finland, August 2006.

[86] B. C. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, London, UK, 1991.

[87] B. C. J. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society Of America*, 74(3):750–753, 1983.

[88] J. M. Moore, D. J. Tollin, and T. C. Yin. Can measures of sound localization acuity be related to the precision of absolute location estimates? *Hearing Research*, 238(1-2):94–

109, April 2008.

[89] S. Münkner and D. Püschel. *Contributions to psychological acoustics: Results of the Sixth Oldenburg Symposium on Psychological Acoustics*, chapter A psychoacoustical model for the perception of non-stationary sounds, pages 121–134. Bibliotheks- und Informationssystem der Universität Oldenburg, Oldenburg, Germany, 1993.

[90] S. Nakagawa and H. Schielzeth. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, February 2013.

[91] M. Naoe, T. Kimura, Y. Yamakata, and M. Katsumoto. Performance evaluation of 3D sound field rerpoduction system using a few loudspeakers and wave field synthesis. In *Second International Symposium on Universal Communication (ISUC'08)*, pages 36–41, Osaka, December 15-16 2008.

[92] G. Ng. *Elevation localization of single- and multiple-band noises.* PhD thesis, Boston University, Boston, USA, 2005.

[93] R. Nicol. *Binaural Technology.* AES Monograph. Audio Engineering Society, 2010.

[94] S. R. Oldfield and S. P. A. Parker. Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions. *Perception*, 13:581–600, 1984.

[95] OpenStax College. Anatomy & physiology, connexions web site. online, January 5 2015.

[96] R. G. Pachella. *Human Information Processing: Tutorials in Performance and Cognition*, chapter The interpretation of reaction time in information processing research, pages 40–82. Lawrence Erlbaum Associates, New York, USA, 1974.

[97] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne. PERCEVAL: Perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society*, 40(1/2):21–31, February 1992.

[98] R. D. Patterson. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society Of America*, 59(3):640–654, March 1976.

[99] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. Technical report, Annex B of the SVOS final report, December 1987.

[100] I. N. Pieleanu. Localization performance with low-order ambisonics auralization. Masters thesis, Rensselaer Polytechnic Institute, Troy, NY, USA, August 2004.

[101] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: principles, algorithms, and applications.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[102] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[103] V. Pulkki and T. Hirvonen. Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, 13(1):105–119, January 2005.

[104] D. Purves and S. M. Williams. *Neuroscience*. Sinauer Associates, Sunderland, MA, USA, 2nd edition, 2001.

[105] D. Püschel. *Prinzipien der zeitlichen Analyse beim Hören*. PhD thesis, Georg-August-Universität Göttingen, Göttingen, Germany, 1988.

[106] B. Rakerd and W. M. Hartmann. Localization of sound in rooms, ii: The effect of a single reflecting surface. *Journal of the Acoustical Society Of America*, 78(2):524–533, August 1985.

[107] L. Rohr, E. Corteel, K.-V. NGuyen, and H. Lissek. Vertical localization performance in a practical 3-D WFS formulation. *Journal of the Audio Engineering Society*, 61(12):1001–1014, December 2013.

[108] M. Rossi. *Audio*. Electricité. Presses Polytechniques et Universitaires Romandes, Lausanne, 2007.

[109] F. Rumsey, S. Zieliński, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares. QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using and artificial listener. In *125th AES Convention*, San Francisco, USA, October, 2-5 2008.

[110] A. Sander. *Psychoakustische Aspekte der subjektiven Trennbarkeit von Klängen*. PhD thesis, Universität Oldenburg, Oldenburg, Germany, 1994.

[111] J. Sanson, E. Corteel, and O. Warusfel. Objective and subjective analysis of localization accuracy in wave field synthesis. In *AES 124th Convention*, Amsterdam, The Netherlands, May 2008.

[112] H. Schielzeth. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1:103–113, June 2010.

[113] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society Of America*, 66(6):1647–1652, June 1979.

[114] S. R. Searle. Parallel lines in residual plots. *The American Statistician*, 42(3):211, 1988.

[115] A. Sekey and B. A. Hanson. Improved 1-bark bandwidth auditory filter. *Journal of the Acoustical Society of America*, 75(8):1902–1904, June 1984.

[116] B. G. Shinn-Cunningham. Influences of spatial cues on grouping and understanding sound. In *Proceedings of Forum Acusticum 2005*, Budapest, Hungary, August 29 - September 2 2005.

[117] B. G. Shinn-Cunningham and K. Kawakyu. Neural representation of source direction in reverberant space. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 79–82, New Paltz, New York, USA, October 19-22 2003.

[118] L. S. R. Simon and E. Vincent. A general framework for online audio source separation. In *10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA) 2012*, Tel-Aviv, Israel, March 12-15 2012.

[119] J. Soumagne, P. Mabilleau, S. Morissette, G. Chouinard, and D. Bennett. A comparative study of the proposed high quality coding schemes for digital music. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '86*, volume 11, pages 21–24. IEEE, April 1986.

[120] T. Sporer. Objective audio signal evaluation - applied psychoacoustic for modeling the perceived quality of digital audio. In *Proceedings of the 103rd AES Convention*, New York, USA, September 26-29 1997.

[121] S. Spors and J. Ahrens. Analysis and improvement of pre-equalization in 2.5-dimensional wave field synthesis. In *128th AES Convention*, London, UK, May 22-25 2010.

[122] S. Spors, R. Rabenstein, and J. Ahrens. The theory of wave field synthesis revisited. In *124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, May 17-20 2008.

[123] E. Start. *Direct Sound Enhancement by Wave Field Synthesis*. PhD thesis, TU Delft, Delft, The Netherlands, 1997.

[124] S. S. Stevens and J. Volkmann. The relation of pitch to frequency: a revised scale. *The American Journal of Psychology*, 53(3):329–353, July 1940.

[125] H. W. Strube. A computationally efficient basilar-membrane model. *Acustica*, 58(4):207–214, September 1985.

[126] J. R. Stuart. Noise: Methods for estimating detectability and threshold. *Journal of the Audio Engineering Society*, 42(3):124–140, March 1994.

[127] T. Thiede and E. Kabot. A new perceptual quality measure for bit-rate reduced audio. In *100th AES Convention*, Copenhagen, Denmark, May 11-14 1996.

[128] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ - the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*,

48(1/2):3–29, January/February 2000.

[129] C. Travis. A new mixed-order scheme for ambisonic signals. In *Ambisonics Symposium 2009*, Graz, Austria, 25-27 June 2009.

[130] E. N. G. Verheijen. *Sound Reproduction by Wave Field Synthesis.* PhD thesis, TU Delft, Delft, The Netherlands, 1997.

[131] E. Vincent. Improved perceptual metrics for the evaluation of audio source separation. In F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, editors, *Latent Variable Analysis and Signal Separation*, volume 7191 of *Lecture Notes in Computer Science*, pages 430–437. Springer Berlin Heidelberg, 2012.

[132] E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 734–741, 2009.

[133] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, July 2006.

[134] E. Vincent, M. G. Jafari, and M. D. Plumbey. Preliminary guidelines for subjective evaluation of audio source separation algorithms. In *Proceedings of the UK ICA Research Network Workshop 2006*, 2006.

[135] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca. First stereo audio source separation campaign: Data, algorithms and results. In *Proceedings of the 7th international conference on independent component analysis and signal separation (ICA)2007*, pages 552–559, 2007.

[136] P. Vogel. *Application of Wave Field Synthesis in Room Acoustics.* PhD thesis, TU Delft, Delft, The Netherlands, 1993.

[137] G. von Békésy. *Experiments in Hearing.* Robert E. Krieger Publishing Company, Huntington, New York, reprint edition, 1980.

[138] D. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis - Principles, Algorithms, and Applications.* Wiley/IEEE Press, New Yory, NY, USA, 2006.

[139] W. Wang, editor. *Machine Audition: Principles, Algorithms and Systems.* IGI Global, Hershey, PA, USA, 2011.

[140] B. T. West, K. B. Welsh, and A. T. Gałlecki. *Linear Mixed Models - A Practical Guide Using Statistical Software.* CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, second edition, 2015.

[141] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. ii: Psy-

chophysical validation. *Journal of the Acoustical Society Of America*, 85(2):868–878, February 1989.

[142] H. Wittek, F. Rumsey, and G. Theile. Perceptual enhancement of wavefield synthesis by stereophonic means. *Journal of the Audio Engineering Society*, 55(9):723–751, September 2007.

[143] S. Yairi, Y. Iwaya, and Y. Suzuki. Influence of large system latency of virtual auditory display on behavior of head movement in sound localization task. *Acta Acustica united with Acustica*, 94:1016–1023, 2008.

[144] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13:863–882, 2001.

[145] S. Zieliński, F. Rumsey, and S. Bech. On some biases encountered in modern audio quality listening tests - a review. *Journal of the Audio Engineering Society*, 56(6):427–451, June 2008.

[146] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society Of America*, 33(2):248, February 1961.

[147] E. Zwicker. *Psychoakustik*. Springer, Berlin Heidelberg, Germany, 1982.

Curriculum Vitae

Lukas Rohr was born in Basel, Switzerland, in 1985. He studied electronic engineering at Ecole Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, Switzerland and received a M.Sc. degree in 2010. The subject of his Master thesis was the localisation of early reflections in room acoustics by means of active acoustic goniometry.

In October 2010, he enrolled as PhD student at EPFL, first at the Laboratory of Electromagnetics and Acoustics (LEMA) and now at the Signal Processing Laboratory 2 (LTS2). His work mainly focusses on the audibility of artifacts from audio source separation in the context of 3D Wave Field Synthesis. Other current fields of interest include sound localisation, auditory perception, psychoacoustic modelling and signal processing.

List of representative publications

E. Corteel, L. Rohr, X. Falourd, K.-V. Nguyen, and H. Lissek. A practical formulation of 3 dimensional sound reproduction using wave field synthesis. In *International Conference on Spatial Audio 2011*, Detmold, Germany, 2011.

E. Corteel, L. Rohr, X. Falourd, K.-V. NGuyen, and H. Lissek. Practical 3 dimensional sound reproduction using wave field synthesis, theory and perceptual validation. In *Acoustics 2012*, Nantes, France, April 23-27 2012.

X. Falourd, L. Rohr, M. Rossi, and H. Lissek. Spatial echogram analysis of a small auditorium with observations on the dispersion of early reflections. In *39th International Congress on Noise Control Engineering 2010 (INTER-NOISE 2010)*, volume 1, pages 505–513, Lisbon, Portugal, June 15-16 2010.

L. Rohr, E. Corteel, K.-V. NGuyen, and H. Lissek. Vertical localization performance in a practical 3-D WFS formulation. *Journal of the Audio Engineering Society*, 61(12):1001–1014, December 2013.

L. Rohr, X. Falourd, M. Rossi, and H. Lissek. Caractérisation acoustique spatiale des salles : étude des premières réflexions. In *10ème Congrès Français d'Acoustique*, Lyon, France, April 12-16 2010.

Lausanne, May 26, 2015