Individualisation of spectral cues for applications in virtual auditory space: study of inter-subject differences in head-related transfer functions using perceptual judgements from listening tests

David Schönstein

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

SMAER

Présentée par

David Schönstein

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

INDIVIDUALISATION OF SPECTRAL CUES FOR APPLICATIONS IN VIRTUAL AUDITORY SPACE: STUDY OF INTER-SUBJECT DIFFERENCES IN HEAD-RELATED TRANSFER FUNCTIONS USING PERCEPTUAL JUDGEMENTS FROM LISTENING TESTS

L'INDIVIDUALISATION DES INDICES SPECTRAUX POUR LA SPATIALISATION ACOUSTIQUE : ÉTUDE PERCEPTIVE DE LA VARIABILITÉ INTER-INDIVIDUELLE DANS LES FONCTIONS DE TRANSFERT RELATIVES À LA TÊTE

soutenue le 12 septembre 2012

devant le jury composé de :

Sabine Meunier	Chargé de recherche CNRS, rapporteur
Gaël Richard	Chargé de recherche CNRS, rapporteur
Brian Katz	Chargé de recherche CNRS, HDR, directeur de thèse
Christophe d'Alessandro	Directeur de Recherche CNRS, co-directeur de thèse
Jean-Luc Zarader	Professeur UPMC, examinateur
Jean-Michel Raczinski	Alliance Manager Arkamys, examinateur

ABSTRACT

The human auditory system has evolved to seamlessly interpret sounds from our environment. As the mechanisms underpinning our ability to accurately determine the nature and location of sound sources in space have become better understood, we have been able to apply this knowledge to technologies that are of benefit to us. One of these technologies is the ability to generate compelling auditory illusions, or a Virtual Auditory Space (VAS). Despite a large emphasis in the past on virtual reality for vision, VAS and its effectiveness in generating a compelling illusion, is now a field of research that is growing along with the many applications of the technology.

The theory behind how the illusion of a virtual sound source is created is in essence quite simple, and determined by the fact that the auditory system has all the information required to correctly interpret a sound source in space embedded in the sound waves that reach the auditory periphery at the level of the tympanic membrane inside the ear canal. Any number of different and competing sound sources can be in a sense smeared onto the basilar membrane and the auditory system is able to extract the most salient cues at the time in order to stream and interpret the auditory scene, as detailed in chapter 2. Thus, in order to provide a compelling illusion of sound sources in space all one has to do is reproduce the exact same signal to the auditory system as if sounds were originating from the desired location. This can be achieved by either using speakers or headphones; the latter being the focus of this body of work. One can imagine the simplest case in which small microphones are placed inside the ear canal and record the signal from a sound source in space, and then this recorded signal is played back to the same listener over headphones, adjusted to the correct level and placed at the exact location of where the microphones were. In this scenario the listener would perceive the sound source as if it originated at the location at a distance, instead of originating from the headphones, as the exact same acoustic information is being presented to the auditory system. The task of rendering any monophonic signal in VAS, known as binaural synthesis, is more complicated as detailed in chapter 3.

The applications of binaural synthesis are far reaching given that it allows any signal to be virtually positioned in space when presented to the listener over headphones. One helpful use of binaural synthesis is for people who are hearing impaired and require hearing aids. The hearing impaired have difficulty in environments where many people are talking at the same time. A standard hearing aid will simply amplify all talkers and not provide information on the sound source location, which becomes almost impossible for the listener to understand since one of the most important ways the auditory system streams different sound sources, and is subsequently able to focus attention on one talker for example, is by using the sound source location as a cue. A hearing aid that can produce a binaural synthesis, providing cues

to sound source location, along with an amplification of the signal, could allow for improved intelligibility in multi-talker environments. There remain however limitations to this technology for the hearing impaired as aids are often characterised by a reduced sensitivity for frequencies that are crucial for locating sound sources.

Another application of binaural synthesis, geared towards the general public, is the generation of immersive auditory experiences for movies and video games. This is akin to three-dimensional visual effects being used in cinemas today and acts as an augmentation of the different senses in order to make the experience more realistic and emotionally engaging. The technology of binaural synthesis has also been adopted by pilots of aeroplanes as an additional aid for negotiating the large amount of feedback from the console during a flight. A virtual sound source can be an indicator of elevation or pitch of an aeroplane as the sound is displaced above or below the listener's head.

While the many applications of binaural synthesis are still emerging, studies are still exploring the perceptually salient cues used by the auditory system in order to effectively interpret an auditory scene. A better understanding of how these mechanisms work will enable an improved experience for listeners in terms of producing a high fidelity rendering in VAS. One of the aspects of how sounds are perceived by the auditory system that researchers have isolated as being a critical component to the production of a realistic binaural synthesis, is the way in which sound sources are filtered by the human head and body, and in particular the outer ear, known as the Head-Related Transfer Function (HRTF). Embedded in the acoustic filtering effects of a listener's morphology, represented by the HRTF, are cues that help resolve any ambiguity relating to the nature and location of a sound source in space, as detailed in chapter 4.

The HRTF has been shown to be linked to the asymmetric shape of the ear, or pinna, in ways that produce acoustic signatures to sound source location. Given that the shape of the ear varies from listener to listener, the way in which our auditory system is tuned to HRTFs also varies from listener to listener. These differences necessitate the use of what is known as individualised HRTFs; the HRTFs themselves are specific to the morphology of the listener (see chapter 5). Here in lies the difficulty in providing binaural synthesis as a technology to the masses; the task of acquiring HRTFs for a particular listener is in fact expensive and laborious. Whilst there are other factors that have an impact on the fidelity of a rendering in VAS, such as the headphones used and compensating for the way in which the headphone impacts the HRTF (a study performed in chapter 6), the problem of providing some form of individualised HRTF for binaural synthesis stands as the major roadblock to making the technology more amenable to the consumer market. Arkamys, the company that has funded this research, was drawn to such a personalised audio experience in their pursuit to provide the highest quality digital audio solutions in the consumer electronics industry.

The pursuit of an elegant solution to HRTF individualisation has been evolving for some time with a range of different approaches available. One

difficulty in finding the best solution, is the question of what constitutes an effective rendering in VAS and how to measure and quantify it? The majority of studies have looked at whether listeners are able to accurately determine the position of virtual sound sources, also known as localisation accuracy, as a measure of the quality of a binaural synthesis. The quantification of localisation accuracy is done by measuring the disparity between the perceived and true location of a virtual sound source. This disparity measure was used in the current body of work in chapter 6 for testing different types of headphones. In reality, localisation accuracy is only one aspect of what makes an effective binaural synthesis, and measuring localisation errors only one way of quantifying the effectiveness. Qualities such as the realism and naturalness of the illusion, the coherence of the auditory image, and whether the sound is perceived at the desired distance or inside the head, are as important as localisation, depending on the application of the technology. The concept of measuring the quality of a binaural synthesis, and quantifying it using metrics other than localisation errors, was explored in chapter 7 via the use of listening tests. The ability of subjects to be able to make consistent perceptual judgements of a binaural synthesis using different HRTFs and listening tests was studied and considered an important precursor to any conclusions drawn with respect to a particular solution to the HRTF individualisation issue.

Once the perceptual limits of subjects in this context was understood to some degree, a study was performed, as described in chapter 8, in order to build on what we know about the most perceptually salient components of the HRTF. An important aspect of this study was to provide a method for determining the best way to describe the differences in HRTFs between listeners based on a perceptual validation. In being able to characterise the perceptual differences between HRTFs, we are a step closer to understanding what it is about a rendering in VAS that makes it suitable for a particular listener, and providing an individualised or personalised HRTF. Of the many different approaches to producing individualised HRTFs that are explored in the literature, the current body of work focused on a technique in which an HRTF was selected for a particular listener from a large database of HRTFs obtained for many different listeners. In order to perform this HRTF selection and avoid the costly procedure of obtaining HRTFs from the listener directly, the link between the subjects' morphology and HRTFs in the database was tested using dimensions of the listener's ear, or pinna, as detailed in chapter 9.

RÉSUMÉ

Le système auditif humain a évolué de façon à pourvoir interpréter efficacement les sons de notre environnement. Grâce une meilleure compréhension de notre capacité à déterminer avec précision la nature et la position des sources sonores, nous sommes en mesure d'appliquer ces connaissances à des technologies bénéfiques pour l'homme. L'une de ces technologies est la capacité à générer des illusions auditives convaincantes, ou sources sonores

virtuelles. Malgré une importante concentration de la recherche sur les illusions virtuelles dans le domaine de la vision, les espaces acoustiques virtuels, ou *Virtual Auditory Space* (VAS) en anglais, sont désormais au cœur des recherches actuelles et ce nouveau domaine développe de nombreuses applications en lien avec cette technologie.

La théorie derrière la manière dont se crée l'illusion d'une source sonore virtuelle est en réalité assez simple. L'illusion se crée du fait que le système auditif dispose de toutes les informations nécessaires pour interpréter correctement une source sonore dans l'espace, contenues dans les ondes sonores présentes au niveau de la membrane du tympan, à l'intérieur du conduit auditif. N'importe quel nombre de sources sonores différentes et concurrentes peuvent être dans un sens plaquées contre la membrane basilaire, et le système auditif est alors capable d'en extraire les signaux les plus saillants à chaque moment afin de diffuser et interpréter la scène sonore, comme l'explique le chapitre 2. Ainsi, afin de fournir une illusion convaincante de source sonore dans l'espace, il suffit de reproduire un signal sonore de façon à ce que le système auditif l'interprète comme s'il provenait réellement de l'endroit désiré. Ceci peut être réalisé en utilisant des haut-parleurs ou un casque audio ; la transmission via casque, plus précisément, sera l'objet de la recherche présentée ici. Il est facile d'imaginer une expérience assez simple où de minuscules microphones seraient placés à l'intérieur du conduit auditif afin d'enregistrer le signal d'une source sonore réelle. Le même signal serait ensuite transmis au même sujet via le casque audio, correctement ajusté au bon niveau et placé à l'endroit exact où étaient les microphones. Dans ce scénario, le sujet percevrait le signal sonore comme s'il émanait d'une source située à une certaine distance et non pas provenant du casque audio en lui même puisque la même information sonore est présentée au système auditif. La tâche de transformer un signal monophonique en VAS, connue sous le nom de synthèse binaurale, est plus compliquée et est expliquée en détail dans le chapitre 3.

On trouve de nombreuses applications de la synthèse binaurale liées au fait qu'elle permet de présenter à un auditeur, via une écoute au casque, n'importe quel signal virtuellement positionné dans l'espace. Cette application se révèle notamment très utile pour les personnes souffrant d'une déficience auditive nécessitant une prothèse afin de pallier aux difficultés rencontrées dans des environnements où plusieurs personnes parlent en même temps. Une prothèse auditive standard se contente d'amplifier les sons émis par tous les locuteurs et ne fourni pas d'indications quant à la localisation de la source sonore, rendant quasiment impossible pour l'auditeur de savoir d'où provient le bruit. Or le système auditif fonctionne de telle façon que la localisation de la source sonore est utilisée comme repère afin de séparer les différentes sources sonores et se concentrer, par exemple, sur une personne qui parle. Une prothèse auditive utilisant la synthèse binaurale qui pourrait procurer une information sur la position de la source sonore, tout en amplifiant les signaux sonores, pourrait alors améliorer l'intelligibilité de façon significative dans les environnements comprenant plusieurs locuteurs. Cette technologie a néanmoins ses limites puisque les prothèses utilisées par les malentendants sont souvent caractérisées par une sensibilité réduite aux fréquences qui sont cruciales pour la localisation des sources sonores.

Une autre application de la synthèse binaurale, orientée vers le grand public cette fois, est la possibilité d'une immersion auditive totale dans les films ou jeux vidéo. Ceci s'apparente aux effets tridimensionnels utilisés au cinéma aujourd'hui afin de rendre l'expérience cinématographique plus réaliste et plus touchante émotionnellement pour le public. La technologie de la synthèse binaurale a également été adoptée par les pilotes d'avion comme une aide supplémentaire permettant de mieux interpréter et gérer la quantité d'informations transmises par la console lors d'un vol. Une source sonore virtuelle peut être alors un indicateur de variation d'altitude ou de tangage d'un avion en générant une source sonore virtuelle de façon à ce qu'elle soit perçue comme au dessus ou en dessous de l'auditeur.

Alors que les nombreuses applications de la synthèse binaurale sont en encore un domaine en pleine émergence, des études explorent actuellement les indices perceptivement saillants utilisé par le système auditif afin d'interpréter la source sonore. Une meilleure compréhension du fonctionnement de ces mécanismes permettrait une amélioration d'écoute pour les auditeurs en terme de production d'un rendu de grande fidélité en VAS. Un des aspects de la perception des sons par le système auditif, identifié comme essentiel à la synthèse binaurale par les chercheurs, est la manière dont les sources sonores sont filtrées par la tête et le corps humain, notamment l'oreille externe, connue sous le nom de fonctions de transfert acoustique ou *Head-Related Transfer Function* (HRTF) en anglais. On trouve des indices concernant la nature et la position d'une source sonore dans l'espace, représentés par les HRTFs, dans les effets acoustiques de filtrage liée à la morphologie de l'auditeur, comme l'explique en détails le chapitre 4.

Il a été prouvé que les HRTFs sont liées à la forme asymétrique de l'oreille, ou du pavillon, de façon à produire des signatures acoustiques de la position de la source sonore. Etant donné que la morphologie de l'oreille varie d'un auditeur à un autre, la façon dont notre système auditif s'accorde avec les HRTFs varie elle aussi. Ces différences nécessitent l'utilisation de ce qu'on appelle des HRTFs individualisés puisque les HRTFs sont spécifiques à la morphologie de l'auditeur comme le détaille le chapitre 5. Ceci représente le plus grand défi rencontré par l'application de la synthèse binaurale pour une écoute destinée au grand public, la tâche de produire un HRTF pour un auditeur spécifique étant coûteuse et laborieuse. De nombreux facteurs influencent la qualité d'un rendu en VAS, et notamment le type de casque utilisé et le fait qu'il faut alors compenser son influence sur l'HRTF; une étude sur ce sujet est présentée dans le chapitre 6. La difficulté de produire un HRTF individualisé pour la synthèse binaurale est un obstacle majeur à la diffusion de cette technologie sur le marché des consommateurs. La société qui finance cette recherche, Arkamys, a été attirée par cette idée d'une écoute personnalisée au cours de leurs recherches pour une meilleure qualité audio de la technologie numérique qui serait destinée à l'industrie électronique grand public.

La recherche d'une solution élégante à l'individualisation de l'HRTF se poursuit depuis un certain temps et selon tout un éventail d'approches différentes. Une des difficultés à trouver cette solution est la question relative à ce que constitue un rendu efficace en VAS, et comment le mesurer et le quantifier. La majorité des études se sont penchées sur la capacité des auditeurs à déterminer avec précision la position des sources sonores virtuelles, aussi appelé précision de localisation, comme une mesure de qualité de la synthèse binaurale. Quantifier la précision de la localisation revient à mesurer l'écart entre la position perçue et la position réelle d'une source sonore virtuelle. Cette mesure de différence entre position perçue et position réelle a été utilisée au cours de cette étude pour tester différents types de casque, comme le montre le chapitre 6. Il est important de noter que la précision de la localisation n'est qu'un des aspects de ce qui rend une synthèse binaurale efficace, et les erreurs de localisation mesurées ne sont qu'une des façons de quantifier sa précision. D'autres qualités, comme le réalisme et la nature de l'illusion, la cohérence de l'image auditive ou si le son est perçu comme provenant de l'endroit souhaité ou perçu comme provenant de l'intérieur de la tête, sont tout aussi importantes que la localisation ; leur importance dépendant notamment de l'application qui est faite de cette technologie. Les façons de mesurer la qualité d'une synthèse binaurale et de la quantifier en utilisant d'autres mesures que l'erreur de localisation sont explorées dans le chapitre 7, grâce à des tests d'écoute. La capacité des sujets à faire des jugements perceptifs consistants d'une synthèse binaurale utilisant des HRTFs différents ainsi que des tests d'écoute est étudiée dans cette recherche et constitue un précurseur de possibles solutions concernant les problèmes de l'individualisation de l'HRTF.

Après avoir compris dans une certaine mesure les limites perceptives des sujets dans ce domaine, une étude a été réalisée, et décrite dans le chapitre 8, avec dans l'idée d'utiliser ce que nous savons concernant les indices perçus comme les plus saillants de l'HRTF. Un aspect important de cette étude a été de créer une méthode, fondée sur une validation perceptive, pour déterminer la meilleure façon de décrire les différences entre HRTFs selon les auditeurs. En étant capable de caractériser les différences perçues entre HRTFs, c'est un pas de plus vers la compréhension de ce qui fait qu'un rendu en VAS soit adapté à un auditeur spécifique, et être capable de fournir un HRTF personnalisé ou individualisé. Parmi les différentes approches utilisées pour produire des HRTFs individualisés décrites dans la littérature sur le sujet, cette étude se concentre sur la technique de sélection d'un HRTF pour un auditeur spécifique parmi une large base de données de HRTFs, obtenues auprès de différents auditeurs. Afin d'effectuer la sélection pour un auditeur spécifique et éviter la production coûteuse et laborieuse d'obtenir un HRTF de l'auditeur lui-même, le lien entre morphologie et HRTF dans la base de données est testé en utilisant les dimensions de l'oreille ou du pavillon, comme détaillée dans le chapitre 9.

LIST OF PAPERS, PATENTS, AND APPEARANCES IN THE MEDIA, RELATING TO RESEARCH

Research from this body of work that has:

- been submitted to academic journals:

D. Schönstein and B. F. G. Katz. (accepted with revision). Variability in Perceptual Evaluation of HRTFs. Journal of the Audio Engineering Society, 2011.

D. Schönstein and B. F. G. Katz. (submitted). Analysis of HRTF inter-subject differences applied to HRTF selection from a database using morphological parameters. Journal of the Acoustical Society of America, 2011.

- been presented at conferences with articles:

D. Schönstein and B. F. G. Katz. HRTF selection for binaural synthesis from a database using morphological parameters. In Proceedings of 20th International Congress on Acoustics, Sydney, Australia, August 2010.

D. Schönstein and B. F. G. Katz. Variability in perceptual evaluation of HRTFs. In Proceedings of the 128th Convention of the Audio Engineering Society, London, England, May 2010.

D. Schönstein and B. F. G. Katz. Sélection de HRTF dans une base de données en utilisant des paramètres morphologiques pour la synthèse binaurale. In Proceedings of the 10th Congrès Français d'Acoustique, number 431, Lyon, France, April 2010.

D. Schönstein. Méthodes pour l'adaptation individuelle de l'HRTF pour le rendu via le synthèse binaurale. In 5ème Journées Jeunes Chercheurs en Audition, Acoustique Musicale et Signale Audio, Marseille, France, November 2009.

D. Schönstein, B. F. G. Katz, and L. Ferré. Comparison of headphones and equalization for virtual auditory source localization. In Proceedings of the ASA and EAA Joint Conference on Acoustics, pages 4617-4622, Paris, France, June 2008.

- been included in a published patent:

B. F. G. Katz and D. Schönstein. Procédé de sélection de filtres HRTF perceptivement optimale dans une base de données à partir de paramètres morphologiques, 2011. Patent No. PCT/FR2011/0508405.

- appeared as newspaper articles:

V. Shannon. Taking sound to a new level. The New York Times, 19 March 2008. viewed 2 May 2012. http://www.nytimes.com/2008/03/19/technology/19iht-ptend20.1.11248265.html?_r=1

D. S. Mis. Des oreilles pour tous, un son pour chacun. Figaro, 6 February 2008. viewed 2 May 2012. http://www.lefigaro.fr/hightech/2008/02/06/01007-20080206ARTFIG00396-des-oreilles-pour-tous-un-son-pour-chacun.php

- appeared as a television show:

Bose, Arkamys: au coeur du son... . video recording. Plein Écran, TFI, Paris, 13 April 2008. viewed 2 May 2012. http://videos.tf1.fr/infos/plein-ecran/plein-ecran-avril-2008-bose-arkamys-coeur-4375180.html

ACKNOWLEDGMENTS

I would like to thank the company Arkamys for its financing of the Ph.D. as part of the CIFRE industry-linked research program. In particular, I would like to acknowledge the involvement of Jean-Michel and Philippe.

To my supervisors Brian and Christophe, thank you for your guidance and hard work over the years at LIMSI and from abroad.

To Simon from The University of Sydney, thank you for your ongoing support from the beginnings of my life as a researcher up until now.

Thanks Zane and Brendan for tackling some of the editing.

To my family, Eva, Peter, and Lisa, thank you for your patience and kindness over the years.

And finally, to my partner Antonia, and son Julian, thank you for your big hearts and tolerance throughout.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

HRTF	Head-Related Transfer Function
VAS	Virtual Auditory Space
ITD	Interaural Time Difference
ILD	Interaural Level Difference
HRIR	Head-Related Impulse Response
HpTF	Headphone Transfer Function

MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
ISSD	Inter-Subject Spectral Difference
PCA	Principal Component Analysis
BEM	Boundary Element Method
MaxIACC	Maximum Inter-Aural Cross-Correlation
OC	Open Circumaural
CC	Closed Circumaural
TI	Tube Intra-aural
BC	Bone Conduction
FR	Frequency Response
SCC	Spherical Correlation Coefficient
CCC	Circular Correlation Coefficient
CC	Correlation Coefficient
LSE	LIMSI Spatialization Engine
GUI	Graphical User Interface
ANOVA	ANalysis Of VAriance
DTF	Directional Transfer Function
MS	Multidimensional Space
PC	Principal Component
FSF	Frequency Scale Factor
SVM	Support Vector Machine

Part I

BACKGROUND AND LITERATURE REVIEW

1 GENERAL INTRODUCTION

The current body of work aims at addressing the challenges to producing a compelling rendering of sound sources in Virtual Auditory Space (VAS) for the listener via the use of personalised Head-Related Transfer Functions (HRTFs). The work provides a potential solution to the laborious and expensive task of recording individualised HRTFs so that the technology becomes more amenable to the consumer market. The research has been broken up into four chapters:

1. *Role of headphones in binaural synthesis* (chapter 6) – assessment of the effect of the hardware used in producing binaural synthesis, namely the type of headphones used, and the effectiveness of a headphone equalisation, via a localisation task

2. *Perceptual judgements of HRTFs using listening tests* (chapter 7) – an iterative approach to developing a listening test that can be used to measure the effectiveness of binaural synthesis for different HRTFs, with an emphasis on reducing the variability in subject responses and producing results with a high degree of repeatability

3. *Salient spectral cues for binaural synthesis* (chapter 8) – analysis of the most perceptually relevant spectral features in the HRTF using results from a listening test (from chapter 7) for a large number of subjects along with the subjects' corresponding measured HRTFs

4. *Significant morphological parameters for binaural synthesis* (chapter 9) – validation of a method for predicting an optimal HRTF for a particular listener from a database, based on the results from chapter 8, using morphological parameters from the listener, along with an analysis of the significance of the different morphological parameters used

Chapter 6 explores an important component of binaural synthesis that is often overlooked; how the headphones used can determine the quality and realism of the rendering of sound sources in virtual auditory space. The following three chapters 7, 8, and 9, form part of a process that aims to be able to select an optimal HRTF from a database for a particular listener. Before developing this selection procedure, it was critical to have subjects be able to reliably evaluate the different HRTFs in the database in order to determine whether a selected HRTF was effective or not. To this end, the listening test developed in chapter 7 established a method for minimising the variability in subject responses. The study in chapter 8, whilst not using the optimal listening test design from chapter 7, involved an analysis establishing the most effective way to describe inter-subject differences based on measured HRTFs. In the process of validating different methods for calculating inter-subject differences, insights into which components of the HRTF were the most perceptually salient were developed. These insights were considered relevant to the application of binaural synthesis to the consumer market given that they were grounded in perceptual judgements from a listening test assessing the *quality* of renderings in VAS. The most effective method for evaluating HRTF inter-subject differences was then used in chapter 9 in order to create a model and a predictive procedure for selecting a personalised HRTF for a listener based on their morphology. Morphological dimensions of the outer ear and body were used from the same database from which the measured HRTFs were taken. This type of predictive procedure has significant implications for applications of binaural synthesis in the consumer market given that it bypasses the laborious and expensive task of recording individualised HRTFs whilst ensuring a realistic rendering of sound sources in VAS.

With respect to the individual research chapters, beginning with chapter 6, an analysis of the effect of headphone type on binaural synthesis was performed and considered important for potential applications of the technology in the consumer market given that different headphones vary in their ability to produce a signal at all audible frequencies. The headphone types tested ranged from circumaural, to intra-aural, and even bone conduction headphones, which have different mechanisms for producing the signal at the listener's ears. In addition, different headphones can add colouration to the signal that might be to the detriment of the quality of the rendering in VAS. It is for this reason that a headphone equalisation was also tested.

A localisation task was used for this study in order to be able to compare results with the large number of studies in the literature using virtual sound sources. This allowed for a quantitative assessment of the effectiveness of each headphone (with and without equalisation) to render a virtual sound source so that it was perceived at the target location in space.

The move from using a localisation task to using a listening test when assessing the quality of a binaural synthesis was made in chapter 7 in order to focus on aspects of a rendering that were particularly relevant to the application of the technology in the consumer market. Perceptual judgements for a number of different attributes of a rendering was considered more relevant localisation accuracy in this context. The iterative approach to designing the listening test used in the study was orientated towards questions relating to whether virtual auditory images appeared as being coherent or whether the rendering in general was realistic for the listener. A number of different listening test designs were used, each time aiming to eliminate biases that might cause variability in subject perceptual judgements across replicates of the listening test. One of the most significant changes made in this iterative process was a reduction in the number of different HRTFs being tested, and the use of specific attributes of the renderings rather a global judgement. The goal was to reach a point in the iterative process at which an acceptable degree of repeatability was observed among the subjects tested. Once this was achieved it would then be possible to use the responses for further analysis (i.e. for testing a method for selecting a personalised HRTF).

The following phase of the research presented in this body of work involved both chapters 8 and 9. Both chapters relate to the design and validation of a selection procedure that predicts an optimal HRTF from a database for the listener so that any rendering in VAS will be personalised. A range of solutions to the problem of HRTF personalisation have been proposed in the literature, yet there is often a lack of a conclusive perceptual validation for the proposed techniques. This was one of the main motivations for the studies presented in these two chapters, in which a large number of subjects were tested in order to produce a statistically significant result.

Chapter 8 focuses mainly on using a number of different methods for describing HRTF inter-subject differences. Each method draws on a different aspect of the HRTF in terms of features in the spectrum that are important for a compelling rendering in VAS. The methods use either specific features, such as the notches that appear in the spectrum of the HRTF, or took into account all features in the spectrum. The inter-subject differences were used to produce a multidimensional space in which all the subjects were represented; each subject corresponded to an HRTF from the database, and the closer the subjects were in the space to each other the more similar their HRTFs. Each of the methods, and corresponding multidimensional spaces, were validated using the perceptual judgements from a listening test.

An important factor in the analysis was the large number of subjects that participated in the listening test. Given that the study from chapter 7 showed that there is a high degree of variability in perceptual judgements for renderings in VAS using different HRTFs, it was important to have a large number of responses and an analysis that would be robust to ’noise’ in the results from the listening test (i.e. the assumed differences in perceptual judgements of the subjects if the listening test had been repeated).

The effectiveness of each of the multidimensional spaces in describing the perceptual judgements from the listening test was an indicator of the significance of the HRTF features used in the analysis. This allowed for insights into the role and salience of some of the most commonly studied components of the HRTF. In addition to these insights, the statistical analysis enabled for an understanding of the most perceptually relevant frequency ranges of the spectrum, based on a method of describing HRTF inter-subject differences that used all features in the spectrum.

The study in chapter 9 used the findings from the previous chapter. In particular, the most effective method for describing inter-subject differences and corresponding multidimensional space was selected and then used as the basis for a predictive HRTF selection procedure. The approach took any subject’s morphology dimensions in order to position them into the multidimensional space. Once a subject was placed in the space, the optimal HRTF was selected as the closest to that subject. The effectiveness of this process was tested by calculating the distances between HRTFs in the space for each subject in the database and comparing these distances to the perceptual judgements from the listening test. In this way, the process could be refined by selecting different combinations of morphological parameters until the most effective subset was chosen. The predictive process of selecting an HRTF from a database for a listener based on morphology was then shown to be statistically effective.

The mentioned subset of morphological parameters provided a further insight into what might be the most perceptually significant dimensions of a listener's ear and body, in terms of their filtering effects in the HRTF. In essence, the analysis aimed to find the link between measured morphological parameters of a subject and HRTFs judged by the subject to be effective in VAS. This enabled a better understand how the body interacts with incident waves from a sound source in space. The significance of the listener morphology was further developed by using only the listening test results and the subject morphological parameters (i.e. no HRTF data was included) and incorporating various machine learning algorithms. Decision trees were built using the perceptual judgements, and the morphological parameters were also ranked based on their relative importance as features in algorithm using support vector machines.

The final analysis of chapter 9 involved the use of HRTFs recorded on a mannequin head with replica pinnae made of rubber, cast from molds of real ears. HRTF recordings were made for a number of modifications applied to the left and right ear replica pinna, such as completely filling a cavity with clay. The differences in the recordings across the different pinna modifications for a large number of locations in space showed how each morphological parameter affected the HRTF. This in turn allowed for insights into the significance of the different modified parameters, given our knowledge of the role of the different spectral features in the HRTF from chapter 8 and other studies. The different pinna modifications were also mapped into multidimensional spaces in order to quantify their differences and understand how they compared to each other.

2 HUMAN AUDITORY PERCEPTION

This first chapter presents the fundamentals of how humans perceive their acoustic environment, and will act as a foundation to understanding the concepts of spatial audio detailed in this body of work. Our understanding of how the human auditory system works begins with an evolutionary perspective. The ability of humans to encode mechanical disturbances in the medium in which they live, what we would call a *sound*, appears to have evolved to promote survival (Erulkar, 1972). It would be advantageous for a species, for example, to determine the location of predator's approach or discern the call of a mate from a distance. To this end the auditory system would need to process sensory information in terms of the 'what' and 'where' of a sound; our current understanding is that of a dual-pathway model for encoding these two aspects with a binding of the two paths in the auditory cortex (see for example Arnott et al., 2004).

Despite the fact that the visual system is a more developed and utilised sense in humans (in fact auditory perception has a heavy reliance on visual cues), the auditory system does a remarkably good job in terms of the perception of sounds in its own right. The human ear contains only a single receptive organ called the basilar membrane on which some 30,000 sensory cells, called hair cells, are responsible for registering the mix of all the different frequencies that make up our auditory environment. This is an impressive feat due to the fact that at any one time there might be a combination of many sounds across different frequencies encoded on the basilar membrane and the auditory system is able to seamlessly interpret what we might call the different auditory objects around us (see Griffiths and Warren, 2004).

Humans have evolved to process a broad range of frequencies between 20 Hz and 20 kHz. The level sensitivity of the human auditory system is so great that if it was any better we would hear the blood flow in our veins. We have highly specialised regions of the brain for detecting the 'what' of sounds, such as for speech, and are capable of determining with much precision the 'where', or location, of sound sources without vision. It is the 'where' component of auditory processing that will be the focus of this body of work, and in particular how evolution has made use of the fact that we have two ears (Schnupp and Carr, 2009). The work endeavours to evaluate how the auditory system is able to perceive sounds in space and what aspects of the sounds might be important for effectively performing the task. Whilst explanations are mainly grounded in physiological and physical terms, it is important to note that the perception of the location of sound sources can be quite psychophysical as well (see for example the ventriloquist effect; Alais and Burr, 2004).

2.1 COORDINATE SYSTEM

In order to accurately describe the 'where' of sound sources in space, a coordinate system must be implemented for the purposes of this research. In the current body of work, and in most studies, two coordinate systems will be used based on different poles. Figure 1 shows the two coordinate systems with the listener represented at the centre of an imaginary sphere. Figure 1(a) shows what is what is known as the hoop coordinate system, in which the azimuth and elevation coordinates correspond to the standard single pole coordinate system where $0\,^\circ$ azimuth and $0\,^\circ$ elevation is directly ahead of the listener and positions to the right and up are positive (ranging from $0\,^\circ$ to $180\,^\circ$ and $0\,^\circ$ to $90\,^\circ$ respectively) and positions to the left and down are negative (ranging from $0\,^\circ$ to $-180\,^\circ$ and $0\,^\circ$ to $-90\,^\circ$ respectively).

Figure 1(b) shows what is known as a lateral/polar coordinate system, in which there is a single pole passing through the two ears. The lateral angle is the horizontal angle away from the midline, which is the vertical plane separating the left and right hemispheres (i.e. the plane represented by the purple circle in figure 2). Lateral angles to the left and right of the listener range from $0\,^\circ$ to $-90\,^\circ$ and $0\,^\circ$ to $90\,^\circ$ respectively. For example, a lateral angle of $20\,^\circ$ will describe all positions, at a fixed distance, on a vertical circle that subtend that angle from the median plane, whether it be in front or behind of the listener. The polar angle is the angle around the interaural axis (the axis connecting the left and right ear – represented by the green line in figure 2); this is the angle away from the horizontal plane on the described circle. Polar angles range from $0\,^\circ$ to $180\,^\circ$ from in front of the listener to behind in the upwards direction, and from $0\,^\circ$ to $-180\,^\circ$ from in front to behind in the downwards direction. The lateral/polar coordinate system is a particularly intuitive one due to the fact that it mirrors how the auditory system determines sound source location (see section 2.2.2).

2.2 AUDITORY CUES TO SOUND LOCATION

The human auditory system has adapted to make use of the most salient acoustic cues from sound sources in our environment. Unlike other spatial senses such as vision where there is a topographic projection from receptor epiphelia into the central nervous system, the auditory system encodes the amplitude of the energy entering the ears as a function of frequency. Differences between the energy at the two ears is translated into information about the sound source's location. As explained by Blauert (1997)

> ... the system does not use every detail of the complicated interaural dissimilarities, but rather derives what information is needed from definite, easily recognisable attributes.

It is also important to note that the auditory system uses a variety of different auditory cues depending on the environment (such as a reverberant room for example), which will be discussed below and detailed in chapter 4.

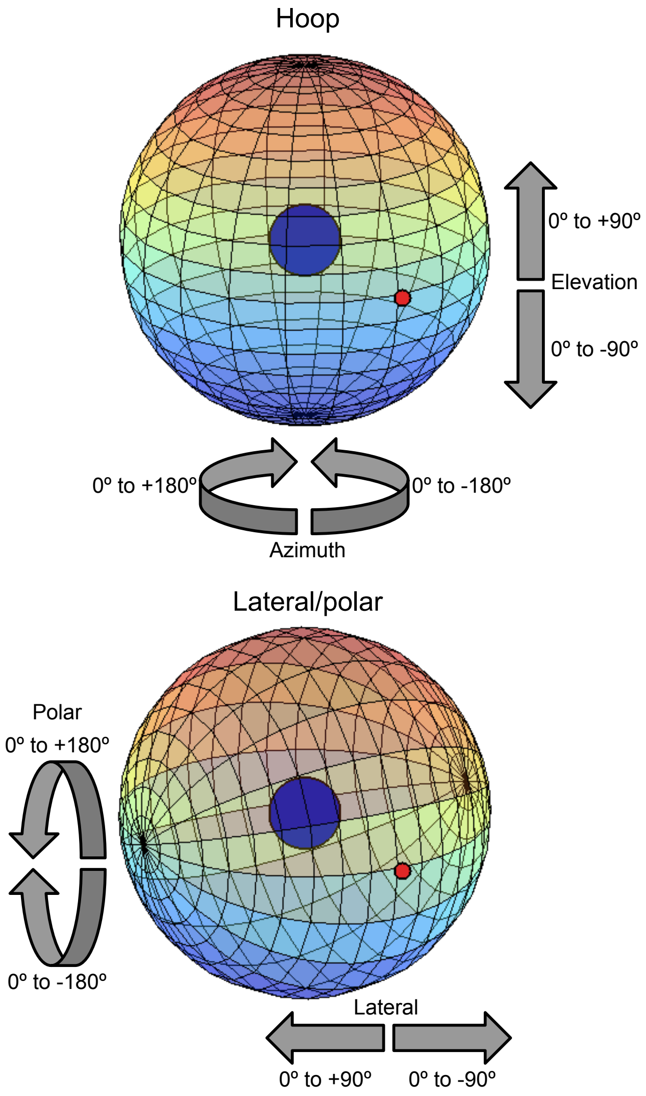


Figure 1: Representation of the two coordinate systems used in the current body of work: (a) hoop coordinate system, and (b) later/polar coordinate system. Red circles represent the position directly in front of the listener, at azimuth 0° and elevation 0°, included in the figure as a reference. The larger blue circle at the centre of the sphere represents the position of the listener's head.

2.2.1 *Dominant interaural cues*

The Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) between the two ears for a single sound source in space are two crucial cues for determining location. ITDs refer to the difference in travel time of incident sound waves between the two ears for sound sources that are not on the midline (i.e. not at a point equidistant from both ears). For example if a sound were to originate to the right of a listener, the incident waves would arrive at the right ear before the left ear. Input from both ears meets at a structure along the auditory pathway known as the superior olivary complex that is sensitive to small time differences. The normal human threshold for detection of an ITD is up to a difference of 10 µs with a relatively large degree of variance between individuals. Experiments conducted using a sphere to model the shape of the head, with a distance of approximately 22-23 cm between the two ears, measured maximum ITDs of approximately 660 µs (Woodworth and Schlosberg, 1965).

Studies measuring the accuracy of listeners' judgements of sound source location, or localisation tasks, using different stimuli suggest that whilst frequency dependent interaural phase differences are detectable for low frequencies (Zwislocki and Feldman, 1956; Palmer and Russell, 1986) subjects are insensitive to them when an ITD is maintained (Kulkarni et al., 1999). Similar tests have shown that ITDs are probably encoded by mostly low-frequency auditory neurons (Middlebrooks and Green, 1990) for frequencies below about 2 kHz (Blauert, 1997). ITDs are also known to dominate ILD cues for broadband stimuli (Wightman and Kistler, 1992).

ILDs are caused by the absorption of energy primarily by the head and also the body for sound sources off the midline, which produces a shadowing of the farthest ear. At low frequencies where the wavelength of sound approaches or is larger than the distance between the listener's ears, the head does not diffract the incident waves and ILDs are quite weak and thus not a salient feature for localisation. Experiments with spherical head models, using sounds with wavelengths much smaller than its diameter (i.e. the distance between the ears), have measured a maximum ILD of 6 dB for a sound source positioned along the interaural axis (Shaw, 1974). The minimum thresholds for ILDs are less than 1 dB (Mills, 1960). The auditory system most probably integrates level differences (and time differences) over discrete frequency channels, using the most salient and easily recognisable features available (Macpherson and Middlebrooks, 2002).

2.2.2 *Spectral cues*

The interaural cues described previously are used by the auditory system to estimate a sound source's position in space. However ITDs and ILDs alone will not provide enough information for localisation in three-dimensional space; this is simply due to the fact that a specific interaural difference can describe any position in space that is equidistant to the listener's ears. Equidistant points to a listener's ears in space can be represented as a plane in the shape

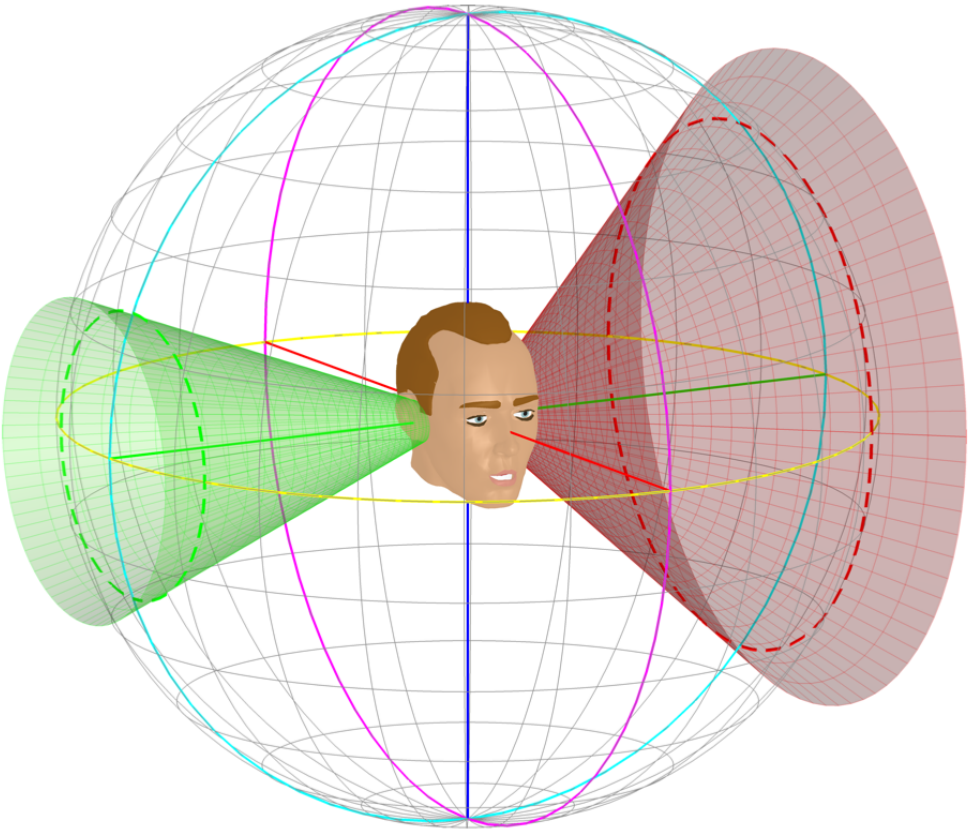


Figure 2: Locations along two example cone of confusions in red and green. All locations along any plane of the two cones of confusion (examples shown by a dotted line) are equidistant from the listener's ears and will have the same ITD. Figure taken from (Guillon, 2009).

of a hyperboloid (a hyperbola rotated around the interaural axis), the so-called cone of confusion, as represented in figure 2 (see Katz et al., 2005, for an analysis of cone of confusion forms).

If any position lying on a cone of confusion produces the same interaural cues, the auditory system is in need of more information in order to resolve the ambiguity. The additional spectral cue used comes from the interaction of sounds with the external ear, namely the pinna. Reflections of sound waves within the folds of the pinna are a perfect candidate for producing the additional spectral information, due to the asymmetrical shape of the ear; the reflections create a filter for sound sources that is dependent on their position in space. This is due to the fact that a change in the reflection path, determined by a change in position of the sound source, will alter the spectral features of the filter. This position dependent information is what codes for the location of a sound source on the cone of confusion (defined by the interaural cues), and is important for determining elevation (up-down) as well as coding for whether the sound is coming from in front or behind the listener (Wightman and Kistler, 1999). Spectral cues can be thought of as mostly monaural as the majority of localisation studies have found no evidence for an interaction between the signals at both ears (Carlile et al., 2005). There is some work however that suggests the auditory system codes for differences between the spectral cues at both ears for specific regions in space (Jin et al., 2004).

An example of the mentioned spectral cues for a selection of positions in space for the left ear of the author is shown in figure 3 as a function of frequency. The locations are all along the midline. Each location is labelled using azimuth and elevation angles in the form (azimuth, elevation). From top to bottom in the figure, locations run from below and in front of the listener to behind and below the listener along an arc. The figure shows that the spectral features of the filters vary with elevation particularly for higher frequencies, above about 3 kHz; in fact the size of the ear is too small to interact with wavelengths for lower frequencies. Also shown are the effects of the head and torso, which are evident at lower frequencies. The frequencies are displayed on a logarithmic scale as this is a good approximation of the resolution on the basilar membrane for different frequencies. Magnitude is also presented on a logarithmic scale as this approximates how the auditory system interprets differences in loudness. The variations in the magnitude spectrum provide different salient cues to location given that humans have a just-noticeable-difference for pure tones of about 1 dB (Zwislocki and Feldman, 1956).

The variations in the filtering effects of the listener's morphology are crucial to coding for the sound source's location in space. Figure 4(a) shows the detail of these variations for the left ear filter imposed by the listener's morphology for all positions in front and to the left of the listener (azimuth locations from 0° to -150° and an elevation of 0°). The colour contours of the two charts represents the magnitude in the spectrum in decibels. One position in space is represented horizontally on the charts, i.e. each horizontal slice represents the variations in the gain of the filter. The auditory system learns the colouration of the spectrum at different positions as a signature for location and will compare the filtering effects of sound sources with those stored in memory. In comparison to the filter colouration, figure 4(b) displays the level of detail that is actually registered by the auditory system when taking into account the effects of what is termed the cochlear filter; this filter is distinct from the one caused by the pinna and represents the frequency dependence of auditory sensitivity (Carlile and Pralong, 1994). Despite the reduction in detail between the two plots in figure 4, the change in spectral cues as a function of azimuth is still clear in figure 4(b); different horizontal slices, will always be perceptually different due to the different colourations across frequencies. A detailed description of the different known spectral cues and their suspected purpose will be provided in chapter 4.

The notion that spectral cues play a crucial role in determining sound source location is generally accepted and can be easily tested. Two studies that tested subjects' ability to determine the location of a sound source whilst either completely covering the outer ear and leaving only the entrance to the ear canal unobscured (Roffler and Butler, 1968), or by occluding cavities within the ear with a mould (Gardner and Gardner, 1973), demonstrated deteriorated accuracy when compared to normal hearing. Subjects also demon-

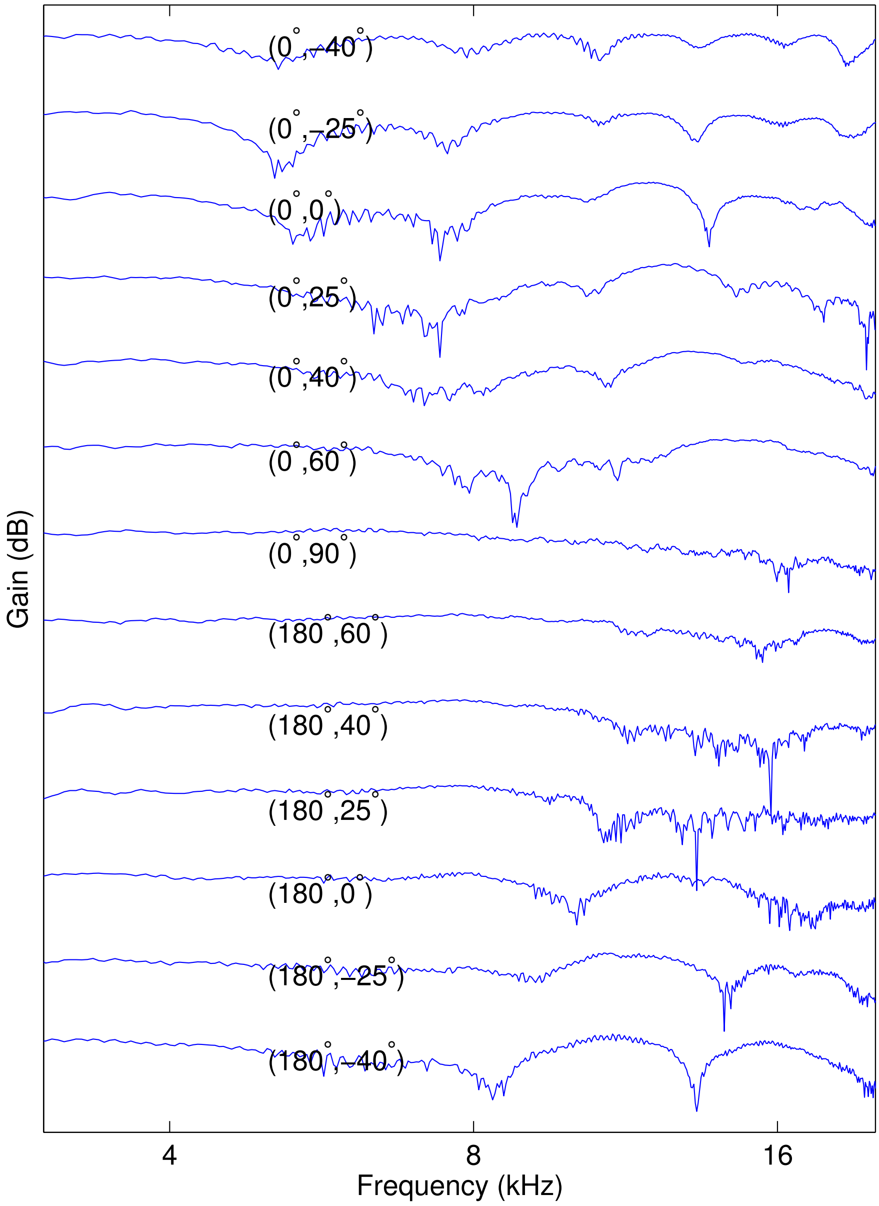


Figure 3: Filtering effects of the left ear for the author of this body of work for various elevations. All locations are along the midline, i.e. on the vertical plane that divides the left and right hemispheres. The azimuth and elevation of the recorded sound source for the different locations is labelled in degrees on the figure.

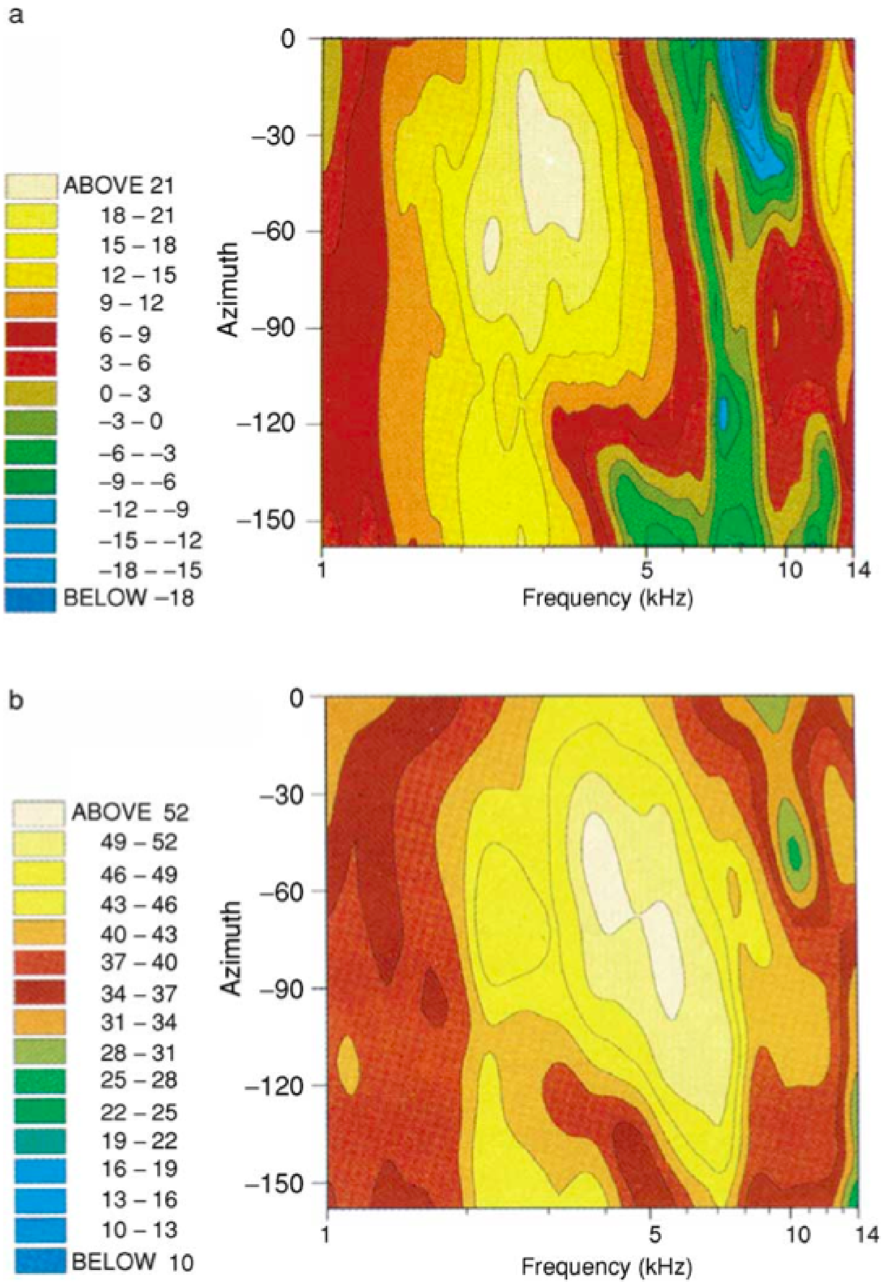


Figure 4: Variation in the spectral cues as a function of location (changes in azimuth) of a sound source for the region of space directly ahead and left of the listener. Spectral variation is shown (a) for recordings from inside the ear canal using specialised microphones, and (b) at the level of detail that is likely to be encoded by the auditory system after passing through a cochlear filter. Frequency is plotted on a logarithmic scale and the gain of the filter is indicated by the color contours, which are arranged in 3 dB steps. Figure taken from (Carlile et al., 2005).

strated difficulty in determining whether the sound source originated from in front of them or behind them; a phenomenon that will be analysed in chapter 6.

2.2.3 *Distance cues*

The perceived distance of sound sources, whilst not critical to determining their location, is another important aspect of how the auditory system interprets an acoustic environment. Variations in the distance of a sound source has numerous effects on the acoustic energy that reaches the listener's ears, and for this reason is interpreted based on a number of different auditory cues.

Humans are in fact quite poor at judging the distance of sound sources in real world environments; listeners are found to consistently and exponentially underestimate true distances for sound sources more than approximately 2 m away, and overestimate for sound sources that are closer than 2 m (Zahorik et al., 2005). Distance is judged based on the principle cues of intensity and direct-to-reverberant energy ratio. The intensity cue relates to the fact that the perceived intensity of a sound source will decrease as a function of the inverse-square law in ideal conditions and with a decreased rate in reverberant environments. The direct-to-reverberant energy ratio distance cue is present when reflections occur and is interpreted as the ratio of the energy that reaches the listener directly to the energy that reaches the listener after reflecting off surfaces in the environment.

More subtle distance cues are provided by a change in the spectrum due to either the absorption of frequencies in the air (for distances greater than about 15 metres) (Blauert, 1997) or reflections (Von Békésy, 1960), and by differences in ITDs and ILDs for sound sources in the acoustic near-field (less than 1 metre from the listener) (Shinn-Cunningham et al., 2000).

2.3 the head-related transfer function (hrtf)

Taken together, interaural, spectral and distance cues can be thought of as the fundamental pieces of information required by the auditory system to be able to perceive sound sources in space in terms of their location. They also relate to other aspects of how we perceive sounds, such as whether an auditory image is coherent, as detailed in the next chapter. These cues are contained in the transfer function, for a linear time-invariant system, that describes the listener's morphology response to the sound source. It is for this reason that we term this acoustic filter Head-Related Transfer Function (HRTF). The time-domain representation of an HRTF is the Head-Related Impulse Response (HRIR).

The importance and role of the HRTF in the perception of our auditory environment will be the focus of this body of work. More specifically the spectral cues, which as previously explained provide location dependent variations in the magnitude of the HRTF at each ear, and their relation to the listener's morphology will be a central theme. The HRTF will thus often refer only to the spectral variations in magnitude as displayed in figure 3 despite the fact that a listener's HRTF contains many other auditory cues. The significance of specific aspects of a listener's HRTF is further detailed in chapter 4.

There are in fact many other cues being used to interpret sound sources in our auditory space; an important feature of the human auditory system is that it has evolved to code for a broad range of auditory cues so that we are able to seamlessly locate sounds even when some cues have been degraded, such as for a reverberant environment. One such example are the cues provided from the movement of the sound source or listener. These dynamic cues are produced by the direction of the change in the interaural cues, which are dependent on the sound source's location and thus act as an unambiguous indicator particularly for determining whether sounds originate from in front of or behind the listener (Wightman and Kistler, 1999). Despite a growing body of evidence detailing how the auditory system functions, the redundancy in the information used by the brain to locate sounds in space makes it very complex and subsequently the study of the role and importance of individual cues difficult.

An important aspect of the HRTF is that the auditory system is continuously recalibrating these location specific auditory filters based on what is seen, or even felt, by the listener. The degree of neuroplasticity involved in this recalibration is directly related to changes in the listener's morphology due to the head or ear growing larger as we age, or simply from a haircut. It has been shown that listeners can completely relearn new HRTFs if need be (Hofman et al., 1998). The corollary of this fine-tuned adaptation to the HRTF is that there exists significant perceptual differences in HRTFs between listeners. The impact of individual differences between listeners' HRTFs is described in detail in chapter 5. This body of work addresses these differences in HRTFs between listeners and presents various methods for quantifying them. These methods can then be used to generate a personalised virtual auditory space, which creates the illusion of sound sources in space over headphones, via what is known as binaural synthesis (next chapter).

3 BINAURAL SYNTHESIS

The aim of this chapter is to describe the different methods for implementing our knowledge of how the Head-Related Transfer Function (HRTF) is interpreted by the auditory system in the real world in order to produce virtual auditory technologies. This implementation is known as binaural synthesis: a simulation of sound sources in space, generally presented to the listener over headphones. The illusion of an auditory scene, or a Virtual Auditory Space (VAS), is possible due to the fact that all the information that a listener needs in order to perceive sound sources in their environment is embedded in the signal entering the listener's ear canals. Thus a near perfect representation of these signals at each ear via a synthesis, played over headphones to the listener, will recreate any acoustic environment. Applications of such a technology are fast evolving and include: a powerful experimental tool, improving hearing aids, creating an audio immersion for gaming and movies, and orientation cues for aircraft pilots and the blind.

One very simple example of a VAS is the recording of an auditory scene, with any number of sound sources, using two small microphones at some position along the ear canals of a model head, and using the recorded signals as playback for a listener over headphones. This technique is known as a dummy-head recording, and can be thought of as a physical binaural synthesis as opposed to a numerical binaural synthesis. A purely numerical binaural synthesis, used to create a realistic illusion of sound sources in space (i.e. not involving a recording of free-field sound sources), is in fact a more complex and difficult task. A binaural synthesis of a virtual sound source requires the convolution of a monophonic signal representing the sound source with a pair of filters that embody the acoustic transfer function between the sound source location and the listener's tympanic membrane inside the ear. These filters are the HRTFs detailed in the previous chapter and have embedded in them all the auditory cues required for perceiving sound source location, namely interaural time and level differences (ITDs and ILDs respectively), and spectral cues. Thus for a high fidelity rendering of sound sources in VAS an accurate representation of the HRTF is needed.

3.1 MEASURING HRTFS

The most obvious method for calculating HRTFs is by recording a sound source in space using microphones at the ear canals of a listener or dummy head. The general methodology involves positioning the microphones, ensuring that the listener's head is in a fixed position, and recording a broadband stimulus played over a speaker that is mounted at a distance in such a manner so that it can sample the space around the listener's head at a high enough resolution. All measurements are made in an anechoic chamber so

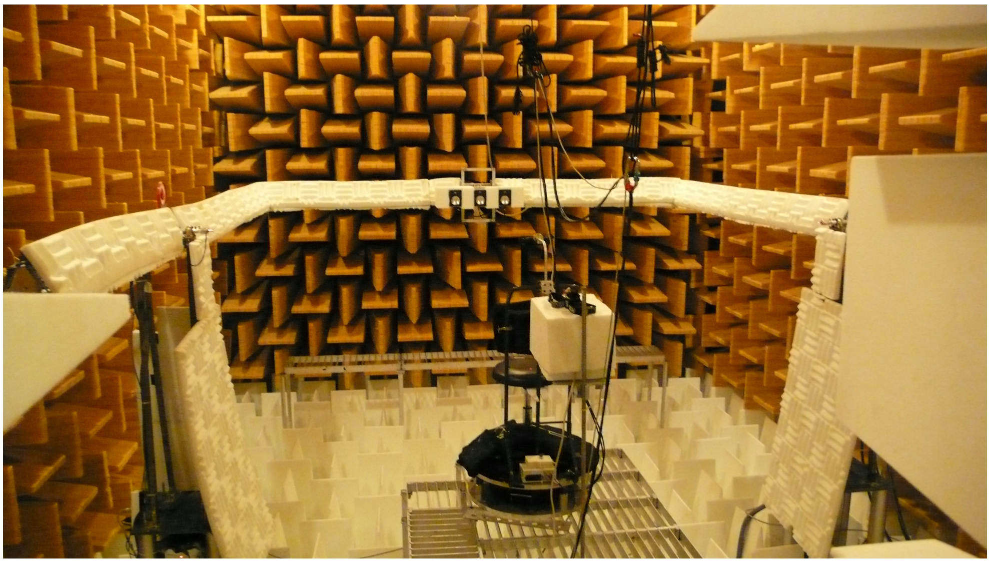


Figure 5: Example of the experimental setup for recording HRTFs in an anechoic chamber. The listener is seated at the centre of the chamber with microphones in both ear canals while the mechanical arm with speakers attached moves up and down for different elevations. Different locations in azimuth are recorded by rotating the chair the listener is seated on.

as to describe only the interactions of the sound waves with the auditory periphery (see figure 5 for an example of the experimental setup). The position of the microphones can be either at the entrance of the ear canal (with the ear canal open or blocked) (Moller, 1992) or inside the ear canal (Pralong and Carlile, 1994). Placing microphones at the entrance of the ear canal, and effectively blocking it, has several advantages relating to safety, a better signal-to-noise ratio, and less potential movement of the microphones during recording, which degrades the fidelity of the registered HRTFs (Wightman and Kistler, 2005). Another advantage of using the blocked ear canal method is the ability to avoid any standing waves within the ear canal since the ear canal resonance is not included in the HRTF measurement, which can be applied later to the response as it is independent of the sound source position (Møller et al., 1995b). In addition, for frequencies greater than the first cross-mode of the ear canal, non-planar waves propagate and spatial information can no longer be decoded as the eardrum is now excited as a modal membrane; the filtering effects of the pinna are no longer only source position dependant. This also means there can be reflections along the ear canal conduit as the sound propagates, as it is no longer a plane wave, meaning different propagation lengths (or multiple paths) for the same incident source direction.

3.2 PROCESSING HRTFS

Before the HRTFs can be used for a binaural synthesis it is standard procedure to implement a processing of the recordings. This is an important step for generating a high fidelity rendering in VAS as any recording system will incorporate imperfections and the raw responses are not representing what is necessarily perceptually relevant to the auditory system.

3.2.1 *Equalisation*

In the first instance the measurements are usually treated so that any artifacts of the recording system are removed. This form of equalisation is done by deconvolving the raw HRTFs with a measurement from the speaker in the absence of the listener's head. A calculated mean of all measured HRTFs, weighted to account for a distribution of the position of the speaker across recordings that might not be uniform, can also be used in the deconvolution. The latter method removes any non-directional information from the HRTFs and for this reason the resulting filters are termed Directional Transfer Functions (DTFs) (see Middlebrooks and Green, 1990).

3.2.2 *Minimum-phase and pure delay*

For the implementation of binaural synthesis using DTFs, the most common approach for generating the complex transfer function is to use a minimum-phase filter and a pure delay. A minimum-phase spectrum provides a Head-Related Impulse Response (HRIR) with the same spectra as the original HRTF, but with the energy redistributed to a single main impulse. A pure delay is assigned to the minimum-phase filter and is a coarse approximation of HRTF phase since it does not depend on frequency. The use of such an approximation is justified as long as the low-frequency Interaural Time Difference (ITD) information is available; the human auditory system is not sensitive to interaural HRTF phase spectra as long as the overall ITD of the low frequency components provides a reliable cue (Kulkarni et al., 1999).

It has also been shown that localisation accuracy for virtual sound sources is not affected by modelling HRTFs as minimum-phase filters and pure delays for most locations in space (Kistler and Wightman, 1992). Minimum-phase modelling works well for frontal sources with relatively small ITDs, but not for sound sources near the interaural axis and behind the listener (Katz et al., 2005).

There are various techniques used in the literature for estimating the pure delay component for a minimum-phase representation of HRTFs (see Nicol, 2010, for a review). Probably the most common method used, and the technique of choice in this body of work, is the maximum of the interaural cross-correlation between the left and right HRIRs. This method is specifically measuring the time shift between the HRIR envelopes and resembles the mechanism used by the auditory system to determine ITDs, which makes it particularly attractive.

3.2.3 *Smoothing of HRTF spectrum*

A smoothing of the irregularities in the HRTF spectrum magnitude is also a common final processing step. The fact that the frequency resolution of the auditory system approximates a logarithmic scale means that much of the fine details in the magnitude of the recorded HRTF are not perceived. The cochlear filter removes many of the features in the HRTF that might have been present in the initial recording that will not be encoded by the auditory system (see Carlile and Pralong, 1994, and figure 4 from section 2.2.2 in the previous chapter). In fact spectral cues are robust, in terms of localisation accuracy, to smoothing that results in a loss in frequency resolution far beyond that imposed by the auditory filtering (see Kulkarni and Colburn, 1998; Macpherson and Middlebrooks, 2003).

3.3 HEADPHONES IN BINAURAL SYNTHESIS

Headphones play a crucial role in effectively rendering sound sources in VAS. There exists an interaction between the listener's outer ear and the signal from the transducer of a headphone that resembles to some degree that of sound waves from a point source in space embodied in the HRTF. In addition, headphones have their own spectral characteristics that can influence the naturalness of a binaural synthesis due to coloration. Chapter 6 studies these interactions and the characteristics of headphones in detail using a localisation task. The following sections provide a background to the role of headphones in binaural synthesis.

3.3.1 *Choice of headphone type*

In real-world environments, as opposed to laboratory conditions, there are a wide variety of headphones used by listeners. Each headphone has its own characteristics that affect the sound produced. The most obvious difference between headphones is their physical ability to reproduce all frequencies in the audible range. For example, headphones that are inserted into the ear have smaller membranes at the transducer that impose limitations for lower frequencies. More specifically, for a binaural synthesis there are particular regions of frequencies in the spectrum of the signals presented to the listener that are more important than others and can be emphasised, as will be described in chapter 4 and will be analysed in further detail in chapter 8. If a headphone does not effectively produce any signal above a certain frequency, say in a region crucial to communicating spectral cues for elevation judgements, sound sources in VAS may appear poorly defined in terms of their perceived position in space. Ideally headphones will produce all audible frequencies at the same level, known as a flat frequency response, so that the filtering effects of the HRTF imposed via the binaural synthesis are not affected. In reality, in order to compensate for the physical limitations of a transducer, many headphone manufacturers choose to produce frequency responses that emphasise some frequencies over others and give the headphones their own spectral flavour and identity.

3.3.2 *Variations in frequency response*

The ability of headphones to reproduce a signal is determined by the frequency response of the hardware itself and for the most part determined by the manufacturer. The most comprehensive investigation of a variety of commercially available headphones was performed by Møller et al. (1995a). In the mentioned study, headphones were organised into three main categories: supra-aural, in which the headphone rests on the ear, circumaural, in which the headphone completely covers the ear making contact only with the head, and free-from-ear, in which the headphone makes no contact with the listener's head or ear but sits close to the entrance to the ear canal. Two more categories could be added to this list in order to encompass the majority of the different types of headphones on the market. The first would be the intra-aural, in which the headphone is small enough to be placed inside the ear canal of the listener, and the second would be that of the bone conduction, in which vibrations up against the bones connected to the inner ear transmit sounds to the cochlea. The latter has been used for augmented reality due to certain advantages it affords, such as being able to leave the ear completely unobscured so that environmental sounds can be heard (Walker et al., 2007).

In the study by Møller et al. (1995a), the frequency response of the headphones was measured in much the same way as HRTFs are; a small microphone is placed at the entrance of the ear canal. A broadband stimulus is presented from the headphones, worn by different listeners, and the response recorded. Results from the comparison of 14 different headphones belonging to the first three mentioned categories demonstrated that there were significant differences between the frequency responses of the headphones. In general, it was found that the headphone responses were not flat, showing large fluctuations with frequency. The frequency responses of the headphones were smooth up until approximately 3 kHz. Above this range responses were characterised by large peaks and notches, similar to those found in subject HRTFs. Kulkarni and Colburn (2000) have shown that these spectral features can be of similar magnitude and bandwidth as those in HRTFs.

A significant amount of the variation between the headphones can be explained by resonances forming inside the headphone cavity, particularly for the supra-aural and circumaural types as shown by Xie et al. (2009). In their study, the frequency responses of two circumaural headphones and one research-grade intra-aural headphone (the same as used in the study described in chapter 6) was measured using a dummy head with small built-in microphones inside the ear canals. The intra-aural headphone displayed less dramatic peaks and notches in its frequency response than the other headphones tested. The two circumaural headphones displayed vastly different frequency responses.

3.3.3 *Headphone transfer function*

The peaks and notches found in the frequency responses of the headphones, for the previously mentioned studies, can only partly be attributed to the headphones themselves. This is due to the fact that the effect of the outer ear must play a role, as the frequency responses were measured using small microphones either inside or at the entrance of the ear canals of a listener or dummy head. The sound waves from the headphones therefore interacted with the pinna and also the ear canal. The only exception being the intra-aural and bone conduction headphones, due to the fact that the signal is played directly to the cochlea, bypassing the outer ear. Yet intra-aural headphones are still affected by the ear canal resonance and impedance. Bone conduction headphones are not affected by the ear canal resonance, but are influenced by the bone conduction transfer function. The acoustic filter that is applied to the signal originating at the headphones is known as the Headphone Transfer Function (HpTF) (see Møller et al., 1995a). The HpTF encompasses a chain of transfer functions from the recording microphone, the headphone, and the impedance of the ear canal. For an accurate description of the combined effects of the headphone and pinna, described by the HpTF, all these elements must be known. HpTFs have been shown to vary significantly between listeners (Møller et al., 1995a), with some authors suggesting the need for an individualised HpTF when compensating for these colourations (Pralong and Carlile, 1996; Wightman and Kistler, 2005). Measurements of HpTFs even display significant variation between replacements of the headphone itself for a particular dummy head (Kulkarni and Colburn, 2000; McAnally and Martin, 2002) or listener (Møller et al., 1995a), yet there is no clear agreement on whether this might be perceived in a binaural synthesis as other than a global coloration effect.

3.3.4 *Headphone equalisation*

The HpTF, albeit a non-directional filter of the acoustic signal, is an unwanted artifact in a binaural synthesis over headphones. It embodies the actual characteristics of the headphone transducer, or the frequency response of the headphone, along with the transfer function between the headphone transducer and the point at which the microphone is positioned in the ear canal. Since it is preferable to use HRTFs that resemble as closely as possible those of the listener, the effect of the HpTF is often removed from a binaural synthesis. This is achieved by using a convolution of the desired signal with an inverse filter, generated from the inverse of the magnitude of the HpTF, before playback over headphones. Headphone equalisation is one of the central themes to a study detailed in chapter 6.

3.4 INTERPOLATING HRTFS

For a high fidelity rendering in VAS there needs to be a seamless representation of sound sources for all positions around the listener. Due to the fact that HRTFs are measured for finite positions in space, usually as a mesh of positions on an imaginary sphere, an interpolation is needed for any other desired position not part of the recording. When performing interpolations there are two related components to creating a perceptually seamless VAS; there is the issue of the density of measurements in space required and the interpolation technique to be used.

Using a reconstruction of HRTFs in the time domain Langendijk and Bronkhorst (2000) performed a linear interpolation and asked subjects to subjectively compare interpolated and original HRTFs in terms of differences in the spectrum (such as timbre) and location of the sound source. The results suggested that spatial resolutions of 6° in azimuth and elevation were adequate for equivalent subjective ratings of timbre, whilst resolutions between 10° and 15° were shown to provide a perceptually valid VAS in terms of localisation. Hartung et al. (1999) compared two different interpolation algorithms both in the time and frequency domain. It was found that the technique using spherical splines, a method that involves a global function using all measured HRTFs to compute an interpolated HRTF, provided the best results in terms of subjective and objective judgements made by subjects. The interpolation was performed in the frequency domain and used a resolution of 10° in azimuth and 15° in elevation. A similar method was used by Carlile et al. (2000), and also proposed by Chen et al. (1995), using a spherical spline interpolation of the weights from a decomposition of DTFs via a principal component analysis. The interpolation was performed for both ITDs and minimum-phase representation of the HRTFs. The minimum resolution of measured HRTFs needed for a continuous VAS was shown to be on the order of 150 recordings spread equally in space.

3.5 MEASURING THE EFFECTIVENESS OF BINAURAL SYNTHESIS

3.5.1 *Localisation*

As previously discussed in chapter 2, humans have the ability to determine the location of a sound source in space using interaural and spectral cues embedded in the HRTF. For binaural synthesis, one of the main ways to gauge the effectiveness of a rendering is to test how accurately the position of virtual sound sources can be determined by a listener. These localisation tasks have been used as the primary tool in studies that examine the role of the HRTF in binaural synthesis.

As will be seen in chapter 6, localisation errors have a certain form that is related to how our auditory system interprets the different cues to sound source location. In particular, there exists peculiarities in the nature of the distribution of responses by subjects performing localisation tasks. In nearly all localisation studies, subjects demonstrate, for a small portion of responses, uncharacteristically large localisation errors where a virtual sound source presented in front of the listener is perceived as coming from behind, and to a lesser degree the reverse. The percentage of responses that are of this nature is usually on the order of approximately 5 to 6% for localisation tasks in VAS (see for example Wightman and Kistler, 1989b; Makous and Middlebrooks, 1990). Some studies have reported higher front-back error rates of up to 13% (Katz and Parseihian, 2012). This is the so-called front-back confusion and has been shown to be distinct from errors of the same magnitude in the up-down directions, given the latter is less common; effectively zero instances in a study by Carlile et al. (1997) compared to 3.2% for front-back confusions. Makous and Middlebrooks (1990) performed a comprehensive analysis of front-back confusions showing that few of these types of errors deviated in terms of their azimuthal angle from a mirrored response across the vertical plane containing the interaural axis (i.e. the errors were almost always mirrored from front to back). These findings show that front-back confusions are not part of a generic cone-of-confusion error and are a distinct class of errors. This suggests that the auditory system is dependent on some fundamentally different cue in order to distinguish between the two hemispheres. Localisation studies have demonstrated that cues in the region of the spectrum between 8 and 16 kHz are important for determining whether a sound source is in front or behind a listener. It has in fact been shown that HRTF spectra display a peak in the frequency region from 11 to 14 kHz that is only present for locations of sound sources in the frontal hemisphere (Carlile and Pralong, 1994; Asano et al., 1990).

The proportion of front-back confusions in localisation tasks has been shown to be significantly higher for a binaural synthesis compared to that using free-field sound sources such as speakers (see Carlile et al., 1997). Front-back errors made whilst listening to a binaural synthesis are larger than free-field front-back errors, which are usually restricted to within 30° of the vertical plane containing the interaural axis (Carlile et al., 1997; Makous and Middlebrooks, 1990).

3.5.2 *Listening tests*

Another way in which the quality of a binaural synthesis can be assessed is via the use of perceptual judgements based on attributes of the sound. Of particular relevance to this body of work is the use of listening tests, in which a binaural synthesis is presented to a listener who is asked to register responses to the auditioned stimulus on some predefined scale. Listening tests aim to address the many other attributes of a virtual sound source, other than its location in space, such as whether the auditory image itself is coherent or diffuse, realistic or unrealistic, externalised or perceived as coming from inside the head.

There are three main techniques in which the evaluation of audio quality via listening tests is commonly performed (see Zielinski et al., 2008, for a review). The first is based on a triple stimulus with hidden reference approach, in which listeners have to access three stimuli at a time (say A, B, and C) and can switch between them at will. The first stimulus (A) is a signified reference, explicitly labeled as a reference offering the baseline audio quality in the test. The two remaining stimuli (B and C) consist of a non-signed reference (i.e. the same as the signified reference only not explained to the listener as such), commonly referred to as a *hidden reference*, and a processed recording (in the case of the this body of work, a rendering in VAS). This type of listening test, standardised by ITU-R BS.1116-1:1997, requires the listener to make a judgement of both B and C on a continuous scale for each presentation of the three stimuli. The scale used on this type of test is often numerical and labelled with language that describes the type of impairment of audio quality, for example *slightly annoying*.

The second main technique, developed to evaluate more distinguishable differences between stimuli than the former, is based on the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method (see for example Macpherson and Middlebrooks, 2000), standardised in the ITU-R BS.1534-1:2003. It is recommended that no more than 15 different stimuli be included in any one trial for this type of test. As the name suggests, there is a hidden reference included as one of the recordings being assessed, along with a signified reference and anchor (i.e. labelled for the listener). In the context of a binaural synthesis, the hidden reference might be a synthesis using the subject's own HRTFs, and the anchor might be a degraded version of the reference. In contrast to the previous method, the continuous scale used is a quality scale beginning at *poor* and ending at *excellent*.

The third technique is most commonly used for the evaluation of speech quality, in which listeners are presented with sequential recordings and asked to evaluate the quality of the recordings using five discrete categories ranging from *poor* to *excellent*. Hidden references are often included so that trials may be compared. Despite the existence of the mentioned three standardised listening tests, many variants are used in studies and a range of biases exist, which will be explored in chapter 7.

Examples in the literature of listening tests that are relevant to this body of work include a study by Usher and Martens (2007) in which the naturalness of a set of renderings in VAS was judged. Subjects were presented with pairs of speech stimulus as a virtual sound source and asked to determine which of the two sounded more natural. Seeber and Fastl (2003) used a selection of different HRTFs and asked subjects to make judgements based on a set of attributes including: whether the virtual sound source was perceived as coming from in front of the listener when presented in that region of space, and whether it was perceived at a constant distance (preferably at a distance that was far away).

3.5.3 *Externalisation*

As HRTFs are measured at a fixed distance and in anechoic conditions, sound source distance cues in binaural synthesis is an effect that needs to be added to the chain of signal processing for a realistic rendering of sound sources in VAS. This is due to the fact that distance cues depend heavily on reverberant conditions. Effectively rendering distance cues, as described in section 2.2.3, is closely linked to what is known as the externalisation of virtual sound sources. The externalisation of sound sources via binaural synthesis is a prerequisite to being able to perceive distance; the inability to properly externalise virtual sound sources translates to the impression that they are located inside the listener's head.

There are many possible causes for a poorly externalised binaural synthesis (see Nicol, 2010, for a review). The basic explanation is tied to degraded or conflicting interaural and spectral cues to sound source location (Hartmann and Wittenberg, 1996). Another possibility is the absence of a dissimilarity between the signals at the left and right ears of the listener; in natural environments there are significant differences in the signals at each ear due to asymmetries between the left and right pinna shape as observed by Searle et al. (1975) and Toole (1970). A binaural synthesis will often remove some of these asymmetries when performing an equalisation, as described in section 3.2.1. As mentioned, the effect of the anechoic conditions in which the HRTFs are recorded can lead to a lack of the impression of space, such as reverberation in a room. The addition of reverberation is one of the most common ways to enhance the externalisation of virtual sound sources. Including dynamic cues in binaural synthesis (see section 2.3), by tracking the listener's head and modifying the HRTFs used in the rendering accordingly, is another way to improve externalisation (Inanaga et al., 1995). As previously discussed in section 3.3, the headphones used in a binaural synthesis can result in poor externalisation due to their effect on the signal before it even reaches the lowest level of the auditory system: the basilar membrane.

The final obstacle to a high fidelity rendering in VAS, relating primarily to an effective externalisation of virtual sound sources, is the more subtle psychological effect of sustaining an auditory illusion. The most obvious difficulty is that in natural acoustic environments humans are used to being able to have some sort of visual feedback for a sound source in space. For binaural synthesis, this is very often not the case and there is a disparity between what is heard and the other sensory modalities. There is no better example of the strong reliance that the human brain has on visual cues to sound source location than what is known as the ventriloquist effect (Alais and Burr, 2004; Choe et al., 1975). For this effect, a conflict occurs between the location of the visual cues of a sound source and its actual location, created by, for example, having a television screen of a person talking and the speech emanating from a speaker at a different location. When such a conflict occurs, listeners will have the illusion that the speech is originating at the television rather than the speaker in an attempt to fuse the two conflicting sensory inputs. In the complete absence of visual cues, there is a tendency, when listening over headphones in VAS, to localise sound sources inside the head, or from behind and above. Familiarity also plays a large role for an effective rendering in VAS due to the synthetic nature of binaural listening over headphones. This means that listeners who have had ample training or who are even experts in listening to virtual sound sources should have much better externalisation; this hypothesis will be explored in chapter 7.

4 SPECTRAL CUES

The previous chapters outlined the acoustic cues that our auditory system interprets for estimating the position of sound sources in space, and how our knowledge of this process has allowed for us to create compelling auditory illusions over headphones. The current chapter delves deeper into what specific aspects, embedded in the Head-Related Transfer Function (HRTF), are important for this seamless processing by the brain to take place. Of particular interest is the relation between spectral cues and the physical dimensions of our body; how sound sources in space interact with our external ears, head, and body, and create signatures to sound source location in the magnitude spectrum that our brain in turn memorises. Spectral cues as outlined in this chapter are any features in the frequency domain of the HRTF that enable a listener to perceive sound sources in space, whether it be in terms of sound source location or externalisation.

4.1 FREQUENCY RANGE OF SPECTRAL CUES

Humans have on average an audible frequency range between approximately 20 Hz and 20 kHz, yet this does not mean that spectral cues occupy this same bandwidth; in reality spectral cues occupy a narrower range of frequencies. A study by Algazi et al. (2001a) has demonstrated that there exist spectral cues in the lower frequency range (below 3 kHz) to the sound source elevation, particularly for sound sources that lie off the midline. However these cues are considered limited, with evidence that the low frequency range of HRTFs do not provide salient spectral cues to sound source location with respect to whether it is heard from in front or behind (Morimoto et al., 2003). Several studies have shown that significant spectral cues exist in the range between 4 and 16 kHz by studying localisation accuracy for sound sources on the midline (Hebrank and Wright, 1974; King and Oldfield, 1997; Langendijk and Bronkhorst, 2002). It is probable that when spectral cues exist in the higher frequency range, those in the lower range under 3 kHz become much less significant. The spectral cues below approximately 3 kHz have wavelengths that are too large to interact with the convolutions of the pinna due to its size and are caused by a diffraction and dispersion of sound waves by the head and torso. The most perceptually relevant frequency range, with respect to perceptual judgements of sound sources in Virtual Auditory Space (VAS), will be explored in chapter 8.

Given that spectral cues have been shown to exist within a specific frequency range, the question of whether specific narrow frequency ranges, or frequency bands, are interpreted differently by the listener is of interest. Blauert (1969) was one of the first authors to look at whether narrowband stimuli were localised differently depending on their centre frequencies. Sound sources in Blauert's study were speakers positioned at locations along the midline. Clear differences were observed as each band, termed "directivity bands", and sources tended to be perceived as originating from specific regions in space; for example the bands with centre frequencies at 4 and 12 kHz were localised in front and behind the listener respectively. Middlebrooks (1992) expanded the findings to all positions in space via a rendering in VAS, finding that elevation responses and judgements of whether the virtual sound source was originating from in front or behind the subject, are mostly affected by the centre frequencies of the narrow-band stimulus used. These auditory illusions however tend to dissipate as the width of the band increases and becomes more broadband, as is the case with most stimuli in our natural environment (Jin, 2001).

4.2 MONAURAL VS BINAURAL SPECTRAL CUES

A number of studies have looked at whether accurate localisation of sound sources is dependent on spectral cues being registered at one or two ears. Whilst there are varying results for unilaterally deaf and normal hearing listeners with one ear blocked, which suggests that some accurate localisation of sound sources can occur with only one ear (Butler et al., 1990; Slattery and Middlebrooks, 1994; Van Wanrooij and Van Opstal, 2004), these studies may have benefited from dynamic cues relating to the fact that listeners were able to move their head in the free-field conditions in order to help them determine sound source location. Wightman and Kistler (1997) was able to show that in reality when monaural spectral cues are presented to the listener in a controlled VAS environment, monaural localisation is very poor.

4.3 TEMPORAL AND LEVEL FACTORS

The ability of humans to localise sound sources is affected by the length of the stimulus and the level at which it is perceived. Hofman and Van Opstal (1998) have shown the importance of temporal factors for sound localisation, demonstrating that a minimum stimulus duration of 80 ms was required for accurate sound source localisation. They proposed that the auditory system does not assess sound source elevation by integrating over the whole length of the stimulus, but rather by consecutive short-term estimates (on the order of a few milliseconds).

The relationship between stimulus duration and level has been assessed by Vliegen and Van Opstal (2004), with findings demonstrating that if the signal intensity was too low or too high, sounds were poorly localised. Macpherson and Middlebrooks (2000) assessed subjects' ability to localise sound sources at varying intensity levels. They used a measure called sensation level, in which the intensity was calculated in decibels above each subject's detection threshold, determined by presenting the stimulus at ear level directivity in front of the listener and controlling the signal level. It was found that localisation accuracy of sound sources presented at sensation levels above 40 to 45 dB suffered. This was a result that has been confirmed by Vliegen and Van Opstal (2004). It was suggested that for low sensation levels (approximately 28 dB), the signal-to-noise ratio was insufficient for providing auditory cues to sound source location, and that for high sensation levels (approximately 73 dB), there would be interfering artefacts and central neural processing mechanisms such as compression and neural saturation respectively (Vliegen and Van Opstal, 2004). Limiting the gain of the stimulus was also important in order to avoid level adaptation and the acoustic reflex (Stapedius reflex) usually observed at levels above approximately 70 dB SPL.

4.4 ROLE OF SPECTRAL DETAIL

One key aspect of studying spectral cues is an understanding of the spectral detail that is perceptually relevant in the HRTF and the amount of detail required for accurate localisation. Establishing the role of spectral detail has an influence on any analysis of HRTF data and their use in binaural synthesis. The obvious upper limit is at the level of the spectral resolution of the cochlea. It has been shown by Carlile and Pralong (1994), using the original auditory filter model devised by Glasberg and Moore (1990), that much of the measured detail in HRTFs becomes smoothed and irrelevant after cochlear filtering (see figure 4 from section 2.2.2).

To the degree that ripples in the spectrum of a broadband stimulus can disturb localisation accuracy, Macpherson and Middlebrooks (2003) has shown that a density between 0.5 and 2 ripples per octave produce substantial errors. In terms of the resolution needed in HRTF filters for binaural synthesis, Kulkarni and Colburn (1998) found that subjects could not distinguish between free-field and VAS until the spectral resolution was reduced to 1,562 Hz (i.e. 16 Fourier coefficients at a sample rate of 50 kHz). However, in the study by Kulkarni and Colburn (1998), they used a linear spacing of coefficients that makes their results incompatible with the logarithmic scale used in the Macpherson and Middlebrooks (2003) study. Senova et al. (2002), who examined the accuracy with which virtual sound sources, generated using frequency-warped filters of various spectral resolutions, could be localised, found that the results were similar to the free-field condition for a resolution that was at least 4.6 points per octave (i.e., 32 coefficients across approximately seven octaves from 0.2 to 25 kHz). This representation of the filters, which was close to a critical band distribution, was comparable to the Macpherson and Middlebrooks (2003) study and was indeed consistent with their finding that ripples above 4 points per octave are not particularly important for localisation, according to a review by Carlile et al. (2005).

More generally, Asano et al. (1990), using a smoothing of the HRTF spectrum via an auto-regressive moving-average, found that the fine details in the spectrum might be used for differentiating front-back location of sound sources in regions below 2 kHz and in the high frequencies. Elevation cues did not appear to be dependent on the "microscopic patterns" of the HRTF, which can be thought of as the fine structures in the spectrum, rather the macroscopic detail above 5 kHz.

4.5 SPECTRAL FEATURES

Many studies have sought to determine what specific spectral features in the HRTF are most important in terms how we perceive the attributes of sound sources in space and VAS, such as position, distance, and externalisation. Most of the literature has focused on spectral information in sound localisation. As previously outlined, spectral cues help to resolve the cone of confusion and can therefore be thought of mainly as an elevation cue (i.e. determining polar angle). There exist two trains of thought on spectral features that divides the research into studies either seeking to outline the importance of a feature in the HRTF related to determining sound source location on a somewhat one-to-one relationship (overt features), or showing that there are features that only become apparent once HRTFs for all locations in space have been considered (covert features). The following sections aim to draw on some of the most relevant findings relating to the role of overt and covert features.

4.5.1 *Overt features*

The most prominent overt spectral feature is that of the frequency notch. Notches in the HRTF magnitude are characterised by narrow (width of up to approximately one octave) deviations and have been the topic of much research (Greff and Katz, 2007; Iida et al., 2007; Rodriduez and Ramirez, 2005). Two of the earliest studies to analyse the role of spectral notches, by Hebrank and Wright (1974) and Butler and Belendiuk (1977), used a localisation task along the midline, given that the interaural cues are invariant for all locations. Their results highlighted the role of the frequency notch in different regions of the spectrum as cues to particular regions in space. The general trend is for the frequency notch to migrate from about 5 to 14 kHz as the sound source moves from below to above the interaural axis. Bloom (1977) demonstrated this by creating the illusion of a sound source moving from low to high elevation by varying the centre frequency of a notch inserted in a broadband stimulus, presented at a fixed elevation. The capacity for humans to detect spectral notches in a broadband stimulus has been shown to be limited when sound sources are static, suggesting that they are used mostly as a dynamic cue when the sound source is moving or the listener is able to move their head (Moore et al., 1989).

4.5.2 *Covert features*

In addition to findings that suggest the auditory system uses a broad frequency range for spectral cues, there is evidence that some cues might only become apparent once HRTFs for all positions in space have been considered. Such cues are termed covert features, and have been highlighted by studies looking at the peak, or covert peak, in the HRTF magnitude that appears maximally for a particular frequency across all positions in space. As previously described, early studies by Blauert (1969) and also Humanski and

Butler (1988) showed that band limited stimuli when presented to listeners over speakers is localised dependent on its centre frequency. Predictive models have been able to determine the localisation responses of subjects using covert peaks with some success when compared to other models using broadband spectrum (Middlebrooks, 1992). It should be noted however that it may be difficult to apply the findings to how humans usually perceive sounds in space given that it is rare to encounter band-limited sound sources in our natural acoustic environment. Butler (1987) and Butler et al. (1990) expanded covert peaks to broader regions of space known as covert peak areas that contained all positions within a range 1 dB from where it was maximal. Using this new definition, Jin (2001) was actually able to show that the predictive model was successful for a range of bandwidths.

Support for the role of covert features has also come from work by Middlebrooks (1999a) in which a global frequency scale factor has been used to describe the differences between HRTFs for many locations in space between subjects. The global frequency scale factor determines how a set of HRTFs can be shifted in the frequency domain so that features in the spectra such as peaks and notches align with another set of HRTFs. This particular analysis method is one that will be used in a study by the author in chapter 8.

4.5.3 *Spectral cues using broadband models*

Whilst studies analysing covert peaks and overt notches have provided valuable insights into the role of spectral cues, any model aiming to describe how the auditory system interprets sound sources in space based solely on one or the other is in danger of simplifying a process that most probably involves a combination of cues over the spectrum.

In terms of the spectral notch, there are findings that suggest that localisation is not solely based on the centre frequency of the most prominent notch. Macpherson and Center (1994) showed, using a simple model describing sound source elevation judgements, that centre frequencies of notches alone was not able to predict the pattern of responses from subjects even after taking account the variation across listeners. It has been shown that the parts of the spectrum that are outside the range of the primary notch are necessary for effective localisation of sound sources (King and Oldfield, 1997) and, as previously stated, listeners are able to localise sound sources to some degree using cues outside the range of frequencies in which we find spectral notches (Algazi et al., 2001a).

Langendijk and Bronkhorst (2002) have also shown that spectral notches, or any feature for that matter, contained in regions of the spectrum of half an octave bandwidth above 4 kHz are not significant for localisation, and that it is likely that broadly tuned features are used. This study highlights that the fine detail in HRTFs is most probably not important for localisation. Their study flattened the spectrum of HRTFs for a variety of bandwidths and tested localisation performance. A model was proposed that assumes the auditory system compares the spectrum of the signal arriving at the eardrums from a source in space with a set of stored spectral templates associated with particular sound source directions. In this model, the best match between a template and HRTF will determine the perceived location of the sound source, assuming that there is some a priori knowledge of the stimulus. This was shown to be an effective model for most of the subjects tested, using a localisation task. Similar models have been proposed in which peaks and notches across the whole spectrum are used to predict localisation (Hofman and Van Opstal, 1998; Chung et al., 2000).

Hofman and Van Opstal (2003) used a rapid presentation of broadband stimuli with random spectral shapes and the tracking of saccadic eye movements toward the perceived stimulus locations to analyse whether specific spectral features could be highlighted as being significant to localisation. An analysis of the data using Bayesian statistics allowed for a reconstruction of spectral shapes, associated with sound sources from specific angles of elevation, that resembled subject HRTFs. Most importantly, the data showed that both types of features, peaks and notches, are employed by the auditory system, over a wide frequency range.

4.6 MORPHOLOGICAL INFLUENCE ON HRTFS

The spectral colouration in the HRTF that was seen in figure 3 and 4 of section 2.2.2 is the result of a complex interaction between incident and reflected sound waves from a source and different parts of the human body, namely the outer ear, or pinna. Blauert (1997) explains that:

> The acoustical effect of the pinna is based upon reflection, shadowing, dispersion, diffraction, interference, and resonance.

The head and body's influence can be mainly attributed, respectively, to the shadowing effect of sound waves, and a reflection from the shoulders of the listener. Studies by Algazi et al. (2002) and Algazi (2001) using a human-like dummy head called KEMAR and a simple model composed of two spheres or ellipsoids to describe the body and head's behaviour, have made observations about the nature of the auditory cues provided. The body was shown to have a comb filter effect on frequencies below 3 kHz; the pinna does not interact with sound waves below this frequency due to the relative size of the wavelength. The interactions of the sound waves with the body causes notches in the magnitude spectrum that vary with elevation and it is suggested that these could act as cues to the up-down location of sound sources, as previously described. Even more prominent patterns occur for the interaction between the head and sound waves. A constructive and destructive interaction of sound waves due to diffraction occurs at the ear furthest to a sound source, and this pattern varies as the position changes in azimuth and elevation.

The other component of the listener's morphology, the external ear, has a profound influence on the HRTF. The asymmetrical shape of each ear allows for a location dependent interaction with sound waves for sources in space. A sound will be filtered in a very different manner if it is originating from above than if it is originating from below the listener. Figure 6 shows

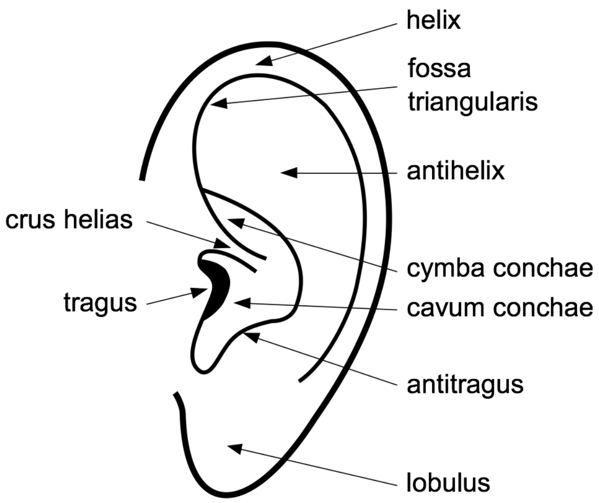


Figure 6: Labelled diagram of the human ear. Image taken from Guillon (2009).

the human ear labelled with terms that will be used throughout this body of work. Some of the most important features are the ensemble of cavities represented by the cymba conchae and the cavum conchae, and the fossa triangularis, represented by the region of the ear above the cavities.

Shaw and Teranishi (1968) were one of the first to study the complex reflections within the pinna that cause variations in the frequency magnitude spectrum, for frequencies above approximately 3.5 to 4 kHz. A set of six modes were described, of which the second and third are of particular importance to the current study; these modes are said to be *vertical*, generating the peaks observed in the spectrum, as they cause significant variations in the HRTF as a function of sound source elevation. These early findings were found to be in good agreement with a later study using a numerical modelling of HRTFs on a model ear and head by Kahana and Nelson (2000).

Gardner and Gardner (1973), in another early study, used the occlusion of portions of the pinna to show that localisation was significantly affected when incident waves were unable to form delayed reflections, contributing to the peaks and notches observed in HRTF spectra. The notion of time-delayed reflections as playing a dominant role in the creation of spectral notches (a cue to sound source elevation) was then suggested by subsequent studies (Hebrank and Wright, 1974; Wright et al., 1974).

Efforts by Lopez-Poveda and Meddis (1996) to isolate specific portions of the pinna and relate them to aspects of the HRTF spectrum, by removing or retaining sections of a replica of the human ear, have been less successful in deconstructing the complex interaction of incident and reflecting sound waves in the pinna. Despite this interdependence of morphological features in the outer ear, the authors were able to relate notches in HRTF spectra to the separation between the cymba concha and cavum concha (see figure 6). The significance of different morphological features in terms of their influence on the HRTF and corresponding binaural synthesis is the topic of a study in chapter 9.

As will be discussed in detail in the following chapter, the geometry of the ear, and subsequently the spectrum of the HRTF, varies significantly from listener to listener and this has a significant impact on how effectively sound sources can be rendering in VAS for individuals.

5 HRTF INDIVIDUALISATION

The current chapter explores the development of high fidelity renderings in Virtual Auditory Space (VAS) and its psychophysical underpinnings. As detailed in the previous chapter, the auditory system has evolved to make use of a range of cues that help humans effectively interpret their acoustic environment. In fact, the auditory system is remarkably tuned to these cues; so much so that small changes in the Head-Related Transfer Function (HRTF), and in particular the spectral cues (see previous chapter), can lead to large differences in how a sound source is perceived. For a particular listener the auditory system has learnt to code for HRTFs as signatures to sound source location; the asymmetrical shape of the pinna is mostly responsible for the variations observed in the spectrum of the HRTF for different locations in space (see section 4.6 from the previous chapter). This dependency of the HRTF on the shape of a listener's ear has lead to a large body of research showing the need for what is termed individualised HRTFs, in which the HRTFs are recorded on the person using them in order to produce a realistic impression of virtual sound sources for a binaural synthesis. Unfortunately the recording of HRTFs, as previously detailed in section 3.1, is an expensive and laborious task, which has lead to a range of approaches at generating individualised HRTFs without the need to record for a large number of positions, or by completely avoiding the recording process all together. The following sections outline the evidence relating to the need for individualised HRTFs, and the many ways in which they can be produced.

5.1 USING NON-INDIVIDUALISED HRTFS

The need for individualised HRTFs can be best demonstrated by using non-individualised HRTFs to produce sound sources in VAS and measuring the effect this has for a number of different subjects. Studies that have looked at the nature of localisation errors for tasks using non-individualised HRTFs (Wightman and Kistler, 1993; Wenzel et al., 1993; Katz and Parseihian, 2012) have demonstrated that whilst little error is observed for responses in terms of lateral angle (see section 2.1), the ability of subjects to determine the elevation of stimuli is greatly disturbed. Wenzel et al. (1993) analysed the rate of up-down confusions in a localisation task, defined as a response that crossed the horizontal plane with respect to the target position, in a study that used a generic representative HRTF from a database for a large number of subjects and showed that confusions was higher than for an individualised HRTF. Katz and Parseihian (2012) have shown that even when an optimal HRTF is chosen for the subject from a database, individualised HRTFs still outperform non-individualised HRTFs. The results showed that elevation errors increased significantly for all but one of the 16 subjects with respect to localisation accuracy in the free-field (i.e. using individualised HRTFs). It is suggested that the lateral angle errors are typically smaller than elevation errors due to the fact that interaural cues vary less between subjects than spectral cues.

Studies that have used binaural recordings and localisation tasks, either produced using a different subject to that being tested or a dummy head, have also demonstrated the same kind of errors as observed in the Wenzel et al. (1993) study (see Møller et al., 1996, 1999; Minnaar et al., 2001; Middlebrooks, 1999a; Gardner and Gardner, 1973). This is a significant result as it shows that the degradation in localisation accuracy for non-individualised HRTFs cannot be solely attributed to the binaural synthesis itself when compared to free-field results (see Carlile et al., 1997).

In addition to an observed degradation in judgements of sound source elevation, studies using non-individualised HRTFs have shown a higher rate of front-back confusions, in which a sound source is mistaken as coming from behind the listener when presented in front. The confusion of a sound source coming from the front when presented behind the listener is less common. Wenzel et al. (1993) found that front-back confusion rates quadrupled when stimuli were presented in VAS using non-individualised HRTFs compared to free-field responses.

Judgements obtained via listening tests for binaural syntheses using non-individualised compared to individualised HRTFs are difficult to find in the literature. One study performed by Usher and Martens (2007) asked subjects to judge the perceived *naturalness* of speech stimuli in VAS using non-individualised HRTFs and the subject's own HRTFs. For this criteria the results were mixed in terms of determining which HRTFs were perceived as the most natural sounding (i.e. the subject's own HRTFs were not always judged as being the most natural sounding). These results suggest that the similarity of HRTFs used in a binaural synthesis, in terms of the magnitude spectrum, to a listener's own acoustic filters might not be as important when considering criteria other than localisation accuracy. This result is significant as many applications of binaural synthesis are not dependent on localisation accuracy per se, but rather the realism of the illusion. Further investigation is needed however before such claims can be made as the study by Usher and Martens (2007) is an isolated finding and tested only nine subjects using headphones that were not research grade. Chapter 7 further investigates the role of individualised HRTFs in perceptual judgements of virtual sound sources.

5.2 METHODS FOR PRODUCING INDIVIDUALISED HRTFS

As can be seen in the previous section, numerous studies have demonstrated the usefulness of individualised HRTFs, particularly in terms of localisation accuracy. Due to the difficulties involved in recording individualised HRTFs, a number of studies have developed unique methods for solving the problem efficiently. The different solutions can be broken up into three distinct categories that will be detailed in the following sections of this chapter. The first category encompasses solutions that still involve the recording of HRTFs on the listener; a reduced number of measurement sequences makes them more amenable to the consumer market. These methods aim to reduce the laborious task of measuring HRTFs at many locations around the listener. The advantage of this solution is that measurements are still recorded using the individual listener, and this provides a good basis for faithfully representing an individualised HRTF. The next category of solutions focuses on adapting or using previously recorded HRTFs (i.e. non-individualised HRTFs) and does not require HRTF recordings for each listener, which is a clear benefit for the consumer market. Whilst these solutions can often be easier and faster than the previous category, they do not provide a strictly *individualised* HRTF, but rather a *personalised* HRTF that is well suited to the listener. Depending on the method used, this personalised HRTF might be perceptually as good as or even better than an individualised HRTF. The final category of solutions avoids the recording of HRTFs entirely and makes use of computer simulations to provide individualised HRTFs.

5.2.1 *Reduced measurement sequences*

5.2.1.1 *Measuring HRTFs via method of reciprocity*

Zotkin et al. (2006) has proposed a method that produces HRTFs for many locations using only one recording via the principle of reciprocity. The process works by performing an elegant modification to the traditional HRTF measurement process, which involves exchanging the loudspeaker and the microphone, that is, to put a miniature loudspeaker in the person's ear and microphones at the positions where the HRTF is to be measured. The reciprocity principle states that if all other factors are held constant and only speaker and microphone are replaced then the measurements obtained would be exactly the same as the direct method in which measurements are made sequentially at many locations. In this way all HRTF measurements are obtained in parallel, resulting in a much faster recording time; only two recordings are needed, one for each ear. The reciprocity method has been shown to be in good agreement with the direct method, yet has drawbacks, due to technical limitations, relating to the narrow frequency band of the speaker used, which lies within the listener's ear; the small size of the speaker leads to poor low-frequency output. This lack of low-frequency output from the speaker means that a correction of the magnitude spectrum in this frequency range is required using approximations of the listener's head and body. In addition, the technique still requires expensive equipment and a technical operator.

5.2.1.2 *Interpolation from a reduced set of HRTF measurements*

This second, and more common HRTF generation technique, uses the interpolation of HRTFs (see section 3.4). A variety of algorithms can be used to predict HRTFs at any location in space using a reduced number of measurements evenly distributed in space. The number of required measurement locations can be reduced from approximately 400 to between 120 and 150 evenly distributed positions in space (Carlile et al., 2000; Martin and McAnally, 2007) or even down to as little as 45 to 65 locations if dynamic cues are provided to the subjects for localisation tasks (Guillon, 2009; Guillon et al., 2008). These studies show that even with a reduced number of positions an interpolation can provide a perceptually seamless VAS.

Hartung et al. (1999) performed a comparison of two common interpolation methods: inverse-distance weighting that uses a nearest neighbour approach taking the closest four positions for interpolation, and spherical splines that considers the entire measurement data to interpolate for any given position. The two methods were tested on both a time domain (impulse response) and frequency domain representation of the HRTFs, with the spherical splines proving to be the most effective in the frequency domain for both a signal and psychoacoustical validation. In a study by Guillon (2009) the effectiveness of spherical splines via a signal and localisation validation was also demonstrated. Whilst interpolation methods produce effective renderings in VAS, without the need for any corrections to the spectrum of the HRTFs in the low frequency range as the for the reciprocity method, it still requires acoustic measurements using expensive equipment.

5.2.2 *Not requiring HRTF measurements on listener*

5.2.2.1 *Tuning HRTFs*

Another solution involves simply letting subjects perform subjective judgements of the quality of a rendering in VAS in order to fine tune HRTFs and produce a personalised output. This method requires subjects to literally attenuate or amplify different frequency regions of the spectrum until a realistic impression of virtual sound sources is produced. In this way, the HRTF is said to be tuned by the listener. This is the first of the detailed solutions that does not explicitly aim to provide HRTFs based on free-field recordings; instead this technique offers an HRTF to listeners that has as its only criteria the ability to produce a convincing VAS, not focusing on whether the HRTFs resemble the acoustic filters of the listener.

A number of studies have worked on accentuating features of the HRTF such as peaks and notches to produce a more realistic binaural synthesis. Zhang et al. (1998) exaggerated the perceptual differences for sounds coming from different directions, and as a consequence were able to emphasize the pinna effects the auditory system uses to locate sound soures. The tuned HRTFs produced lower rates of front-back confusions when compared to unmodified non-individualised HRTFs. Silzle (2002) had an expert modify the phase and amplitude of the HRTF and measured localisation performance along with subjective judgements from a panel of listeners. The results suggested that the tuning had indeed helped reduce coloration and had improved localisation accuracy. The improvement in localisation accuracy was somewhat limited to the region behind the listener with the frontal region showing only modest improvements. Tan and Gan (1998) showed that front-back confusions can be reduced if the listener is passed through a chain of psychoacoustic judgements, adjusting the magnitude spectrum of a selected

HRTF across specified frequency bands. Finally, a genetic algorithm has been used by Runkle et al. (2000), along with a tuning of the HRTFs in order to generate an effective and customised rendering.

5.2.2.2 *Frequency scaling of HRTFs*

Another technique, which involves using non-individualised HRTFs and having them personalised for the listener in the form of an adaptation, is that proposed by Middlebrooks (1999b), termed *frequency scaling*. HRTFs are adjusted, or shifted, in the frequency domain in order to produce a set of HRTFs that better resemble those of the listener by aligning peaks and notches of the chosen HRTF to where they might exist for the listener. The HRTFs are shifted by a calculated optimal scale factor for a particular listener, which can be correlated to some particular morphological feature such as the size of the head. This method will be the focus of a study in chapter 8 in which the perceptual salience of the different spectral cues embedded in the HRTF is analysed.

The process for calculating an optimal frequency scale factor, described in detail by Middlebrooks (1999b) and in section 8.2.3.3, involves calculating the disparity between two HRTFs from a pair of subjects by scaling them both upward and downward in the frequency domain in steps that are a fraction of an octave. For each shift and for each position in space, a metric called the Inter-Subject Spectral Difference (ISSD) is calculated. The frequency scale factor that minimises the ISSD across all positions in space is then chosen as the optimal frequency scale factor between any pair of subjects. The calculated frequency scale factors can then be perceptually validated, and then correlated to a dimension of listener's morphology so that they can be predicted and used as a step in the process for producing personalised HRTFs. Middlebrooks (1999a) has shown that there are significant improvements in localisation accuracy when HRTFs have been scaled using the optimal frequency scale factor, and that the scale factors can be correlated to pinna-cavity heights and head widths.

5.2.2.3 *Selection of HRTFs from a database*

Following from the previous section, which has shown that non-individualised HRTFs can be modified and adapted to produce an improved experience in VAS in terms of localisation accuracy, another method is detailed in this section that will be a focus of the studies presented in chapters 8 and 9. This technique involves intelligently selecting HRTFs from a database that contains recordings for a large number of different people, and for this reason is a solution that offers a personalised rather than an individualised HRTF. The assumption is that if HRTFs can be adapted to suit any listener, as with the frequency scaling for example, then with a large enough database one should be able to find an HRTF that corresponds well to the listener's own HRTF, given that the HRTFs in the database are a sample that is representative of the general population. The difficulty with this technique is finding a method for selecting the optimal HRTFs for listeners based on some criteria that are easily obtained.

Seeber and Fastl (2003) presented a selection method from a database that involved the subject taking an active role in making perceptual judgements for a number of different HRTFs. A rendering of a broadband sound source in VAS, describing a horizontal trajectory, was judged based on a few key attributes including externalisation, a match between presented and perceived sound source position, whether there was a focused virtual auditory image (i.e. not diffuse), and a reduction in the rate of front-back confusions. This form of listening test proved to generate a selection of HRTFs that minimised localisation errors when tested, and is in fact the basis of a study presented in chapter 7. The test itself took on the order of ten minutes to complete, and relied on the fact that a naïve listener is able to make consistent perceptual judgements of a binaural synthesis (an assumption that is challenged in the mentioned chapter).

Iwaya (2006) has presented a similar methodology, but rather than having the subjects listen and compare a large number of HRTFs at the same time in order to establish an optimal selection, a simplified task was given to subjects in which pairs of HRTFs in the database were compared using a Swiss-style tournament; any HRTF judged as the worst of the pair twice lead to its elimination from the set. The results of this study were limited however to judgements in the horizontal plane, which as explained relies on interaural cues that have been shown to vary less between different HRTFs than the critical spectral cues.

In order to further simplify the selection process, Shimada et al. (1994) generated a reduced number of HRTFs to choose from, beginning with a large database of HRTFs, via a clustering algorithm. The HRTFs themselves were reduced to the order of 16 points from 512 using cepstrum parameters, or the Fourier transform of the logarithm of the HRTF spectrum, which were in turn grouped into eight clusters based on the Euclidean distance between the different HRTFs in a 16-dimensional space. Subjects were then able to more easily select from these distinct classes than if they had to listen to a much larger number of HRTFs that might be very similar to one another. A similar reasoning will be detailed in chapter 7, and has been presented in a study by the author of this body of work (Schönstein and Katz, 2011), in which a subset of distinct HRTFs were chosen for a listening test based on perceptual judgements by a large number of subjects rather than a purely signal-based algorithm.

Other approaches exist in the literature for reducing, or distilling, the HRTF information in databases similar to the cepstrum method previously mentioned. Wightman and Kistler (1993) performed what is known as a Principal Component Analysis (PCA) (a technique for analysing HRTFs that will be used in the study presented in chapter 8) to draw out the most significant patterns (principal components) in the spectrum, along with a set of weights for each principal component, and determined similarity between different HRTFs. Subjects were shown to localise virtual sound sources using HRTFs that were similar to their own, compared to using those that were calculated to be different. Jin et al. (2000), and later other studies (Zeng et al., 2010; Hugeng et al., 2010, 2011), used a very similar approach incorporating a PCA, but extending the procedure to being able to predict an HRTF from a database based on a regression of some basic morphological parameters to the principal component weights. Both these methods postulate that due to the significant role the pinna plays in generating the characteristic peaks and notches of the HRTF, one might be able to draw a link between some key morphological parameters and an optimal HRTF.

The results from the study by Zeng et al. (2010) have shown that this procedure can select an HRTF that improves localisation accuracy and reduces front-back confusions. However, the study only assessed localisation in the horizontal plane and did not measure elevation errors, which are key to any evaluation of HRTFs given the dependence of elevation accuracy on spectral cues; horizontal plane localisation would mostly be used to test interaural cues such as ITDs.

This particular method, using a PCA and regression using the morphology of the listener, will be explored in detail in chapter 9. It was determined by the author of this body of work as an effective solution to the problem of HRTF individualisation as it does not require expensive equipment, nor a perceptual task by the listener; the process can be automated and requires only a few measurements of the ear, possibly obtained via a photograph.

Finally, Zotkin et al. (2003) have shown that HRTF data itself can be removed completely from the selection criteria, using only morphological parameters and the similarity of these dimensions from one listener to another. HRTFs were selected for a particular listener based solely on how well a set of morphological parameter dimensions corresponded on another subject's dimensions in a database. The findings from this study however were not as convincing due to large localisation errors in elevation for some subjects, and as stated by the study's author, require further perceptual validation.

5.2.2.4 *Learning to use non-individualised HRTFs*

A completely different angle to take for attacking the problem of individualised HRTFs, is to consider the ability of the auditory system to learn to use new acoustic filters for interpreting the acoustic environment. The plasticity of the auditory system is something that comes as no surprise due to the fact that the humans must have an ongoing spatial calibration as they grow and the geometry of the body changes and by consequence the HRTF. In addition, we as humans would be ill equipped if every time our pinna was damaged or we have a haircut the auditory system were unable to adapt to the changes. It is thought that the auditory system relies heavily on the visual system to calibrate the acoustic localisation process by providing accurate spatial feedback.

Hofman et al. (1998) showed that the auditory system can indeed relearn spectral cues to sound source location. The study monitored localisation accuracy for a number of subjects after placing moulds inside the ears of a number of subjects, which disturbed spectral cues and lead to an initial increase in localisation errors. After a few weeks, subjects were able to accurately localise sound sources to a level of performance equivalent to when they were tested without moulds in their ears. Interestingly, the learning of new HRTFs did not disrupt localisation accuracy using the subjects' normal pinnae (i.e. without the moulds). The learning of new HRTFs still needs to be tested in terms of the persistence of the learning effects, nevertheless it offers a novel approach to the HRTF individualisation problem. The main disadvantage of this technique, particularly with respect to its application in the consumer market, is the potential for a poor quality of renderings in VAS in the initial learning phase.

5.2.3 *Not requiring HRTF measurements*

5.2.3.1 *Acoustic modelling*

The following solutions in this section do not require any acoustic measurements, which is an attractive feature if the rendering of sound sources in VAS is perceptually similar to those that require measurements. Acoustic modelling is a technique that uses solely the geometrical data of the head and pinna of a listener to produce an individualised HRTF. One such application has been described by Katz (2001) and Kahana and Nelson (2007) using what is termed the Boundary Element Method (BEM), in which both studies used accurate laser optical scans of the listener's head and pinna to generate mesh models. Once the mesh is created, a computer simulation of the acoustic response is calculated in order to produce HRTFs for a sound source at any desired location in space. Kahana and Nelson (2007) provided a comprehensive signal validation, following on from studies by Katz (2001), of the modelled HRTFs, comparing the results to measurements obtained in the free-field. It was concluded that individualised HRTFs could be produced using the BEM technique.

A major drawback to this solution, demonstrated in both the previously mentioned studies, is the required detail of the mesh used and subsequent computational costs this creates; the study by Katz (2001) was limited to frequencies below 5 kHz and excluded the torso in the mesh due to limits in the computational power available at the time. Kahana and Nelson (2007) were able to include all audible frequencies but had to reduce the mesh to only the head and pinna of the listener, neglecting reflections from torso and other parts of the body, due to computational costs. Whilst, this computational barrier becomes less of an issue with the exponential growth of processing power, the method still requires very expensive equipment for generating the scans. One possible avenue for this type of solution is to generate some form of parametric model of the human ear and vary it according to each user, or build up a library of individualised HRTFs using BEM and match corresponding ear scans to those of a listener. More details of these possible solutions are provided as future research options in chapter 9.

Part II

RESEARCH WORK

6 ROLE OF HEADPHONES IN BINAURAL SYNTHESIS

The signals used in a binaural synthesis are produced via a chain of processing techniques as described in the chapter 3. The processed signals are presented to the listener's left and right ears using headphones in an effort to create the illusion of sound sources in space. An often overlooked aspect of this illusion is the transmission of the signal from the headphone to the level of the tympanic membrane for auditory processing. The interaction between the outer ear and the signal from the transducer of a headphone, along with the headphone's own spectral characteristics, can alter the rendering of a binaural synthesis to the point where it may introduce artefacts that can diminish the overall effectiveness or realism of the Virtual Auditory Space (VAS). If these interactions are well understood, they can be compensated for to some extent, or equalised, in the signal processing chain. This compensation can even work to improve some of the physical limitations of a chosen headphone for binaural synthesis. The purpose of this chapter is to elaborate on and study the effect of headphone choice and equalisation with respect to binaural synthesis.

6.1 BACKGROUND

A general explanation of the influence of headphone choice, spectral characteristics, and equalisation, on renderings in VAS has been provided previously in section 3.3. With respect to the current study there exists five broad categories of headphones: supra-aural, in which the headphone rests on the ear, circumaural, in which the headphone completely covers the ear without making contact with it (i.e. contact only with the head), intra-aural, in which the headphone is small enough to be placed inside the ear canal of the listener, open-canal intra-aural, which is exactly the same as the intra-aural except that the headphone sits just outside the entrance of the ear canal so that it is not blocked, and bone conduction, in which vibrations up against the bones connected to the inner ear transmit sounds to the tympanic membrane. Table 1 shows some examples of these types of headphones specific to this study. Each of these headphones have differences in the way the signal is presented to the listener's ear depending on their functional requirements, which causes them to have signature frequency responses according to make and model. How these differences influence the localisation of sound sources in VAS is one of the main topics of this study.

The other focus of this study relates to methods of compensating for headphone characteristics. When headphones are used in a binaural synthesis there is also the interaction between the outer ear and the signal being presented, as previously described in section 3.3. This interaction resembles in some manner the interactions between sound waves from a point source

ID	IMAGE	TYPE	MANUFACT./MODEL
OC		Open circumaural	Sennheiser/HD570
CC		Closed circumaural	Sennheiser/HD265
iPod		Intra-aural	Apple/iPod 5th Gen.
TI		Tube intra-aural	Telex/CES-1
ER·2		Tube intra-aural	Etymōtic Research/ER·2
BC hi		Bone conduction	Oiido/—
BC low		Bone conduction	Vonia/EZ-80P/S20

Table 1: Table of the eight different headphones used in this study. Each headphone's abbreviated name or ID, image, type, manufacturer, and model is displayed. Note that there is a headphone *TI op.* that is the exact same headphone as *TI* just that it sits at the entrance of the ear canal without blocking it.

in space and a listener's auditory periphery, described by the Head-Related Transfer Function (HRTF). There is a fundamental difference however with respect to the HRTF, in the sense that the Headphone Transfer Function (HpTF) is non-directional, meaning that it will not vary for virtual sound sources presented at different locations. In addition, the headphone membrane acts as an excitation in a resonator cavity, and a modal membrane at higher frequencies, as opposed to a point source at a distance, as is the case for the HRTF. The resonances inside the cavity would mostly be due to the volume and geometry of the cavity, impedance conditions, and ear canal, for frequencies above approximately 500 Hz; the cavity is too small to produce any modifying resonances for lower frequencies.

Despite their differences, the frequency response of a headphone and its HpTF has been suggested to negatively impact renderings in VAS due to their influence on the HRTF (Pralong and Carlile, 1996; Wightman and Kistler, 2005). In order to counter these adverse effects, an equalisation can be used to effectively remove any spectral colouration produced by the headphone itself or the interaction of the signal with the outer ear. This equalisation is achieved by imposing an inverse filter on the stimulus along the signal processing chain. The effectiveness of a particular type of headphone equalisation, aimed to compensate for a headphone's frequency response (but not the HpTF), is central to the current study presented in this chapter.

6.2 method

The following section details a study performed by the author testing a variety of different types of headphones on the consumer market. These headphones ranged from bone conduction, to a more common intra-aural headphone being used commercially; the iPod headphone, to expensive intra-aural headphones that are designed for research applications. Each of the headphone types were tested in terms of one subject's (the author of this body of work) ability to accurately determine the position of a sound source rendered in VAS. The different headphone types were tested with the many applications of binaural synthesis in mind, and to this end the study aimed to validate the headphones and determine which of them might be amenable to consumer applications. In a similar vein, non-individualised HRTFs were used in this study as they are the best solution for consumer markets (given the current technologies available) since their use avoids the laborious task of generating individualised HRTFs (see chapter 5).

The effectiveness of a headphone equalisation to improve localisation accuracy of virtual sound sources, for each of the different headphone types, was also assessed. The type of equalisation used was non-individualised, meaning it did not compensate for the combined effects of the headphone and subject pinna, but rather only for the specific headphone characteristics (see section 3.3). This was tested in order to establish whether this type of equalisation would be a viable option for consumer applications in an effort to enhance binaural syntheses. A non-individualised equalisation is a significantly less involved procedure compared to an individualised equalisation due to the fact that the morphology of the individual listener does not need to be taken into account.

6.2.1 *Experimental procedure*

The results for this study were obtained for one subject, the author of this body of work, in terms of localisation accuracy for a sound source presented over headphones in VAS. The subject used a non-individualised HRTF for the localisation task, which was selected based on results from a listening test in which a number of HRTFs were judged (section 6.2.4). The subject used a number of different headphones, with or without equalisation, for a set of randomly ordered locations using a virtual sound source.

6.2.2 *Stimulus duration and level*

A broadband Gaussian noise was used as the stimulus for the localisation task in this study. The duration of the broadband stimulus was 130 ms, with a 5 ms onset and offset Hanning ramp, which was deemed an appropriate length given that localisation performance peaks and plateaus for stimulus length of 80 ms and above (Vliegen and Van Opstal, 2004; Hofman and Van Opstal, 1998), as described previously in section 4.3. A measure called sensation level was used to determine the level of the stimulus, in which the intensity was calculated in decibels above the subject's detection threshold. The threshold was determined by presenting the stimulus at ear level directivity in front of the listener and controlling the signal level. The sensation level threshold was calculated for the subject by beginning at a level that was inaudible and incrementally increasing the level 1 dB at a time. Once the subject had indicated that the signal was audible, the level was decreased until inaudible, and then increased and so on, until a clear threshold was reached. Once the threshold was found, a sensation level of 50 dB (50 dB above the threshold) was used for the all presented stimuli over headphones. This level was slightly higher than the optimal sensation level (between 40 to 45 dB) suggested by authors assessing the role of stimulus level on localisation accuracy (Macpherson and Middlebrooks, 2000; Vliegen and Van Opstal, 2004), due to the fact that the detection thresholds in this study were not calculated to the level of accuracy used in the cited studies (i.e. not using a two-interval, two-alternative, forced-choice task; see Levitt, 1971) , and probably incorporated some measurement error. In addition, a sensation level of 45 dB was perceived by the subject to be at a level that was difficult to hear, most probably due to the mentioned measurement error. The current study was not specifically assessing the effect of stimulus level, and therefore the main objective was to avoid presenting the stimulus at levels that may induce interfering auditory mechanisms such as the acoustic reflex. To this end the threshold detection method served its purpose and a sensation level of 50 dB, within the bounds of measurement error, was deemed acceptable.

6.2.3 *LISTEN HRTF database*

The HRTF recordings used to render the broadband stimulus in VAS were taken from the online public database of the LISTEN project[1]. The methods employed for measuring the transfer functions of the subjects have been described in detail on the project website. The microphones were positioned in the subjects' ears using the blocked-meatus method (see section 3.1) and recorded the sound source using a resolution in space of 15° in azimuth and elevation, with a slightly reduced resolution at high elevations. There were a total of 187 measurements performed for each of the subjects' left and right ears, each recording was 8,192 points long and recorded at a sampling rate of 44.1 kHz. Unless otherwise stated all signal processing was performed using the Matlab software environment using a library of code developed by the LISTEN project team.

The subject's Interaural Time Differences (ITDs) were preserved by separating HRTFs in the database into minimum phase and ITD; individual ITDs were synthesized for the subject based on a principal component analysis of the estimated ITDs over the entire LISTEN database using the method of Maximum Inter-Aural Cross-Correlation (MaxIACC) and a linear regression to head morphological parameters. Head circumference was chosen as the morphological parameter for calculation of ITD as it provided the most stable regression and the least error with regards to measurement repetition variation. The estimated ITDs for the subject, calculated via a measurement of head circumference, were combined with the different minimum phase spectral components so that each HRTF was evaluated with the individual subject's own ITDs. This equates to, as a simulation, changing the ears of each subject while keeping the same geometry of the head.

The HRTF recordings used in the current study were raw recordings, and had not been free-field or diffuse-field equalised. There are benefits, especially in terms of the externalisation effect, associated with using raw HRTF recordings for a binaural synthesis (see section 3.5.3). The threshold was calculated separately for each headphone used in the study.

6.2.4 *HRTF selection*

There were HRTFs for a total of 45 subjects in the database. A listening test was used to select an optimal set of HRTFs to be used for the localisation task in the current study. Again, the software environment Matlab was used and the test was developed by the LISTEN project. The signal used for the test was a binaural synthesis of a broadband white noise of 0.23 seconds, modelled using a Hanning window. The test signal was presented at fixed positions along two paths presented in sequence:

1. A circle in the horizontal plane (elevation = 0°) in increments of 30°. The path started at 0° azimuth and 0° elevation and made two rotations (duration 6 seconds).

[1] See http://recherche.ircam.fr/equipes/salles/listen for details.

2. An arc in the median plane (azimuth = 0°) from elevation −45° at the front to −45° at the back in increments of 15°. The path started at the front elevation of −45°, and the elevation was varied to the rear and then made to come back the same route to the starting position (duration 9 seconds).

A slider, via a graphical user interface presented on a computer, was used for judging each of the subject HRTFs in the database in terms of how well the VAS rendering resembled the described trajectories. A judgement at the far left end of the slider represented a poor correlation between the trajectory heard and that described, and a judgement at the far right represented an excellent correlation. Judgements in the listening test were analogous to responses that are made by subjects during a localisation task, in the sense that the position of the virtual sound source was the most salient attribute. Yet judgements were also based on other features of the binaural synthesis such as the extent of externalisation. The listening test judgements required the subject to make a spatially weighted response, averaged over the whole trajectory, which is a demanding task due timbral variations over time. This type of listening test often leads to poor reproducibility (see Zielinski et al., 2008, for a review, and Schönstein and Katz, 2010a, for a study of response variance), an issue that is the focus of the next chapter and will be explored in detail.

The subject was able to play the test stimulus as many times as desired, and in any order. Judgements were made for all the HRTFs in the database using the first trajectory before making judgements for the second trajectory. The test duration was about 35 minutes, and upon its completion a mean score for both trajectories was used to determine the optimal subject HRTF, which was calculated to be number 1032 in the LISTEN database. Using the described procedure provides a somewhat personalised HRTF in the sense that it was chosen as the best from the database in terms of rendering sound sources in VAS.

6.2.5 *Headphone types used*

The following experiments were then completed by the author. A total of eight different headphones were compared in the current experiment. Table 1 shows an image for each headphone used including its abbreviated name (or ID), type, manufacturer, and model. The two circumaural headphones differed in terms of the outer casing that contained the transducer; one had a perforated casing meaning that the transducer was open to environmental sounds (open circumaural; OC condition) and the other had a casing that created an almost airtight seal around the ear meaning that the space between the ear and transducer was closed (closed circumaural; CC condition). As previously discussed in section 3.3, these differences greatly affect the resonances produced within the headphones and thus have a bearing on the frequency response of the headphone.

There were four different types of intra-aural headphones used in the study. The first of these was a commonly used intra-aural headphone on the consumer market: the iPod headphone (*iPod* condition). The remaining three intra-aural headphones, can be better described as tube intra-aural headphones, and varied significantly in their market value. These headphones are different from the iPod headphone as their transducer is not located at the entrance of the ear canal where the headphone is inserted into the ear; rather it is at some location away from the ear with a tube being used to transmit the rarefactions and compressions of the sound to the ear canal. The first tube intra-aural headphone (*TI* condition) is the least expensive model and used in the consumer market by security guards or television commentators. The second tube intra-aural headphone (*TI op.* condition; a free-to-ear type) was the same model headphone as the *TI* except that the end of the tube was not inserted into the ear canal using a rubber tip, effectively blocking it, but rather was positioned at the entrance of the ear canal without blocking it using a fixture that wrapped around the ear. This meant that the *TI op.* condition left the ear canal unobscured so that sounds from the surrounding environment could be heard. This is desirable for applications in which a high situational awareness is needed, such as augmented reality or tactical situations commonly found in the military. The visually impaired rely heavily on auditory environmental cues and therefore are also in need of headphones that leave the ear canal open. The third tube intra-aural headphone (*ER·2* condition) was a significantly more expensive model than any other used in the study. It is a headphone specifically designed for research purposes, and was used as a reference headphone in the study.

The final class of headphones used in this study was that of bone conduction headphone. In the study there were two commercially available bone conduction headphones, *BC hi* and *BC low* conditions, the first being more highly priced in the market than the second. Bone conduction headphones are useful for the same reasons previously mentioned with respect to being able to perceive ambient sounds in the environment whilst listening over headphones, and have been shown to be effective for binaural synthesis (Walker et al., 2007).

6.2.6 *Headphone frequency responses*

Each headphone was characterised in terms of how its transducer reproduced frequencies over the audible range. The headphone Frequency Response (FR) described in this study differs from the previously described HpTF in section 6.1. The most significant difference related to how the headphone is coupled to the microphone for recording. For the HpTF, headphones are mounted to a human listener or a dummy head, and the microphone is placed at the entrance of the ear canal much the same way HRTFs are recorded. The HpTF encompasses a chain of transfer functions from the recording microphone, the headphone, and the impedance of the ear canal (Møller et al., 1995a). An accurate description of the combined effect of the headphone and pinna, known as the HpTF, can only be described if all these elements are known. The FR, in contrast, does not take into account the effect of the pinna and ear canal resonance and describes purely the frequency response of the headphone transducer almost as if it was a speaker in space. For this study the frequency transfer characteristics of the microphone used to record the FRs were not taken into account; the microphone was known to have a flat response across all frequencies.

The FRs for each of the headphones were measured by placing each headphone on a metal plate as shown in figure 7, held in place using a thin strip of elastic rubber, and recording a stimulus played over the headphone with a microphone that was flush to the metal plate. A 2 cm piece of foam was inserted between the microphone and the headphone in order to avoid a seal between headphone and metal plate. This was performed due to the fact that circumaural headphones rarely form a perfectly airtight seal around the listener's ears, as was the case for headphones being placed up against a metal plate in this study. The stimulus played over the headphones was a 500 ms sweep signal ranging from 60 Hz to 20 kHz. Using a sweep signal instead of a broadband Gaussian signal has several advantages when measuring the frequency response of a transducer, whether it be for a headphone or speaker, due to the fact that not all the frequencies are produced at once. Measurements using a sweep are considerably less vulnerable to distortion and time variance. The sweep can thus be fed to the loudspeaker with considerably more power without introducing artefacts in the acquired impulse response (Farina, 2000; Müller and Massarani, 2001).

The measured impulse response recorded by the microphone was then converted via a fast-fourrier transform into the frequency domain and the real component was extracted in order to analyse the magnitude response of the headphones. Since an artificial mastoid, which consists of a mechanical simulation of the head, is needed for measuring frequency responses of bone conduction headphones, FRs for the two used in this study were obtained directly from the manufacturer (see figure 10). The FRs for these two bone conduction headphones are more akin to the described HpTF as they involve the effects of the human head, yet it is unsure whether the microphone characteristics used on the artificial mastoid have been taken into account. The type of microphone, make and model of the artificial mastoid used by the manufacturer is not known. Despite these differences, the transfer functions of the bone conduction headphones will be referred to as FRs in this study for the sake of continuity.

Figure 8 shows the FRs for each of the headphones along with the amount of variance observed in the magnitude of the spectrum, displayed in decibels, for 10 repositionings of the headphones. The difference in the variance between the FRs with and without equalisation was approximately 28 dB^2 for the open circumaural (*OC*) headphone, 26 dB^2 for the closed circumaural (*CC*) headphone, 48 dB^2 for the iPod headphone, and 19 dB^2 for the tube intra-aural (*TI*) headphone. The results demonstrate significant deviations from a flat spectrum, and that these colourations vary from headphone to headphone. This is not any type of manufacturing error from the point of view of the headphone makers; headphones are usually specifically designed to

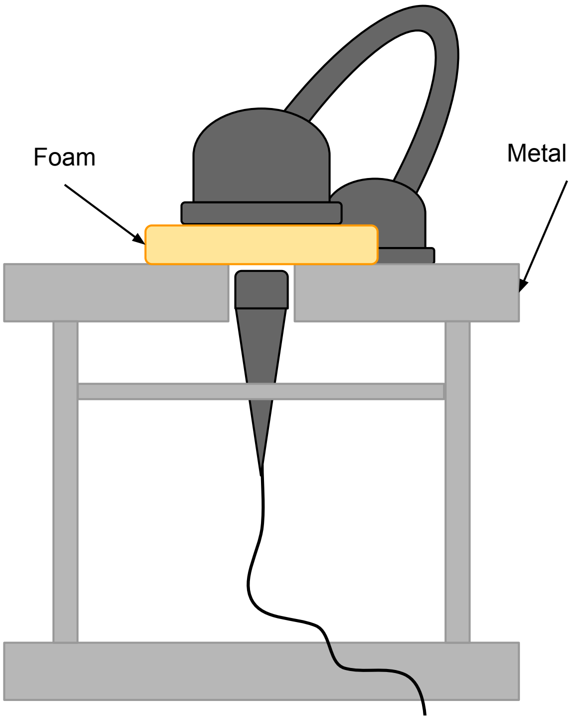


Figure 7: Diagram of the apparatus used to record the headphone frequency responses.

have a frequency response that is not flat at the tympanic membrane in order to produce some sort of timbre signature. Furthermore, when coupled to a human ear the frequency response, or HpTF, might be very different and possibly flat for any of the headphones tested due to the interaction of the sound waves inside the headphone cavity when compared to measurements performed on a metal plate.

One way to gauge the potential differences between the HpTF of each headphone and the FRs measured in this study is to analyse the frequency response for the reference headphone *ER*·2. It is known that this headphone is rigorously tested by the manufacturer to have a flat HpTF at the level of the tympanic membrane, as it is a widely used research tool. Thus, any deviations from the flat frequency response for this headphone in the current study might be due to the fact that the coupler is not a human ear or calibrated dummy head, and to a lesser degree due to the transfer function characteristics of the microphone used. Assuming this is true, it can be seen that there would be significant boost in gain in the region from 0 to 6 kHz, which corresponds well with the the ear canal resonance observed using measurements with rubber replicas pinna, concha, and auditory meatus (Shaw and Teranishi, 1968). There might also exist a boost in gain from approximately 10 kHz, yet the frequency response given by the manufacturer for the *ER*·2 headphones describes a flat frequency response for frequencies up to 16 kHz.

With respect to the variance in the FRs, it is evident that only small differences are observed between headphone repositionings. One exception may be the closed circumauaral headphone (CC) that showed some variance in the

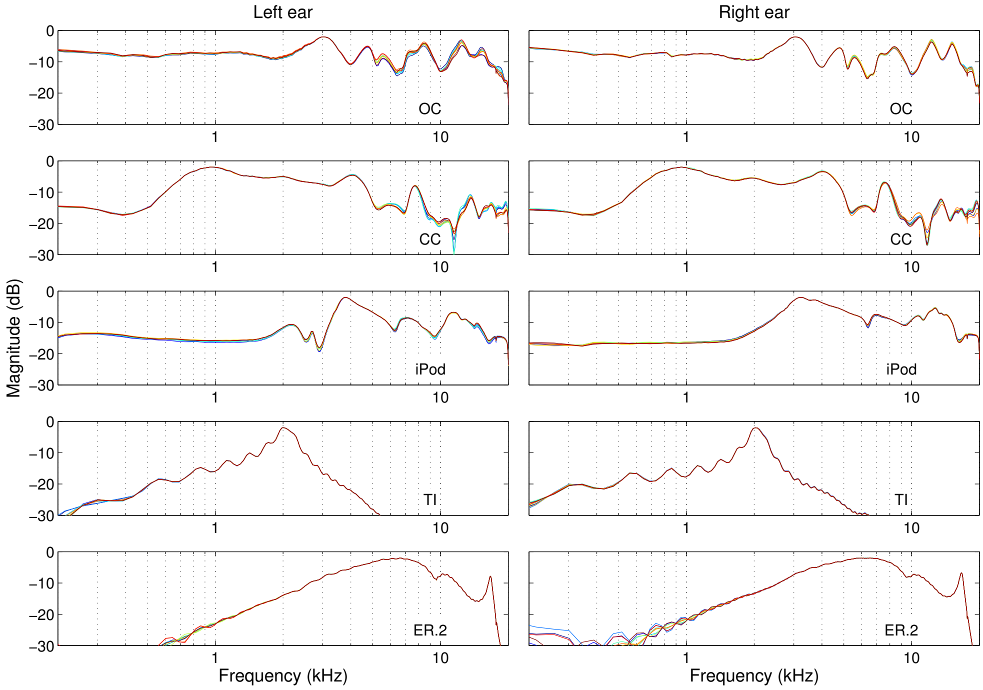


Figure 8: Headphone frequency responses for all the headphones used (except for the bone conduction) for the left and right ear. The magnitude responses for 10 replacements is shown using different colours on the plots. Only a small degree of variation is observed.

depth and position of notches above frequencies of approximately 10 kHz. These results are in contrast to findings in previous studies, as described in section 3.3, which demonstrated a significant amount of variability in the characterisation of HpTFs when removing and replacing headphones for repeated measurements. This can be explained by the fact the transfer function of the pinna, which produces peaks and notches in the spectrum, was not incorporated into the results in this study; a metal plate was used for the measurements instead of a dummy head or human listener. The introduction of the pinna in the HpTF creates a non-directional colouration of the spectrum in the form of peaks and notches due to resonances and reflections in the convolutions of the ear. However, as previously discussed at the beginning of the chapter, the resonances inside the cavity of a headphone with the pinna would mostly be due to the volume and geometry of the space, impedance conditions, and ear canal, for frequencies above approximately 500 Hz. The coloration imposed by the pinna itself would thus have less of an influence.

6.2.7 *Headphone equalisation*

The deviations in the FRs measured in this study change the spectrum of HRTFs when using a binaural synthesis. This might disturb spectral cues embodied in the HRTF used by the auditory system in a binaural synthesis. They resemble very much the HpTFs measured in studies by Møller et al. (1995a) and Wightman and Kistler (2005) in which the authors have argued that since features of the HpTFs are similar to those found in subject HRTFs and introduce a significant colouration of the acoustical stimulus, they may possibly effect a binaural synthesis, particularly in terms of a listener's ability to determine the position of virtual sound sources (Pralong and Carlile, 1996). However due to the fact that the FR acts as a non-directional filter and for the reasons discussed in the previous section, it is thought that the effect of the FR will be minimal unless it is drastically distorting particular frequency ranges or causing them not to be audible at all. It should be noted that despite the similarity between the spectra of HpTFs and the FRs measured in this study, they differ in a fundamental manner; the HpTF incorporates the filtering effects of the pinna whilst the FR does not (see section 3.3.3). The pinna imposes an individualised coloration of the spectrum as does the ear canal resonance, whilst the HpTF is specific to the headphone used and not the listener.

The deviations in the spectrum represented in the FR (measured on humans or dummy heads), in the form of notches and peaks, are caused by the headphone transducer characteristics and resonances formed in the space between the ear canal and headphone. It was the purpose of this study, along with a comparison of the different types of headphones used for binaural synthesis, to test the effect of headphone equalisation. The equalisation aimed to eliminate the effects of the FR and produce a spectrum from the transducer of the headphones that was as flat as possible. The effectiveness of the equalisation can be assessed on a logarithmic scale that is perceptually relevant, as presented in figure 8.

The equalisation was performed by processing the broadband stimulus with the inverse of the filtering effects imposed by the FR for playback in VAS. The mean of the previously described magnitude response, obtained from the FR, for the 10 measured repositionings was assumed to best characterise each headphone. The only exception was for the bone conduction headphones in which the single frequency response provided by the manufacturer was used. The inverse of the magnitude values was then converted to a scale that sampled higher frequencies at an increased resolution, somewhat mimicking the frequency resolution of the basilar membrane and providing detail along the spectrum in a perceptually relevant manner. This varied sampling rate was achieved by an infinite impulse response digital filter designed using a least-squares fit to the scaled inverse of the mean FR. The coefficients from the filter were then converted to the equivalent second-order section form and used to filter the broadband stimulus as a cascading series. This filtered stimulus was therefore adapted to the characteristics of each headphone in an effort to produce a flat frequency response from the transducer. For each presentation of the broadband stimulus in the localisation test, the filtered or unfiltered signal was convolved with the HRTF for the desired position, corresponding to the condition with and without headphone equalisation respectively.

Figure 9 shows the spectra of a broadband stimulus recorded with and without equalisation. An exemplary flat frequency response for the equalised

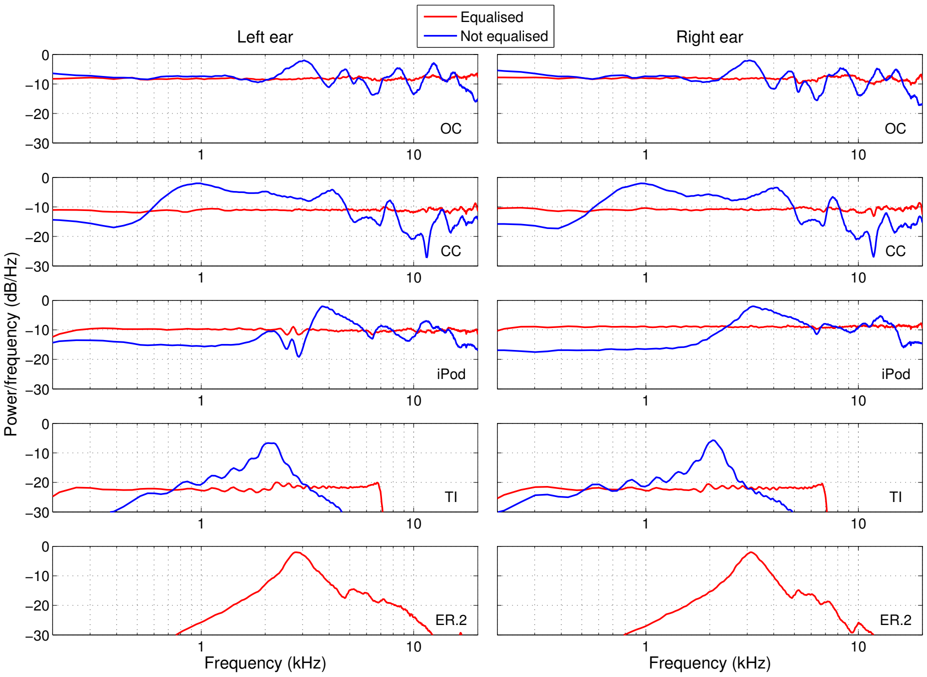


Figure 9: Exemplary recordings of a broadband gaussian noise using the headphone equalisation (in red) for each of the headphones tested, except the bone conduction headphones. The recordings for the broadband stimulus using headphones without equalisation is shown in blue.

condition is seen for each headphone, except for the *ER*·2 headphones. *TI* has a drop in the magnitude at 7 kHz due to the physical limitations of the headphone not being able to produce a signal at high frequencies. The mean reduction in variance between the FRs with and without equalisation for the left and right ear was approximately 28 and 26 dB^2 for the circumaural headphones *OC* and *CC* respectively. The *iPod* headphone showed a reduction in variance of approximately 48 dB^2. The *TI* headphone showed a reduction of approximately 19 dB^2.

Figure 10 shows the frequency response of the two bone conduction headphones, provided by the manufacturers, along with the inverse magnitude response and corresponding inverse filter used for equalisation.

6.2.8 *Localisation task*

The author of this body of work, a male aged 25 years old, was the only subject that participated in the study, and was shown to have clinically normal hearing. The localisation tests were conducted in a sound dampened room with foam padded walls. The audio programming environment Max (version 4.6.3) was used for the localisation task. A patch previously created at the LIMSI research centre was modified and used to run the localisation tasks and produce the audio signals for the headphones.

The subject's head and a pointer were continuously monitored in terms of its position in space using two receivers and an electromagnetic tracking

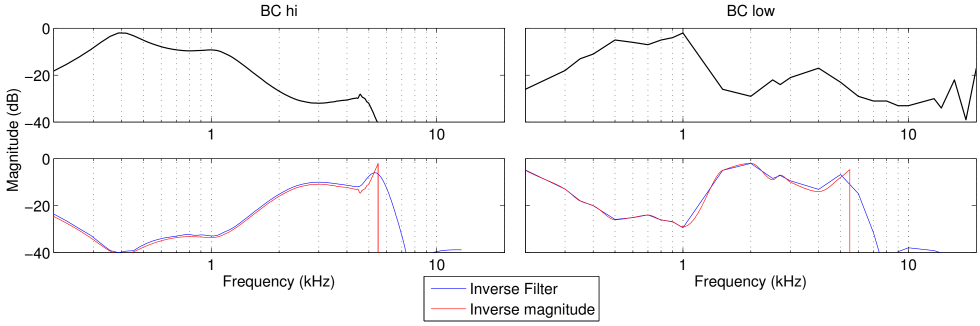


Figure 10: The headphone frequency response for the two bone conduction headphones used in this study, provided by the manufacturer, shown in the top panels. The inverse of the frequency response (in red), along with the modelled inverse filter (in blue) is shown in the lower panels.

system (Ascension Flock of Birds®). The head receiver was attached to a headband worn by the subject, and the hand receiver, equipped with a plastic tip to serve as the pointer, was held by the subject. The subject's task was to determine the location of virtual sound sources, presented using the eight different headphones with and without equalisation. The pointer was to be positioned at the judged location of the virtual sound source and the subject's response recorded by pressing down on a MIDI foot pedal. The position of the pointer was always calculated with respect to the head receiver position the moment the sound was played to the listener.

The procedure for each localisation test involved an initial calibration of the tracking system in which the orientation of the head receiver was adjusted for. The subject was made to stand in a specified position in the room, and the pointer was used to register the position of the tip of the subject's nose and entrance to the ear canal of both ears. Using the coordinates of these three points in space relative to the position of the head receiver it was possible to calculate the centre of the subject's head along with an orientation of the coordinate system for the tracker. The plane connecting the ears and nose was used as the horizontal plane, and the position with coordinates 0° azimuth and 0° elevation (i.e. directly in front of the subject) was taken as the point perpendicular to the interaural axis (i.e. the axis that joins the left and right ear canals) and equidistant to the left and right ear, along the horizontal plane. The pointer was held at a full arm's length when registering responses, and therefore the judged distance of the virtual sound source varied slightly for each target position tested. For this reason, and due to the fact that the presentation of all the target virtual sound sources were at an equal distance, the subject's judged distance was not recorded and responses were always expressed in terms of their direction, using only azimuth and elevation angles. Thus, for each target position, represented by a pair of target azimuth and elevation values, a corresponding pair of azimuth and elevation response values were recorded.

There were a total of 24 positions used in the localisation tasks. Using the hoop coordinate system (see section 2.1), there were four elevations tested: −30°, 0°, 30°, and 60°. There were six azimuths tested: −135°, −75°, −15°, 45°, 105°, and 165°. This set of positions was chosen randomly from three distinct subsets of possible azimuths and elevations so that the subject did not know the exact positions of the target locations. As mentioned, all target locations were of equal distance from the listener describing positions on an imaginary sphere with radius 1.95 m in VAS. This distance corresponded to that used during the measurements of the HRTFs from the LISTEN database. For each localisation test, in which one particular headphone was used, the subject listened to a total of 72 broadband stimuli (i.e. three repeats of each of the 24 positions tested). The position of the stimulus was randomly generated. Each headphone (except for the *ER*·2 reference headphone) was tested with and without the inverse filter, thus there were a total of 15 headphone conditions tested. Each localisation test of 72 positions was selected randomly and repeated twice (144 observations for each headphone). This produced a total of six repeats of each of the 24 positions tested for each of the 15 headphone conditions. The tests were performed over two days with a pause between every test. There were at least five trials performed for each of the headphones before the results were recorded, which served as ample training in the localisation task. The order of the headphones tested was randomised.

6.2.9 *Measurement of localisation accuracy*

In this study two coordinate systems were used to describe the positions of the target broadband stimulus rendered in VAS: hoop coordinates and lateral/polar coordinates (see Leong and Carlile, 1998, and section 2.1 for details).

The lateral/polar coordinate system is a perceptually relevant manner to code for the position of sources in space due to the fact that it mirrors aspects of the mechanisms used by the auditory system for the localisation of sound sources. The lateral angle of a sound source in space is determined for the most part using interaural cues and describes any position on a circle, given that the distance to the sound source is known. The polar angle resolves the exact position of the sound source on this circle and is coded by the auditory system using spectral cues embedded in the HRTF. Hoop coordinates, using azimuth and elevation, are useful for describing locations in space given that there are more intuitive (see Carlile, 1996, for discussion). The lateral/polar angle coordinate system is better suited to describing response errors for the reasons described above.

Localisation accuracy was assessed by measuring the disparity between the target position of the broadband stimulus (the sound source in VAS) and the subject's response, or indicated position, registered using the pointer. The simplest expression of this difference is the spherical angle error, represented by the angle of the arc joining the two target and response vectors. Each of these vectors can be represented as originating at the centre of the VAS (at the centre of the subject's head) with direction and magnitude towards the position of the target and subject response (see Carlile et al., 1997, for an example). An effective global measure of localisation errors across all positions tested was taken as the mean of the spherical angle errors for a particular headphone condition.

A more detailed measure of localisation accuracy is to break down the spherical angle error into the two components: difference between the lateral and polar angle for target and response locations. These measures are perceptually relevant and as previously described draw on different processes along the auditory pathway. Mean lateral and polar angle errors were used to compare localisation accuracy across the 15 headphone conditions.

The rate of front-back confusions (see section 3.5.1) was also measured in this study. Front-back confusions are by definition any response for a target location that is in the opposite hemisphere, mirrored across the plane that contains the interaural axis. Front-back confusions represent a distinct type of localisation error and for this reason have been dealt with separately from the relatively smaller response errors. However, errors on or very close to the interaural axis should not be considered as front-back confusions due to the fact that although they represent an error that strictly crosses the separation between the front and back hemispheres, their magnitude suggests that the error is not related to the issue of resolving a cone of confusion for sound sources originating from either in front or behind the listener. For these reasons, front-back confusions were defined as any localisation error that had a response in the opposite hemisphere and was not within a region close to the interaural axis. A region defined by all positions within 10° (lateral angle) was arbitrarily used as it represented a relatively small portion of the VAS and has been used in similar studies (Leong and Carlile, 1998). In addition, a front-back confusion, as defined in this study, was allowed to cross over the midline by only 10°. This equates to not defining localisation errors as front-back confusions, if, despite having a response that is in the opposite hemisphere, are clearly on the opposite side of the midline (i.e. response the left side of the listener for a target location on the right side, and vice versa).

There are two methods for dealing with front-back confusions in the analysis of localisation errors: they are either removed completely from the analysis (Jin et al., 2004; Carlile et al., 1997), or resolved and incorporated into the analysis (Wenzel et al., 1993; Makous and Middlebrooks, 1990; Wightman and Kistler, 1989b). Resolving the errors is achieved by mirroring the responses across to the opposite hemisphere, effectively accounting for the error in the azimuthal angle by reducing it, whilst maintaining the response elevation angle. In this study, the recorded responses meeting the previously mentioned definition of a front-back confusion were removed from the calculations of the mean spherical angle error and the global accuracy measure described in the following paragraph as they constituted a different type of localisation error, and their magnitude would disproportionately skew the measures mentioned, and create a bimodal distribution of errors. The percentage of front-back confusions was recorded and compared across the different headphone conditions.

Whilst the mentioned measures are effective descriptors of the magnitude of localisation errors across tested positions for a headphone condition, they do not offer any insight into the generalised localisation accuracy by grouping responses by the target locations tested. To this end, a metric termed the Spherical Correlation Coefficient (SCC) was used, which collapsed the localisation data over the sphere to a value between −1 and +1, ranging respectively from a complete negative and positive correlation between target and response locations (see page 232 Fisher et al., 1987, and Wightman and Kistler, 1989b, for implementation). All target and response positions were expressed as unit length vectors for simplicity in the calculation, reflecting the fact that only the direction of positions in space were analysed.

The first step in the calculation of the SCC was to group subject responses by target position. As detailed previously the subject repeated six randomly ordered localisation judgements for each target position in an effort to build up a robust description of localisation accuracy for tested positions. The repeated responses were pooled to create what is termed as a response centroid, representing the subject's average response for a particular target location. The response centroid was a unit vector with direction computed from the resultant response vector; the vector sum of all the unit length response vectors. The next step in the calculation of the SCC, was to take the direction cosines (i.e. the Cartesian coordinates in a three-dimensional space) of the response centroids Y and corresponding target vectors X and organise them into an $n \times 3$ matrix where n corresponds to the number of positions tested for each of the 15 headphone conditions in the localisation test. The determinant of the summed product of X and Y was then calculated, representing the orientation of the bases of the direction cosines in X with respect to those in Y. The SCC is expressed as:

$$\rho = \frac{\det\left\{ \sum_{i=1}^{n} X_i Y_i \right\}}{\sqrt{\det\left\{ \sum_{i=1}^{n} X_i X_i \right\} \det\left\{ \sum_{i=1}^{n} Y_i Y_i \right\}}} \tag{1}$$

This equation yields a value bound between −1 and +1, which is a measure of how well the Y_i's (the subject response centroids) can be matched with the X_i's (the target vectors) by performing an orthogonal transformation of the Y_i's. The optimal alignment of the target and response matrices X and Y is positive for a rotation, and negative for a reflection, in any axis.

6.3 RESULTS

6.3.1 *Results of localisation tasks*

Most of the analyses of the localisation task responses in this study were performed using spherical plotting and analysis routines provided as part of the publication by Leong and Carlile (1998). The numerical computing software Matlab along with the Spherical Package (SPAK), found on the second author's website[2], was used to draw conclusions on the types of localisation

[2] http://www.physiol.usyd.edu.au/~simonc

errors made by the subject. The results were expressed in terms of the localisation errors, which were broken down into measures that incorporated front-back confusions, such as the lateral and polar angle errors, and measures such as the mean spherical angle error and global measure of fit (the SCC), that did not include responses defined as front-back confusions.

6.3.2 *Lateral angle errors*

An analysis of the localisation results showed that the lateral angle component of response errors was minimal, whilst the polar angle component was not. The individual raw localisation responses, *including* those that were deemed front-back confusions, for all 15 headphone conditions are presented as scatter plots in figure 11. In order to clearly represent the number of responses as a function lateral angle, responses within 10° by 10° grids on the plot have been grouped. The size of the circles in the plot correspond to the number of responses in a particular grid.

The more accurate the subject's responses the more the response angles should cluster on the diagonal line in the scatter plots. The diagonal line represents the relation in which values on the vertical axes are equal to values on the horizontal axes, and thus represents ideal localisation accuracy. The correlation coefficient between target and response angles was calculated for each headphone condition in an effort to quantify the magnitude of the errors, and is displayed on each plot. It was shown that there were strong correlations between the lateral target and response angles. This is an expected result due to the heavy reliance of judged lateral angle on interaural cues, particularly ITDs. The main difference between the headphone conditions was in terms of their frequency response, as demonstrated in figure 8, whilst little variation would be expected between the temporal aspects of the signals reaching the two ears for the binaural synthesis. Despite the fact that the temporal and phase characteristics of the headphones were not measured in this study, it can be inferred from the results that they were not significantly disrupted by any of the headphones and that differences between the arrival times of the envelopes of the signals at the left and right ear were accurately presented. One exception might be the less expensive free-to-ear tube intra-aural headphones without and with equalisation (*TI op.* and *TI op. eq.* respectively). The lateral angle responses for these two headphone conditions showed relatively large lateral angles errors, which translates into lower correlation coefficients.

In addition to the correlation coefficient, a linear fit was performed on the results for each headphone condition producing a slope and intercept value, describing any bias in the data. The coefficient estimates (slope) and constant terms (intercept) are shown in table 2. None of the slope values for the different headphone conditions were below 1 indicating that lateral angles were generally localised slightly further from the midline than the target lateral angle. This bias was most pronounced for the tube insert and bone conduction headphones, with a maximum slope of approximately of 1.32 for the equalised bone conduction low (*BC low eq.*) condition. This type of bias might

CONDITION	SLOPE	INTERCEPT
OC	1.09	-4.18
OC eq.	1.09	-7.07
CC	1.07	-6.59
CC eq.	1.16	0.17
iPod	1.16	-4.25
iPod eq.	1.23	-3.51
TI	1.11	-1.29
TI eq.	1.13	-7.64
TI op.	1.18	9.51
TI op. eq.	1.21	-1.29
BC hi	1.21	-3.46
BC hi eq.	1.18	4.9
BC low	1.24	-2.06
BC low eq.	1.32	-6.95
ER·2	1.08	-10.14

Table 2: Regression coefficient estimates and constant terms for lateral angle response and target angles for each headphone condition.

exist as an artefact of the method being used to register subject responses; when pointing with the hand the subject may have had a tendency to exaggerate his movements towards to sound source effectively overshooting its location. Testing with an increased number of subjects would determine if this was subject specific or not. The intercept values were generally negative, with the largest negative value for the reference headphone condition (*ER·2*) at approximately $-10.14°$. This intercept value represents a global bias in responses and would most probably be due to the measurement system for subject responses or the signal produced via the headphones. The fact that the intercept is negative for most headphones suggests that this global bias is due to the calibration of the subject response equipment. The correlation coefficient is thus the most appropriate measure of accuracy due to the fact that it ignores a global bias resulting from a different intercept value, i.e. no matter the intercept value for response and target angles, the correlation coefficient will be the same.

6.3.3 *Polar angle errors*

The polar angle errors are presented as circular hair plots (see Jin et al., 2004), in which all the responses, including front-back confusions, were collapsed across all lateral angles tested and only the polar angle component displayed (figure 12). Scatter plots are not well suited to displaying polar angle errors

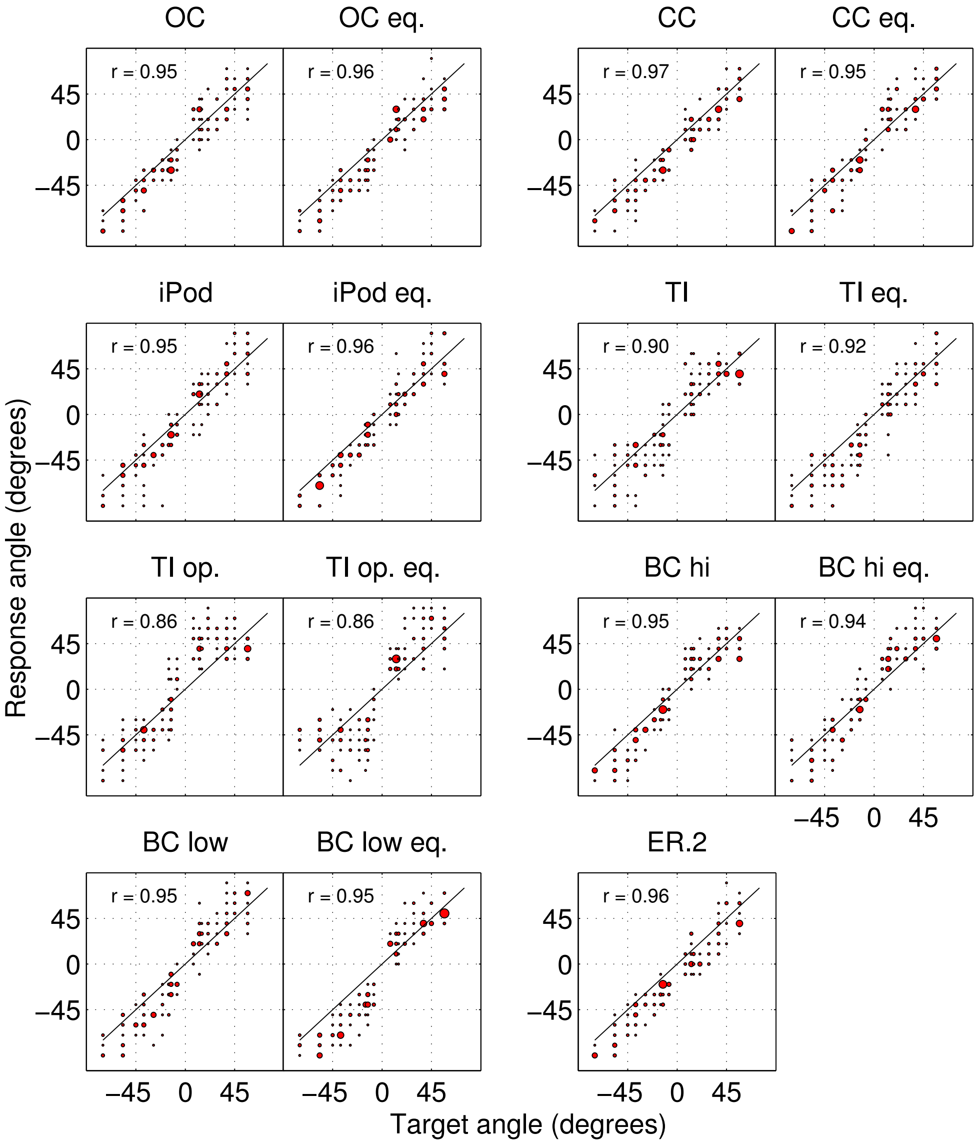


Figure 11: Scatter plots of lateral angle responses for all the different headphones tested. Response are presented on the vertical axes and target positions on the horizontal axis in degrees. The diagonal line corresponds to ideal responses. The size of the circles on the plots corresponds to the number of responses in a particular 10° by 10° grid.

due to the fact that circular data requires special statistical methods (see Fisher, 1995) and has no discontinuity in the range of possible angles. For example, for a fixed lateral angle of say 0°, a polar angle of −170° and 170° are actually two locations that are near one another (i.e. both almost directly behind the listener). Yet this will not be communicated as such using a linear representation such as a scatter plot. Furthermore, it can be seen that an arithmetic mean of 0° is not a correct calculation for these angles as it does not take into account the continuous nature of polar angles at −180° and 180°. This is in contrast to lateral angle, which has clear limits at 90° and -90°.

For each headphone condition and corresponding circular hair plot, a circle was drawn representing the many cone of confusions on which the target was presented from. One can imagine a side-on view of the subject at the centre of the circle as represented in the figure legend with the labels for locations in front, behind, above, and below. The target polar angles were presented as small dots on the circle. Short line segments originating from the dots of the target angles represented the polar angle responses. The direction of the segments can be used to determine localisation accuracy; each segment is directed to the point on the circle where the subject made a response to the corresponding target. That is to say, if each segment were extended out to where it intersected with the circle, the point of intersection would represent the polar angle response location on the cone of confusion. A segment that is at a tangent to the dot thus corresponds to a response that perfectly matches the target. A front-back confusion would be represented by a segment that points towards the opposite hemisphere, indicating a polar angle response on the half of the circle that is directly opposite. Using circular hair plots leads to a visualisation of the data in which accurate responses are less visible as tangent lines, thereby emphasising the larger polar angle errors. A Circular Correlation Coefficient (CCC), proposed by Fisher and Lee (1983), was used to quantify the errors made by the subject and was displayed on each plot. The CCC is calculated as follows:

$$\mathrm{CCC} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin\left(t_i - t_j\right)\sin\left(r_i - r_j\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin^2\left(t_i - t_j\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin^2\left(r_i - r_j\right)}} \tag{2}$$

where n is the number of data points, t is a target polar angle and r is the response polar angle.

Polar angle responses showed substantial variation between the different headphone conditions and significantly larger errors than lateral angle responses. The differences between the conditions can be accounted for by the degree to which the headphones were able to present a signal to the ears that effectively communicated the spectral cues used for determining sound source elevation. If these cues are deteriorated then the auditory system is unable to resolve the cone of confusion and the subject will be unable to determine the virtual sound source's location. Failure to effectively resolve this type of ambiguity leads to large polar angle errors, as they are a di-

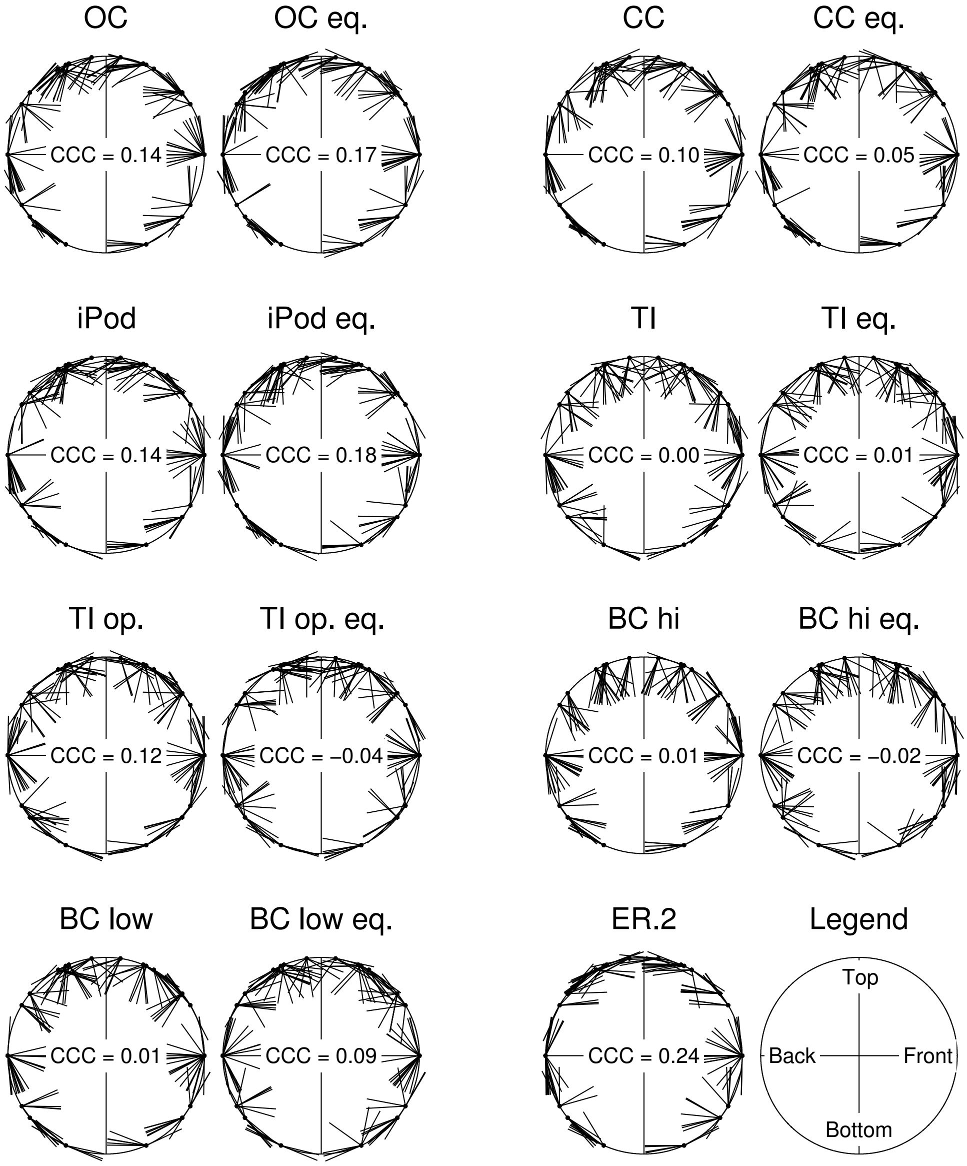


Figure 12: Circular hair plots of polar angle errors for all the headphone conditions tested. The subject responses have been collapsed across all lateral angles. The target positions are represented by the small dots on the circles and the responses by the small segments originating at these points. The direction of a response is determined by the direction that the small segment makes; the location of a response corresponds to the point at which the extended segment intersects the circle. An ideal response is thus a segment that is at a tangent to the small target dot on the circle, and a front-back confusion corresponds to a segment directed towards the opposite side of the circle. The value at the centre of each circular hair plot represents the circular correlation coefficient.

rect measure of the subject's uncertainty in terms of the broadband stimulus' location on a particular cone of confusion.

The headphone conditions *ER·2* (reference), *iPod eq.* (equalised iPod headphones), and *OC eq.* (equalised open circumaural headphones), enabled the most accurate localisation for the subject in terms of measured CCCs with values of 0.24, 0.18 and 0.17 respectively. These values are still relatively low when compared to the correlation coefficients in the lateral angle analysis. This is due to the fact that the calculation of CCCs included many large errors (front-back confusions) that were not present in the correlation calculation for lateral angle errors. The *ER·2* highlights a level of accuracy not seen in any other condition in terms of the localisation of target positions originating from above the listener (polar angle closest to 90°). This is displayed in figure 12 for the *ER·2* condition by the larger number of hairs that are tangential in this region when compared to other conditions, and is also evident from its higher CCC value. The *ER·2* condition did however still produce a number of poorly localised target locations directly in front of the subject (see figure legend), shown by the hairs that are not tangential. The results show that the subject localised the majority of target positions in this region at locations from behind, represented by the hairs pointing to locations on the circle behind the subject. The conditions *iPod eq.* and *OC eq.* demonstrate the large front-back confusion rate even more clearly for target locations with a polar angle of 0° (i.e. directly in front of the subject), with a clear distinction between target positions localised in front (more tangential hairs) and directly behind (hairs tending towards being normal to the circle), particularly for the condition *OC eq.*.

Other headphone conditions demonstrated localisation performance for polar angles that was almost random, such as the *TI*, *TI eq.*, *TI op. eq.*, *BC hi*, *BC low*, *BC hi eq.*, corresponding to the tube intra-aural (with and without equalisation), free-to-ear tube intra-aural without equalisation, both bone conduction without equalisation, and the more expensive bone conduction with equalisation, respectively. The results for these conditions are characterised by a large proportion of localisation responses in the opposite hemisphere (hairs directed normal to the circle) for target locations from above the subject (polar angles larger than 0° and smaller than 180°). These localisation errors lead to CCCs for these conditions that were near-zero or negative. A large proportion of errors for target locations above and behind the subject, is a finding also demonstrated by Makous and Middlebrooks (1990) for free-field localisation using non-individualised HRTFs.

6.3.4 *Global measures of localisation accuracy*

For the global measures of localisation accuracy it was important to differentiate between errors that were front-back confusions and errors that were due to more localised deviations from the target position. This is important due to the described differences in the nature of these two errors both in terms of their distribution and the auditory processing involved. Front-back confusions were thus removed from the calculations for global measures of accuracy and noted as a separate metric in the analysis.

A summary of the spherical correlation coefficients, mean spherical angle errors, and percentage of front-back confusions is presented in table 3, along with the correlation coefficients from the lateral angle analysis and CCCs from the polar angle analysis. The results show that across all headphone conditions the percentage of front-back confusions was high, with a mean of 43%. Some headphone conditions such as *BC hi*, the more expensive of the bone conduction headphones, had over half of the responses removed as front-back confusions. The lowest rate was observed for the reference headphones at 38%. This rate of front-back confusions for the *ER·2* was consistent with previous studies using non-individualised HRTFs for localisation tasks in VAS, such as that by Wenzel et al. (1993) in which a mean front-back confusion rate of 31% was reported across 16 subjects. Wightman and Kistler (1993) also reported high rates of front-back confusions for HRTFs chosen to be different from subjects' own measured HRTFs. The rates in this study must be taken into account when drawing on the remaining two global measures of localisation accuracy for comparisons between the headphone conditions, due to the fact that there was a varying proportion of responses removed for each condition that has a direct effect on any further analysis of the remaining data. It is also important to note that the lateral angle correlation coefficients and polar angle CCCs are measures that included the front-back confusions in their calculation. Given the high degree of front-back confusions observed for the CCCs, if they were removed from the calculations the correlation would increase significantly; the lateral angle correlation coefficients would be relatively unaffected due to the fact that a front-back confusion generally produces an accurate lateral angle response shown in figure 11 by the number of responses close to the diagonal.

With the front-back confusions removed, the mean spherical angle errors for the remaining data showed some variation across the different headphone conditions, ranging from approximately 23° for the reference headphone to 48° for the bone conduction headphone *BC hi*. These results indicate that even once front-back confusions were removed from the analysis the localisation errors were still relatively large for the poorly performing headphone conditions. The SCCs, for the same data (with front-back confusions removed), confirm this result with only five of the 15 headphone conditions demonstrating positive values above 0.67. These headphone conditions, ordered in decreasing SCC, were the *iPod*, *ER·2*, *OC eq.*, *iPod eq.*, and *OC*, corresponding to the iPod, reference, open circumaural equalised, iPod equalised, and open circumaural without equalisation. Globally these headphones significantly outperformed the others and represent a subset that most effectively rendered the broadband stimulus in VAS. The other headphone conditions lead to localisation accuracy that was only barely better than chance performance (i.e. a SCC of 0.00).

CONDITION	SCC	SPH.	% FRONT-BACK	LAT. CC	POL. CCC
OC	0.67	30	44	0.95	0.14
OC eq.	0.75	28	38	0.96	0.17
CC	0.30	33	44	0.97	0.10
CC eq.	0.20	36	42	0.95	0.05
iPod	0.79	30	44	0.95	0.14
iPod eq.	0.69	31	40	0.96	0.18
TI	-0.07	43	40	0.90	0.00
TI eq.	0.13	40	41	0.92	0.01
TI op.	0.55	37	38	0.86	0.12
TI op. eq.	0.52	37	50	0.86	-0.04
BC hi	-0.45	48	52	0.95	0.01
BC hi eq.	-0.26	45	42	0.94	-0.02
BC low	0.38	33	48	0.95	0.01
BC low eq.	0.26	34	41	0.95	0.09
ER·2	0.77	23	38	0.96	0.24

Table 3: Global measures of localisation accuracy for the different headphone conditions tested. For each headphone the spherical correlation coefficient (SCC), mean spherical angle error (SPH.), rate of front-back confusions (% FRONT-BACK), lateral angle correlation coefficient (LAT. CC), and polar circular correlation coefficient (POL. CCC), is given. The colour-coding represents whether the equalised condition had an improved value (green), worse value (red), or equal (black), for a given metric, relative to the condition without equalisation.

COND.	SCC	SPH.	% FRONT-BACK	LAT. CC	POL. CCC
Eq.	0.31 (0.44)	36 (5)	44 (7)	0.93 (0.04)	0.07 (0.07)
No eq.	0.33 (0.35)	36 (4)	42 (6)	0.93 (0.04)	0.06 (0.09)

Table 4: Global measures of localisation accuracy for all headphone conditions grouped by whether or not an equalisation was used. Standard deviation is shown in parentheses. For each headphone the spherical correlation coefficient (SCC), mean spherical angle error (SPH.), rate of front-back confusions (% FRONT-BACK), lateral angle correlation coefficient (LAT. CC), and polar circular correlation coefficient (POL. CCC), is given. The colour coding of the values is the same as for table 3.

6.3.5 *Effectiveness of the headphone equalisation*

Globally, measuring the different metrics across all headphone conditions, grouped by whether or not an equalisation was used, showed that the equalisation resulted in a slight increase in the measured SCC and decrease in front-back errors (see table 4). The other measures had equal or near equal values for with and without headphone equalisation.

For the measures of localisation accuracy used in this study, summarised in table 3, comparisons of headphone conditions with and without the equalisation showed varied results. Improvements in localisation accuracy, in terms of SCC, rate of front-back confusions, mean spherical angle error, lateral correlation coefficient, and polar CCC, are represented in green and degraded localisation accuracy in red, for each headphone condition. The results suggest that there was an improvement for the open circumaural headphone (conditions *OC* and *OC eq.*), but not for the closed circumaural headphone (conditions *CC* and *CC eq.*). The less expensive bone conduction headphone (conditions *BC low* and *BC low eq.*) show some substantial improvements, particularly for polar angle errors, despite a lower SCC, representing a lower correlation between target and response centroids. Similarly, the iPod headphone (conditions *iPod* and *iPod eq.*) showed some slight improvements for the percentage of front-back confusions and metrics used for the lateral and polar angle analysis, yet had a degradation for the SCC and mean polar angle error. This makes it difficult to make an overall judgement on the effectiveness of the equalisation for this headphone. Given the conflicting results from the different measures and that the magnitude of the improvement or degradation was somewhat low (with maybe the exception of the SCC), it would be fair to say that the equalisation had little effect on localisation accuracy for this headphone.

The reference headphone (condition *ER·2*) has been equalised by the manufacturer for scientific research purposes using a dummy head that approximates the ear canal resonance of humans. This headphone, not surprisingly, outperformed or was parity to all other headphone conditions with or without equalisation, for almost all the measures used. The reference headphone showed significantly more accurate localisation, particularly in terms of mean spherical angle error and polar CCC. The equalisation for the open tube intra-aural headphone (conditions *TI op.* and *TI op. eq.*) had a negative effect on localisation accuracy. The other headphone conditions, with or without equalisation, demonstrated localisation accuracy that was too poor to confidently assess whether the equalisation had any effect.

The lack of significant improvement in localisation accuracy with headphone equalisation for the tube intra-aural and bone conduction conditions (*TI* and *BC* respectively) can be explained in terms of their measured frequency responses; these headphones had a response in the higher frequency ranges that was almost non-existent (see figure 9 and 10). The tube intra-aural headphone, with and without equalisation, was unable to produce a signal for frequencies above approximately 7 and 4 kHz respectively. The bone conduction headphones displayed similar results as noted by the manufacturer; the *BC hi* and *BC low* models displaying a cutoff for frequencies above approximately 5 and 6 kHz respectively. These frequency responses lead to a lack of spectral cues in the high frequency ranges that are essential for determining sound source elevation.

6.4 DISCUSSION

The magnitude of the localisation errors reported in this study were comparable to those reported in similar studies in which non-individualised HRTFs were used (Wenzel et al., 1993). The high rates of front-back confusions and elevated polar angle errors in the current study suggest the spectral cues needed to resolve the cone of confusion were not correctly decoded by the subject. In contrast, the analysis of lateral angle errors showed that interaural difference cues were effectively synthesized. This was an expected result, given numerous studies showing the large variation in HRTFs between individuals (Møller et al., 1995b; Mehrgardt and Mellert, 1977; Middlebrooks et al., 1989; Shaw, 1966; Wightman and Kistler, 1989a) and the reliance of the auditory system on specific spectral cues, as discussed in chapter 4. Despite these performance errors, localisation accuracy was still a good indicator of the effectiveness of the different headphone conditions to render sound sources in VAS. Furthermore, the purpose of this study was to test different headphone types, with or without equalisation, and thus key to this analysis was how the headphone types compared to each other rather than their overall performance. Finally, it is important to note that since this study was tailored to applications of binaural synthesis in consumer markets, the use of a non-individualised HRTF was an important and essential characteristic of the results.

Of the different headphone types tested, the open circumaural, iPod, and reference headphones showed the most accurate localisation responses. The open circumaural headphone (without equalisation) outperformed the closed circumaural headphone (without equalisation) for most of the metrics assessed. Open headphones are often considered better suited to binaural synthesis than closed headphones because of open headphones' lower acoustic impedance at the ear (Vorlander, 2000). This tends to provide a more realistic rendering of virtual sound sources via better externalisation. However, some studies have found closed circumaural headphones to perform better than open headphones (Boren and Roginska, 2011). It was suggested however that this was due to the fact that the closed circumaural headphones had the flattest frequency response, as measured on dummy heads. Despite the fact that these frequency responses included pinna effects and the current study did not, the findings would support the role of a flat frequency response and thus an effective equalisation. In the current study, a flat frequency response might have only been achieved for the open circumaural and not for the closed circumaural headphone due to the deep spectral notches as previously described.

For the tube intra-aural and bone conduction headphones the equalisation had little to no effect. As previously mentioned these headphones were unable to produce a signal for frequencies above approximately 6 kHz. It is known that spectral cues, essential for judgements of sound source elevation and front-back location, exist at frequencies above 6 kHz and that localisation accuracy can be significantly affected if they are compromised (Langendijk and Bronkhorst, 2002). In addition, studies have shown that ITDs, used by the auditory system as a dominant cue to sound source lateral location, are encoded by mostly low-frequency auditory neurons (Middlebrooks and Green, 1990) for frequencies below about 2 kHz (Blauert, 1997). This would explain the preserved lateral angle responses for the headphones that otherwise showed large polar angle errors.

With respect to the equalisation performed in this study, it was shown to only have a positive effect for the open circumaural headphones, and possibly for the less expensive bone conduction headphone, yet the improvement was less clear. A somewhat negative effect on localisation accuracy was seen for the closed circumaural and open tube intra-aural headphones, with no significant effect seen for the iPod headphone. The result for the closed circumaural headphone might be explained by the fact that its frequency response was characterised by relatively deep and narrow spectral notches for frequencies above 10 kHz. The notches for this particular headphone were shown to vary more than the other headphones for the measurement repositionings, and this variance might have been be amplified when the headphones were placed on the listener when introducing the filtering effects of the pinna. These deep notches and variations would have been difficult to compensate for, and the equalisation could have easily increased the degree of the deviation from the desired flat frequency response. The open circumaural headphone by contrast did not have such prominent spectral features, and showed less variance for the repositionings, which might explain the effectiveness of the equalisation. For the iPod headphone, the lack of effect of the equalisation might be a consequence of the limitations of the measurement apparatus, given that the frequency response of intra-aural headphones is very much dependent on ear canal resonances. Specific couplers are needed to properly characterise this type of headphone, whereas in this study the same process involving the use of a metal plate was used for all headphones.

In addition, as previously described, the equalisation used in the current study did not take into account the effect that the outer ear might have on the signal produced from the headphone transducer. This is particularly relevant to the circumaural headphones that produced a signal that was filtered by the pinna in much the same way a HRTF filters sound sources in space. Many studies have suggested that an individualised headphone equalisation is mandatory for an effective VAS synthesis (Pralong and Carlile, 1996; Wightman and Kistler, 2005; Møller et al., 1995a). The lack of an individualised headphone equalisation may have reduced the benefits of the equalisation in this study to the point where the differences were difficult to measure.

There are obvious limitations with the current study due to the fact there was only one subject performing the localisation task. Despite the fact that only the author participated in the study, the results can be assumed to be robust given the experience and large amount of training the subject received during the experimental design. In addition, the author has been shown to have the highest degree of repeatability among six subjects tested, for a listening test in the next chapter. If repeatability in a listening test is comparable in some manner to reliable results in a localisation task, which is likely given that they both involve the evaluation of a binaural synthesis, then the author can be thought of as a reliable subject. The author also has a lot of experience with localisation tasks from previous studies prior to this body of work (see for example Carlile and Schönstein, 2006a,b).

Future experiments should however be performed using a number of subjects in order to make the observations about the effect of headphone equalisation more statistically robust. Learning effects would have also played a role in the current study, however, as previously mentioned, it would not be expected to be significant given the numerous repeats of the localisation tasks performed by the author. If there were learning effects, they would be equal across all headphones given that the headphone conditions were chosen at random.

6.5 CONCLUSION

The purpose of this chapter was to introduce the crucial role that headphones play in the presentation of virtual sound sources via a binaural synthesis. A study of the different types of headphones used in VAS, along with an assessment of the usefulness of a headphone equalisation, was presented. The headphones were assessed in terms of localisation accuracy for one subject. It was found that despite the fact that localisation tasks were characterised by relatively large response errors, there were significant differences between the headphones. Three of the eight headphones tested showed reasonably good localisation performance, comparable with previous studies using similar methodologies.

In terms of the headphone equalisation, the results showed that the equalisation had varying effects on the different headphones, with a significant improvement reported for only one of the headphones. Despite only improving results for one headphone, it must be stressed that many of the other headphones tested were of low quality. Some of the headphones were unable to produce a signal at high frequencies. The only other high fidelity headphone tested with the equalisation, other than the open circumaural that was shown to benefit from the equalisation, was the closed circumaural, which as the results suggest is a difficult headphone to equalise due to resonances in the closed cavity. Thus, the type of headphone equalisation used in the current study has merit based on the open circumaural headphone results, and due to the fact that the procedure does not involve customising the headphone equalisation for each listener, i.e. the equalisation was performed on a metal plate rather than the listener's head.

7 PERCEPTUAL JUDGEMENTS OF HRTFS USING LISTENING TESTS

This chapter aims to discuss and study how listeners perceive the quality of a binaural synthesis. In the previous chapter, certain aspects of how a binaural synthesis is delivered to the listener, namely the type of headphone and headphone equalisation used, were assessed using localisation accuracy for sound sources in Virtual Auditory Space (VAS). In this chapter, a study is presented that uses listening tests (see section 3.5.2) as an alternative to evaluating the quality of a rendering in VAS using a particular Head-Related Transfer Function (HRTF).

7.1 BACKGROUND

As noted by Martens (2003):

> While a primary goal for binaural synthesis is to spatially position an auditory image, methods typically employed to study the ability of human listeners to spatially localize an actual sound source address only a narrow subset of the issues that are important to the development of adequate spatial auditory display technology.

In a similar vein, the current study was performed in order to measure aspects of a binaural synthesis that were particularly relevant to their application in consumer markets and did not focus on localisation accuracy as a means of evaluation. The experimental design of the listening tests used in this study were orientated towards questions relating to whether virtual auditory images were coherent or whether the rendering in general was realistic and well externalised. These questions were framed in the context of HRTF customisation techniques for listeners; perceptual judgements of different HRTFs are key to validating any procedure that aims to select an optimal non-individualised HRTF for a particular listener, as will be seen in the next chapter.

Studies that have used listening tests to evaluate HRTFs have had subjects judge varying attributes, such as perceived naturalness (Usher and Martens, 2007), tone timbre (Matsui and Ando, 2009), externalisation (Seeber and Fastl, 2003), or localisation (Huopaniemi et al., 1999; Iwaya, 2006). The different studies used a variety of listening test types, yet did not address a fundamental question relating to the validity of the obtained results; the question of whether the subject responses were reliable or reproducible. The only known study that has extensively looked at the reproducibility of perceptual evaluations of HRTFs by a number of subjects has shown that judgements of an optimal HRTF from a selection of 32 different sets of HRTFs were essentially random when repeated ten times (Yairi et al., 2008). It is unclear whether these findings were a consequence of known biases in the listening test design or due to the inability of subjects to make definitive judgements on the quality of renderings in VAS using different HRTFs.

One of the major impetuses of the current study was thus to establish whether reliable subject responses were possible when evaluating different HRTFs. The aim was to develop, through an iterative process, a listening test that was free from known biases and that could be used to evaluate the reproducibility of subject responses. Ideally, it is only after an acceptable degree of repeatability has been observed among the subjects being tested that the HRTF judgements can be used for further analysis.

7.2 OUTLINE

This study develops three broad designs of a listening test: Listening Test 1, Listening Test 2, and Listening Test 3. For each listening test design, different iterations were performed that led to improvements in the results. A summary of these iterations along with their purpose is summarised as follows:

- Listening Test 1 - 46 HRTFs were judged as either *excellent*, *fair*, or *bad* for virtual sound sources describing two different trajectories.
- Listening Test 2 - the same number of HRTFs were judged using a completely redesigned interface with a visual cue to sound source location. The most significant change was the use of a continuous scale for judgements instead of a discrete scale (i.e. the categories *excellent*, *fair*, and *bad*). A continuous scale allowed for a scaling of responses, enabling a more detailed comparison across subjects.
  - Listening Test 2.1 - the exact same design as used in Listening Test 2 was used, with the author of this body of work performing three replicates, in order to asses the reproducibility of judgements made for each HRTF. The reproducibility of responses was shown to be poor.
  - Listening Test 2.2 and 2.3 - the number of HRTFs was reduced to 26 and then again to five in an effort to improve reproducibility. The results showed that when using only five HRTFs, judgements began to become consistent across replicates.
- Listening Test 3 - the same design was used as in listening test 2.3 (using five different HRTFs), however judgements were made for three attributes separately as opposed to one global measure as in previous versions of the listening test. It was hoped that the judgement of specific attributes would make responses more reliable. Reproducibility was shown to improve for some of the attributes tested.
  - Listening Test 3.1 - the number of replicates performed was increased from three to four. Although different subjects and a varied set of HRTFs were used, the results showed the best reproducibility observed for any version of the listening test.

- Listening Test 3.2 - this was the final design of the listening test, which aimed to address some biases observed in Listening Test 3.1; changes were made to the headphones, stimulus, and design of the listening test. The effect of subject expertise on the reproducibility of the perceptual judgements was also assessed.

7.3 LISTENING TEST 1

The first listening test used in this study was not performed by the author; it was a component of the LISTEN project that also produced the public online database of HRTFs (see section 6.2.3). Subjects judged 46 different HRTFs from the database with respect to how well a virtual sound source described defined trajectories in space. The large number of subjects taking part in the project allowed for the use of statistical techniques that were robust to noise (unreliable responses) within the data.

7.3.1 *Listening Test 1 procedure*

The HRTF recordings, methods used in the binaural synthesis of the stimuli, and procedure of the listening test are all described in detail in section 6.2.3 and 6.2.4. In this listening test, subsequently known as Listening Test 1, subjects were asked to respond, on a category scale, to the quality of VAS renderings of pre-described sound trajectories (in the vertical and horizontal planes) with a judgement of either *excellent*, *fair*, or *bad* (labelled as *excellent*, *moyen*, or *mauvais* respectively in the test as it was performed on French listeners). Both trajectories were had the virtual sound source presented at the same distance. Subjects were asked to judge approximately 10 HRTFs as *excellent* in an effort to maintain some degree of similarity in the distribution of responses across the different subjects for subsequent analysis of the results.

A total of 45 subjects participated in the listening test, corresponding to all but one of the subjects for which there were HRTF recordings in the database at the time. As mentioned, all subjects were French speaking adults; a mixture of male and female. Each subject judged the 46 HRTFs from the database including their own. Subjects were not told which of the HRTFs judged was their own.

7.3.2 *Listening Test 1 results*

The results from Listening Test 1 have been presented in a paper by the author (Schönstein and Katz, 2010b) as part of a separate study that will be detailed in chapter 8. A summary of these results, showing each subject's judgements of all HRTFs in the database, is displayed in figure 13. The vertical axis represents the subject that made the response and the horizontal axis represents the subject's response. The colour of each of the circles relates to whether the subject judged a rendering, using a particular subject's HRTFs from the database, as either *excellent*, *fair*, or *bad*, corresponding to green, yellow, or red, respectively. The circles on the diagonal, corresponding to when a subject judged his or her own HRTFs, are emphasised by a slightly enlarged circle size. Subject 10 did not take part in the listening test.

It is apparent from the results that the majority of the subjects (all subjects except three), judged their own HRTFs as *excellent*. This result is consistent with findings that have shown listeners more accurately determine the location of virtual sound sources using individualised compared to non-individualised HRTFs (Møller et al., 1996; Wenzel et al., 1993, see also chapter 5). Studies using listening tests to evaluate other subjective or descriptive attributes of binaural syntheses have shown that in general subjects judge their own HRTFs as most effective (Schönstein and Katz, 2010a), yet in some cases subjects have been shown to not judge their own HRTFs with the highest preference value (Usher and Martens, 2007). Judgements made by subjects in Listening Test 1 were mostly based on externalisation and source position stability; the main task was to judge whether the sound source followed a circular path in the horizontal plane and was thus not strictly a localisation task. The judgements most probably also incorporated some degree of affective judgements as will be detailed in the following sections.

The mean number of HRTFs judged as *excellent*, *fair*, or *bad*, was approximately 9, 19, and 18 respectively, with standard deviations of approximately 4, 8, and 9 respectively, suggesting that there was a significant degree of variation across subjects in the number of HRTFs assigned to each of the three possible categories. These types of variations among subjects are not uncommon in the field of descriptive sensory analysis, due in part to the fact that subjects will use the judgement scale differently and have varying sensitivities to the stimulus. In addition, the subjects' expectations of the realism of renderings in VAS will affect their judgements and produce differences across subjects (Zielinski et al., 2008). These types of variations can usually be accounted for using various statistical techniques (see Naes, 1990) as will be seen in the analysis of results for future listening test versions. However, due to the fact that Listening Test 1 used only three categories for subject responses it was difficult to compensate for the variation and normalise the results across all subjects in some way for the benefit of further analysis.

7.4 LISTENING TEST 2

The second listening test, Listening Test 2, was performed as part of this study by the author. This listening test was similar to Listening Test 1, with the most significant difference relating to the change from a category scale for responses (i.e. *excellent*, *fair*, or *bad*) to a continuous scale, in which judgements anywhere between two end-point descriptors could be made. Listening Test 2 was conducted in an effort to improve on the results from Listening Test 1 in terms of the ability to draw conclusions from the judgements of the different HRTFs. In particular, as previously described, it was of interest to be able to scale subject judgements for a statistical analysis of the results, so that, for example, for each subject it was possible to ascertain the 10 most effective HRTFs from the database.

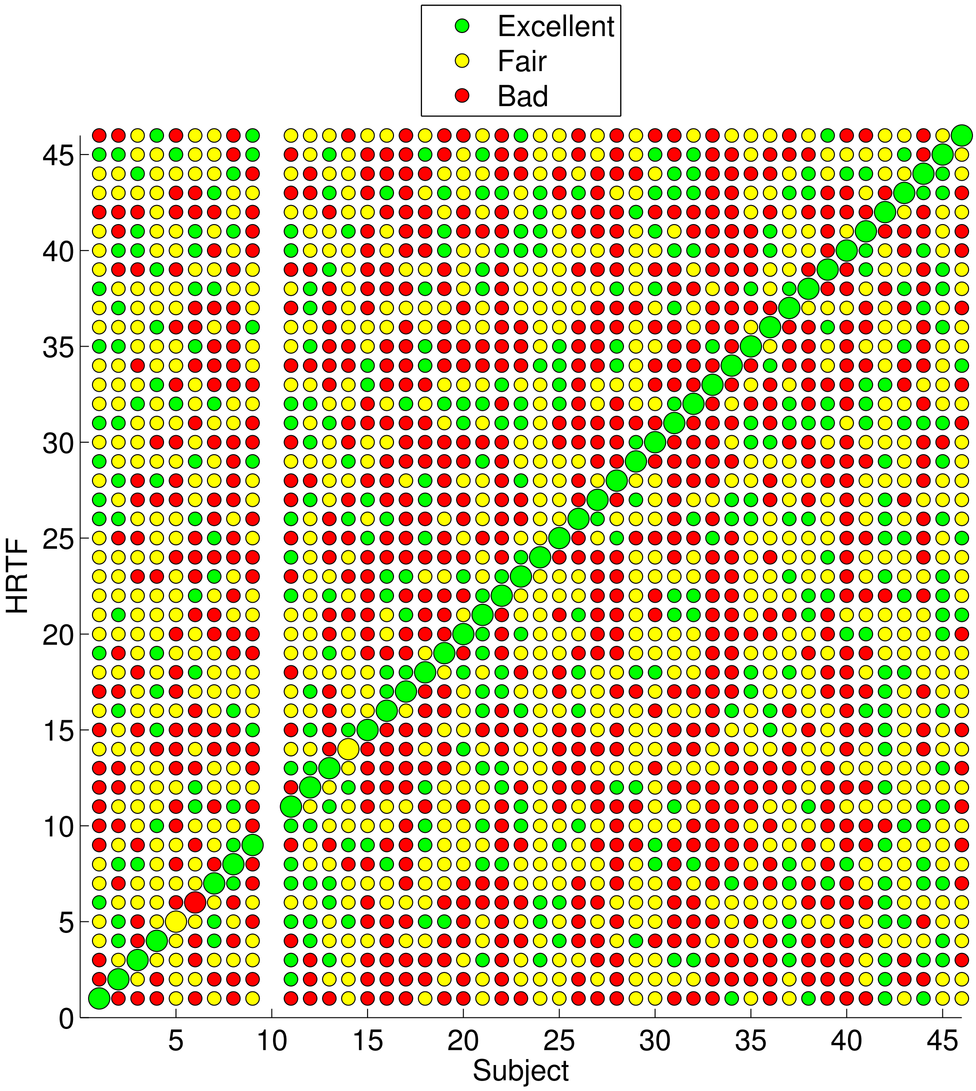


Figure 13: Results from Listening Test 1 in which judgements of either *excellent*, *fair*, or *bad*, corresponding to colours green, yellow, and red respectively, were made by each subject for the 46 HRTFs in the LISTEN database. The subject that made the judgement (subject number) is represented on the horizontal axis and the HRTF judged (also indicated by subject number) is represented on the vertical axis. Subject number 10 did not make any judgements.

7.4.1 *Listening Test 2 procedure*

A total of seven subjects, including the author of this body of work, performed Listening Test 2. The subjects were comprised of six male and one female adult, and all presented positive results for a pure tone audiometry test (standard ISO/IEC IS 8253-1:1989).

The listening test took place in a sound dampened room using a pair of open circumaural headphones (model HD570 by manufacturer Sennheiser) shown to be effective by the author of this body of work for the presentation of binaural synthesis in the previous chapter. No form of headphone equalisation was used, as it was shown in the previous chapter, that for non-individualised HRTFs the effect of a non-individualised equalisation lead to only minor improvements in terms of localisation accuracy. All stimuli were presented from an external soundcard (model Fireface 400 by manufacturer RME) running off a portable computer. Using the same method for setting stimulus level as described in the previous chapter (see section 6.2.2), stimulus level was set to 50 dB sensation level. This represents the level in decibels above a subject's detection threshold for the virtual sound source presented directly in front of the listener.

The binaural synthesis used in this listening test made use of the same random Gaussian white noise (normally distributed broadband flat power spectral density), of length 0.23 seconds, modelled using a Hanning window (with length of the stimulus), as was used for Listening Test 1 (see section 6.2.4). The modelled broadband noise was repeated to form a monophonic signal of duration three minutes in order to give the listener ample time to judge the rendering; subjects almost never listened to the full three minutes. The broadband noise was not freshly generated for each burst, i.e. the same noise burst was continuously repeated. This was not the ideal stimulus to present subjects as it may lead to learning effects. For later versions of the test the noise bursts were freshly generated in order to avoid this issue.

This signal was rendering in VAS using a real-time binaural synthesis engine called the LIMSI Spatialization Engine (LSE) (Katz et al., 2010), constructed and run in the audio programming environment Max (version 4.6.3). Virtual sound source positions were fed into the engine and in real-time the signal was convolved with corresponding Head-Related Impulse Responses (HRIRs) to create a dynamic binaural synthesis. A dynamic synthesis, as opposed to presenting virtual sound sources in fixed locations, provides the listener with additional cues to sound source location (see section 2.3). The result is a virtual sound source that is perceived as having a well defined location in space, which improves the overall effectiveness of the binaural synthesis in terms of realism.

In order to make for a seamless percept of the sound source moving in space, HRIRs were interpolated to sample the space around a listener every 5° in azimuth and elevation, which represented an increase in the number of positions in space from 187 to 2,016. Due to the fact that studies have shown the minimum audible angle of a sound source in azimuth and elevation to be approximately 1° and 4° respectively (Perrott and Saberi, 1990), it was assumed that subjects could not perceive the shift from one HRIR to another in space for this higher resolution. The interpolation used was what is termed as the a nearest neighbour technique, in which the nearest four HRIRs to the desired interpolated position are used to make an estimate (see Hartung et al., 1999, for a copmparison of different interpolation techniques, and section 3.4). In the current study, this involved the supposition of pairs of the closest HRIRs after aligning them via the maximum of the cross-correlation, and using gain factors relating to the distance of the measured HRIRs to the desired position in space. Thus for the four closest HRIRs to an interpolated position in space, aligned and weighted HRIRs were interpolated for a pair above and below the desired location, and then once more for these two interpolated HRIRs.

A total of 51 HRTFs were used in the listening test; these HRTFs represented the same 46 HRTFs described as part of the LISTEN database (see section 6.2.3), with the addition of five sets of HRTFs that were recorded in exactly the same manner after the LISTEN project had finished. The HRTFs were neither diffuse nor free-field equalised. As previously explained in section 3.5.3, there are externalisation benefits associated with using raw HRTF recordings for a binaural synthesis.

For each of the auditioned HRTFs, two trajectories (figure 14) were described in sequence using the broadband stimulus as a virtual sound source:

1. An arc over the listener, that crossed the midline, beginning behind and below at approximately $135°$ azimuth and $-45°$ elevation and finishing at approximately $-45°$ azimuth and $-45°$ elevation in front and below, and then returning to its original position.

2. A horizontal circle at the level of the ears of the listener (i.e. at $0°$ elevation) originating at $135°$ azimuth and rotating twice around the listener.

Positions in Trajectory 1 were chosen as they did not vary in azimuth and were therefore ideal for evaluating the perceived elevation of the virtual sound sources. Trajectory 2 was chosen for a similar reason; positions did not vary in elevation and allowed for an analysis of the perceived laterality. Making judgements of locations in trajectories along azimuth and elevation (in which only one of the two vary) is perceptually relevant as different cues are used by the auditory system to determine location along these dimensions, namely interaural and spectral cues respectively. The virtual sound sources were presented at equal distances for the two trajectories, as for Listening Test 1, corresponding to the distance at which the original HRTFs were recorded in the LISTEN database.

The 51 binaural renderings, corresponding to the 51 HRTFs being evaluated, were presented to the listener via a Graphical User Interface (GUI) on a portable computer using a software application called Sonic Mapper. A screenshot of the different listening test components is displayed in figure 15. Each HRTF was numbered from 1 to 51 and subjects were able to play each one in any order and for as long as they wished up to the full three minutes.

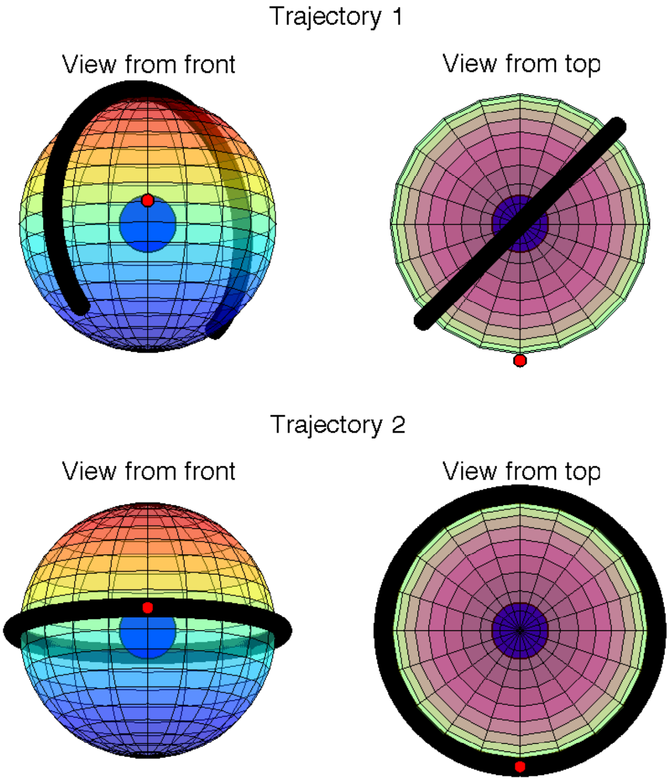
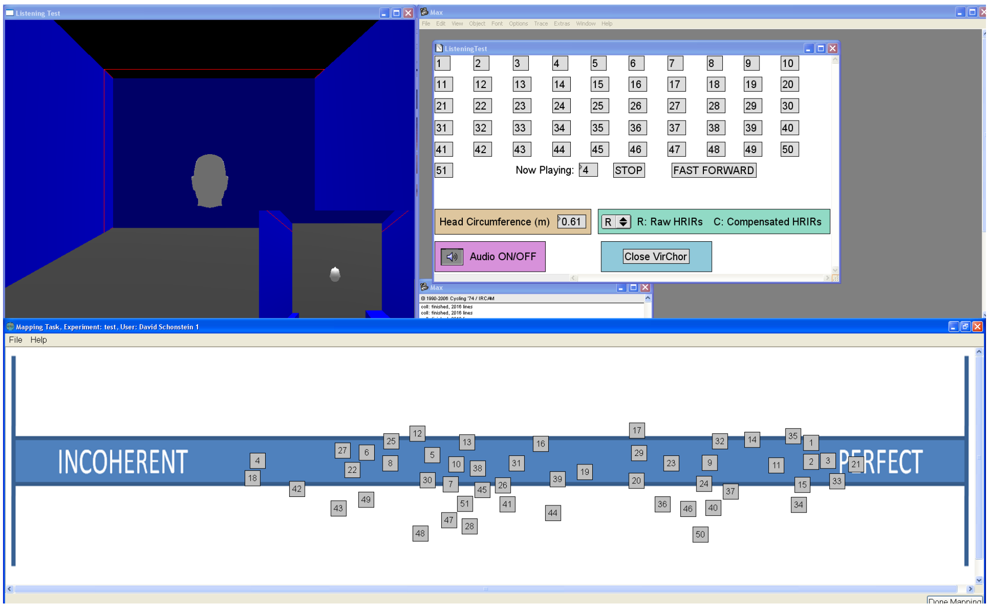


Figure 14: Virtual sound source trajectories in Listening Test 2. The black line represents rendered positions. Red circles represent the position directly in front of the listener, at azimuth 0° and elevation 0°, included in the figure as a reference. The larger blue circle at the centre of the sphere represents the position of the listener's head.



Mapping Task. Experiment: test. User: David Schonstein 1

Now Playing: 4

Head Circumference (m) 0.61

R: Raw HRIRs C: Compensated HRIRs

Figure 15: A screenshot of the graphical user interface used for Listening Test 1.

The HRTF number being played was always displayed to the subject. HRTFs were auditioned by selecting a number between 1 and 51 and clicking on the *PLAY* button. It was possible to skip to the second trajectory in the sequence (the horizontal circle) by clicking the *FAST FORWARD* button.

As can be seen in figure 15, subjects were also provided with a visual aid on the computer screen in the form of a two-dimensional and three-dimensional video showing a sphere describe the given trajectory around a head that was meant to represent the listener. This visual display was created using an open source programming environment for real time graphic and sonic scenes called VirChor[1]. A view from behind the listener and from above was displayed, corresponding to the the ideal location of the virtual sound source being presented over headphones. The position of the sphere in the display being generated by VirChor was fed directly to the binaural synthesis engine in terms of azimuth and elevation angles (distance was held constant). These angles were used to index the correct interpolated HRIR for the location presented.

This real-time visualisation of ideal virtual sound source location was provided so that subjects would be able to more easily evaluate any discrepancies between what they saw and heard in the listening test. It was possible however that in the process there would be a biased audio-visual effect as listeners have a tendency to fuse an audio event with a presented visual origin of the sound source (e.g. the ventriloquist effect; Choe et al., 1975, described in chapter 2). It has also been shown that the audio-visual interaction between two-dimensional videos and spatial audio (in the form of wave field synthesis) provide a mismatch between perceived auditory and visual sound directions when the listener is not in the ideal viewpoint, i.e. from the viewpoint of the head in the video of the current study (de Bruijn

[1] See http://sourceforge.net/apps/mediawiki/virchor for details.

and Boone, 2002). It has been shown that a three-dimensional graphic image, as opposed to a two-dimensional image, can improve localisation (Pernaux et al., 2003), however this is when the visual is used as a feedback to a subject's localisation response and not as an aid in a listening test. Despite the possibility of an audio-visual mismatch, subjects reported after completing the listening test that the visual cues were not used throughout the entire task; rather the visual cues were used as an aid in generating a mental representation of the desired trajectory whilst listening to the first couple of HRTFs and then ignored. In addition, given that the same visual display was provided for all 51 HRTFs, any audio-visual interaction can be assumed to have the same effect across the different renderings and thus have little impact on the relative comparisons made by the subjects.

Subject responses to the stimuli were made via the GUI by dragging each individually numbered HRTF and placing it on a continuous scale between two end-point descriptors: *incoherent* and *perfect*. These two descriptors corresponded to a rendering in which it was impossible to determine the virtual sound source location, and in which the virtual sound source corresponded perfectly to the trajectory displayed in the video, respectively. The software used reported the location of each of the positioned numbers (corresponding to the different HRTFs) on the continuous scale. Subjects were also asked to judge each binaural synthesis in terms of the realism of the auditory illusion, and whether the virtual sound source was effectively externalised (i.e. not perceived inside the head). It was explained that responses should be absolute, and that they did not need to perform pairwise comparisons. They were also told to imagine that the virtual sound source was being played in an anechoic environment, much like the sound dampened room in which they were performing the task. A preferred strategy was also described to the subjects, in which the audition of three or four HRTFs provided an idea of the range of differences that could be expected between the different syntheses, leading to a somewhat rapid placement of the HRTFs on the continuous scale.

It was asked that subjects spend approximately 30 minutes, briefly listening to the renderings and arranging the 51 HRTFs in this first phase of the test, and a further 30 minutes to perform a more comprehensive audition and rearrangement. Subjects were able to place a number at the same position on the scale as another number by simply dragging the one above the other; only the horizontal location of the numbers was registered in the test. By doing this, subject judgements avoided what is known as the stimulus frequency bias (Zielinski et al., 2008), in which responses tend to occupy a larger range, if listeners respond to a stimuli being almost but not exactly the same, on a scale when in fact they are perceived as equal (Poulton, 1989).

More generally, the continuous scale used for judgements in Listening Test 2, in accordance with Note 1 in recommendation ITU-R BS.1116-1:1997, afforded it some advantages in terms of known biases for the evaluation of audio quality with respect to Listening Test 1. In the latter, responses were fixed on a labelled and categorised scale, which in reality should be perceptually linear. Different subjects would consider the three categories to lie along the perceptually linear scale at different locations. Thus judgements in Listening Test 1 should be treated as ordinal in which only rank is preserved; responses were quantised due to the categories *excellent*, *fair*, and *bad*. The use of only two labels at the ends of the scale in Listening Test 2, as displayed in figure 15, allowed for a continuous measure of the effectiveness of the HRTF renderings without a quantisation effect due to the presence of ticks, labels, or the like, between the two end-points (see Conetta, 2007).

Finally, Listening Test 2 used differences in the language of its labels to avoid any bias due to affective judgements of the stimuli, defined as responses relating to the listener's opinion of a sound's character. The labels in Listening Test 1 had an inherent hedonic component to them in the sense that they were not grounded in pure descriptive terms of the stimulus; rather they were examples of language that would be used to express whether the sound was liked or disliked. It is known that affective judgements are vulnerable to non-acoustic factors such as poor long-term stability (Kirk, 1957), bias due to emotions and mood (Vastfjall, 2004), and ambiguities in the meaning of the sound (Zimmer et al., 2004). In addition to these factors and most relevant to this study, affective judgements are known to depend largely on subject expectations of a stimulus, which can vary between subjects. For example, Rumsey (1999) found a bimodal distribution of preferences for up-mixing algorithms (original two-channel recordings compared to five-channel up-mix) using a listening test, which was linked to differences in the subjects' expectations. The degree of variation in subject responses demonstrated in Listening Test 1, in terms of different numbers of HRTFs judged as either *excellent*, *fair*, or *bad*, highlight the different internal standards of the subjects and that they might have had different expectations about how realistic the illusion of a binaural synthesis should be. These internal standards can vary between subjects and can even vary for the same subject over time. Listening Test 2 reduced the affective component in the judgements made by using language such as *incoherent* and *perfect*, which are more descriptive in their nature.

7.4.2 *Listening Test 2 results*

The judgements made by the seven subjects are presented in figure 16. Each plot shows the position of each of the 51 HRTFs placed on the scale by the listener. The results show that for four of the subjects a reduced range of the scale was used, most probably due to differences between the subjects in the expectations of the VAS renderings and sensitivity to them as previously discussed. One subject appears to have created graduations on the scale, presumably to make the task easier by reducing the number of comparisons needed to classify the different HRTFs. The responses for this subject obviously incorporate a significant mapping bias in which the judgements are quantised; a bias that leads subjects to respond at a higher frequency near gradings on a scale (Conetta, 2007). The variation between subjects was an expected result, given the reasons outlined previously, and that different subjects had different levels of expertise with respect to perceptual judge-

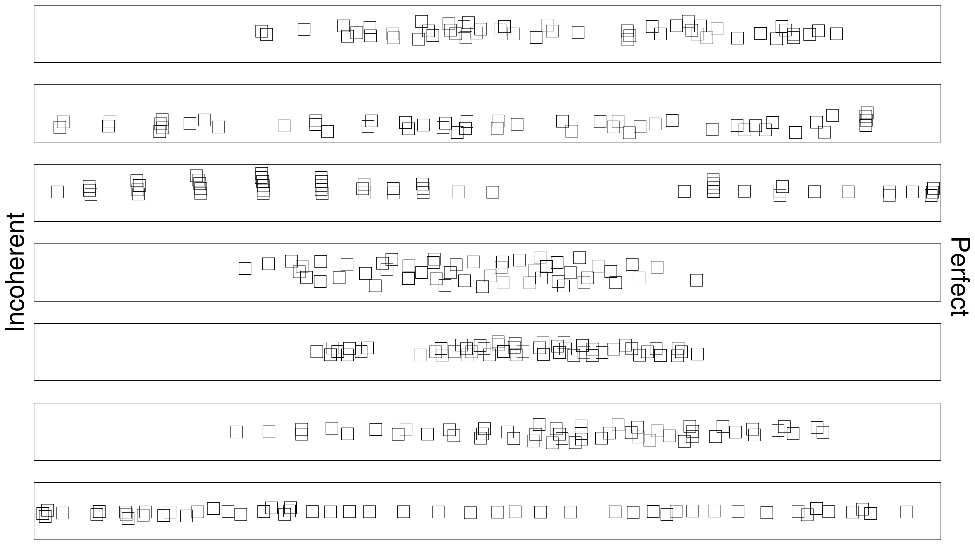


Figure 16: Subject responses for 51 different HRTFs using Listening Test 2. Each plot displays judgements from one subject for all HRTFs between end-point descriptors *incoherent* (far left) and *perfect* (far right).

ments of renderings in VAS (a topic that will be explored in section 7.10.6). As described in the previous section, whilst the design of a listening test can help reduce between subject variations, the continuous assessment scale must be viewed as relative rather than absolute. The fact that there was variation between the subjects was not seen as a major issue, given that these differences can be accounted for, as long as responses were reliable.

Many techniques are available to handle individual differences in descriptive sensory analysis, which is afforded to the results in Listening Test 2 due to the continuous scale used (see for example Romano et al., 2008). As originally described by Naes (1990), there are different ways in which assessors can differ in their use of a continuous scale, of which only two are relevant to this listening test: differences in *where* subjects put their judged HRTFs (level effect), and differences in the span of the scale used for judgements (range effect). Any other individual differences, such as subjects having a varied sense of dissimilarity between pairs of HRTFs, must not be compensated for due to the fact that the judgements of HRTFs are by definition specific to the listener and it is in fact the very purpose of this study to draw out these results. Other techniques used to handle individual differences such as a procustes rotation are not relevant for this listening test as they are used to account for confusions in the attributes of the stimulus being judged, which is very unlikely given the simplicity of the task and the end-point descriptors used.

Given these observations, the simplest form of adjustment of the listening test results was used, known as *standardisation*, which involved a stretching or shrinking and translation of responses. In this way, judgements on the continuous scale are adjusted to have a mean of zero and a standard deviation of one. Other methods of adjustment such as that proposed by Berge (1977), in which scaling is used to make individual scores between assessors

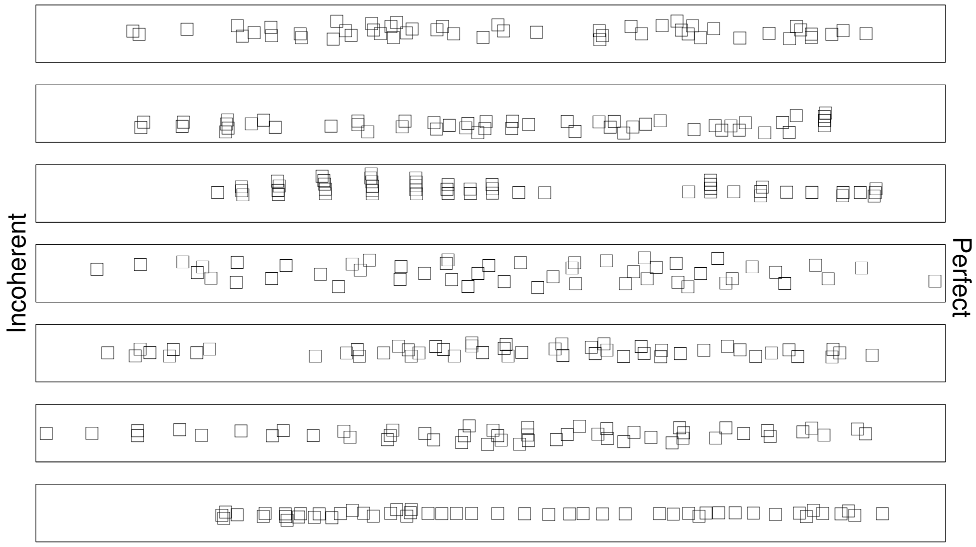


Figure 17: Standardised subject responses for 51 different HRTFs using Listening Test 2. Each plot displays judgements from one subject for all HRTFs between end-point descriptors *incoherent* (far left) and *perfect* (far right). Responses for each subject have been scaled and translated to have a mean of zero and a standard deviation of one.

as similar as possible, were not appropriate for reasons described above. Figure 17 shows the standardised judgements for each subject, with a spread of responses that was more amenable to the desired statistical analysis. In particular, it was possible to divide the judgements into any number of categories or simply use the rank of each HRTF across all subjects, which is a significant improvement over Listening Test 1.

7.5 LISTENING TEST 2.1

Before the results from Listening Test 2 could be used in an analysis, there was one final aspect of the judgements that needed to be assessed: subject reproducibility or repeatability, gauged by evaluating the within assessor variability. Results provided by a subject that demonstrate a large degree of variation for judgements of the same stimulus across multiple tests are not reliable and should be separated from results that show good reproducibility. Testing subject reproducibility, or detecting within subject variation, can be assessed in typical descriptive sensory analysis by comparing an individual subject's average response, the average response across all subjects, and replicates by the subject for each attribute tested (Naes and Solheim, 1991). The differences between these measures allows for a view of a subject's performance relative to the other subjects taking the test. For the type of listening test used in this study however, comparisons of the judgements of specific HRTFs across subjects are somewhat meaningless as an HRTF will be perceived differently for each listener. Thus, an analysis of response replicates was only possible for individual subjects.

7.5.1 *Listening Test 2.1 procedure*

Due to the fact that Listening Test 2 sometimes took in excess of one hour to complete, it was not possible to ask each of the seven subjects to do replicates of the listening test. Therefore only the author of this body of work was assessed across three replicates of the listening test. The author would be likely to show the highest degree of reproducibility given increased familiarity with the listening test stimuli and procedure. The author can be assumed to be a reliable subject given the level of reproducibility shown for later versions of the listening test (see section 7.10). There remains however a limitation for this version of the listening test due to the fact that there is only one subject and the results obtained cannot be compared for a number of subjects.

7.5.2 *Listening Test 2.1 results*

The variability in standardised subject responses for each HRTF judged is presented in figure 18 for the author. The different coloured circles correspond to the three replicates of the listening test. The horizontal axis corresponds to the HRTF that was judged, ordered in increasing positive judgement from left to right according to the third replicate. The vertical axis corresponds to the standardised judgement. The results suggest that there were very few HRTFs showing reproducible responses across the three replicates.

A measure of the reliability of a subject's responses, originally proposed by Gabrielsson (1979), is the error variance, which represents the within-group variability, or unexplained variance, for an ANalysis Of VAriance (ANOVA). In the context of this study, the error variance represents the sum of the squared differences from the mean response for each HRTF, across the replicates, divided by the degrees of freedom. The degrees of freedom is simply calculated as the total number of responses made across all HRTFs and replicates minus the total number of HRTFs judged. The error variance, is equivalent to the mean-square-error, which can also be calculated as the mean response variance for each of the judged HRTFs across replicates. Thus the error variance can be expressed as:

$$\sigma = \frac{1}{a}\sum_{i=1}^{a}\left(\frac{1}{1-n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \tag{3}$$

where n is the number of replicates of the listening the subject performed, and a is the number of HRTFs judged. X_i represents the subject response for one HRTF, and $\overline{X}$ represents the mean response for that HRTF across replicates.

For Listening Test 2.1 the error variance was calculated to be 1.09 for the author (the same subject that participated in Listening Test 2). This variance was large given that responses for any given replicate were standardised to have a standard deviation of one, i.e. the error variance, or mean-square-error, across the judged HRTFs was greater than the standard deviation within one replicate. The standard deviation of the variances was calculated to be 0.83, which is almost as large as the error variance; the maximum variance was 3.23, which represents an HRTF that was judged as both near end-point *perfect* and *incoherent* across replicates. This was not an isolated event; the results show responses for a number of HRTFs that were judged closest to the descriptor end-point *incoherent* in one trial followed by responses closest to the end-point descriptor *perfect* in a subsequent trial, and vice versa.

There were some HRTFs that showed small variance such as number 48, 38, and 16. With the exception of number 48, HRTFs that had small variance were generally judged in the middle of the continuous scale, i.e. between *incoherent* and *perfect*. More variance was observed for HRTFs judged at the extremes of the continuous scale, which suggests that it was difficult to judge an HRTF as conclusively good or bad. It is possible that for the HRTFs with small judgement variance, the judgement criteria, relating to whether the sound source was well externalised or accurately followed the desired trajectory, was well defined for the subject.

Given the lack of reproducibility observed for most of the judged HRTFs for this version of the listening test (Listening Test 2.1), results from Listening Test 2 for the seven subjects were deemed unreliable and not fit for further analysis. This conclusion could not be extended to responses made by subjects in Listening Test 1 given that the task of assigning one of the three categories *excellent*, *fair*, or *bad* would have probably been an easier task, ensuring some degree of reproducibility. In addition, the fact that most subjects judged their own HRTFs as *excellent* offers an indication that the judgements were somewhat reliable. The author's own HRTFs were not included in the current listening test given that they were recorded using different procedures and equipment. The subject's own HRTFs were added to future listening tests once an equalisation had been performed so that all HRTFs resembled each other.

The observed variation in responses was not an unexpected result, as subjective feedback from the six other subjects relating to the difficulty of the task in Listening Test 2 implied that it was a very demanding task and that responses were almost random at times. One possibility for the demanding nature of the task was the sheer number of HRTFs to be compared; subjects would need to listen to 1,275 pairs of HRTF renderings if they wanted to make all 51 pairwise comparisons. Even though subjects were specifically asked to try and make absolute judgements and hence were not required to make all pairwise comparisons, it can be assumed that an excessive number of comparisons, albeit not the entire 1,275, needed to be made, particularly since subjects were asked to complete each trial in under an hour. It is very likely that subject fatigue leading to reduced concentration would also have played a role in producing unreliable responses.

7.6 LISTENING TEST 2.2

In an effort to increase the reproducibility of the listening test, a reduction was made in the number of HRTFs to be judged. The number of HRTFs was re-

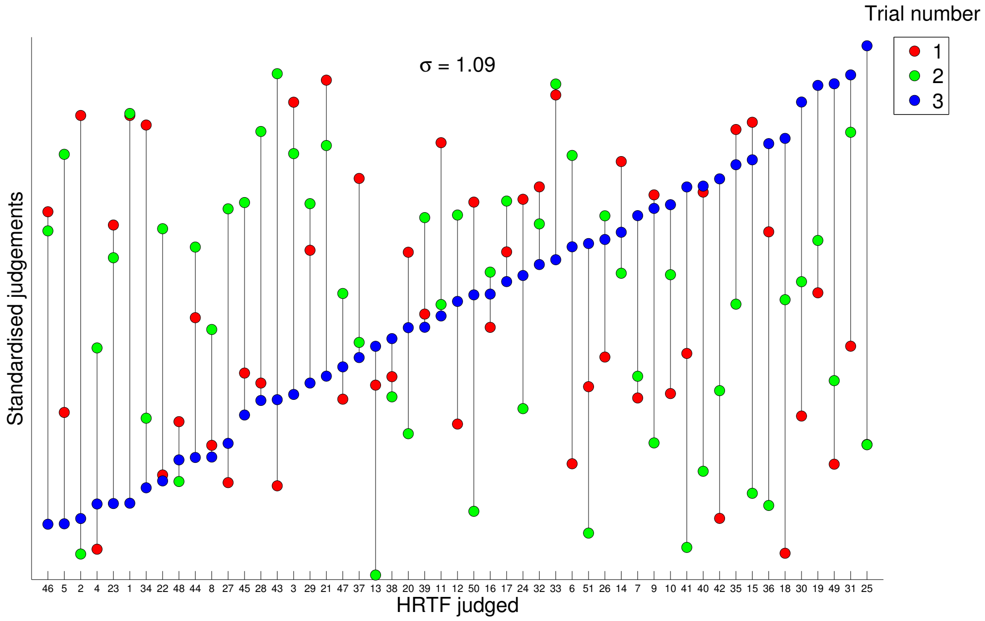


Figure 18: Three replicates of the author's responses for the 51 different HRTFs using Listening Test 2.1. The standardised judgements are represented on the vertical axis increasing from bottom to top, corresponding to the endpoint descriptors *incoherent* and *perfect* respectively. The judged HRTFs are represented on the horizontal axis and are ordered in increasing positive judgement from left to right according to the third replicate. The responses for each HRTF are joined in order to represent the degree of variation between trials.

duced in Listening Test 2.2 and again in Listening Test 2.3 until a reasonable degree of reproducibility in the responses was observed.

For Listening Test 2.2, after two trials the number of HRTFs was deemed to be still too large with respect to the reproducibility of the results (see next section), and thus was further reduced to only five. This reduction in the number of presented stimuli was in line with recommendations for standardised evaluation of audio quality for multiple stimuli that exhibit intermediate levels of audio quality. These intermediate levels in audio quality, such as those produced by different compression algorithms, may be comparable to the differences between HRTFs in the listening test. It is suggested that no more than 15 items should be used for such stimuli using this type of listening test (see ITU-R BS.1534-1:2003 recommendation). If reducing the number of HRTFs did not affect the reproducibility of the responses in Listening Test 2.3 then the source of the within subject variance might be due to the stimuli itself, for example the renderings in VAS for the different HRTFs being indistinguishable from each other to the subjects. In this case, a listening test designed for the evaluation of very small audio impairments might be needed, such as the ITU-R BS.1116-1:1997 recommendation that uses a triple-stimulus with hidden reference approach (see section 3.5.2 for more details on listening test types).

7.6.1 *Listening Test 2.2 procedure*

The number of HRTFs used for Listening Test 2.2 was halved from 51 in Listening Test 2 to 26. The choice of HRTFs to include was made arbitrarily; the first 26 HRTFs recorded from the database were selected. Again, only the author took part in this version of the listening test.

7.6.2 *Listening Test 2.2 results*

As for the results from Listening Test 2, the standardised responses are shown in figure 19 for two trials using Listening Test 2.2, in which 26 HRTFs were judged; the horizontal and vertical axes representing the HRTF judged and the standardised responses respectively. The variance between the two replicates was still large, representing poor reproducibility of subject responses and possibly suggesting that the number of HRTFs was still too large, making the task too demanding.

7.7 LISTENING TEST 2.3

Given the large variance observed in Listening Test 2.2, the number of HRTFs to be judged were further reduced in order to investigate whether the variance could be reduced by using a small number of HRTFs to judge.

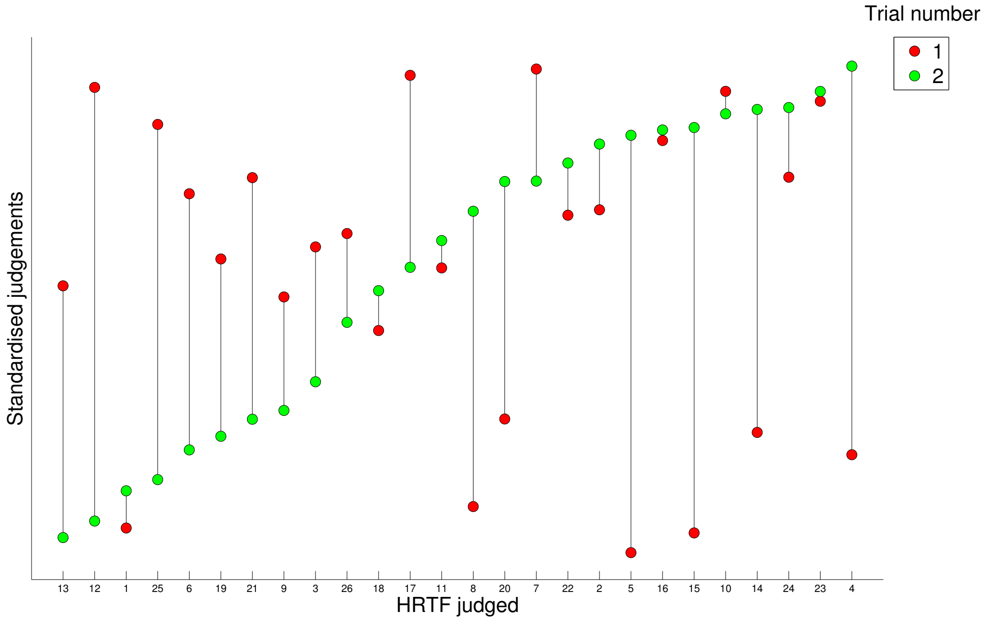


Figure 19: Two replicates of the author's responses for 26 different HRTFs using Listening Test 2.2. The standardised judgements are represented on the vertical axis increasing from bottom to top, corresponding to the end-point descriptors *incoherent* and *perfect* respectively. The judged HRTFs are represented on the horizontal axis and are ordered in increasing positive judgement from left to right according to the second replicate. The responses for each HRTF are joined by a line in order to represent the degree of variation between trials.

7.7.1 *Listening Test 2.3 procedure*

For this version of the listening test (Listening Test 2.3) the selection of HRTFs was not arbitrary; a subset of the HRTFs from the LISTEN database were selected via an analysis of the results from Listening Test 1. This selection was not performed by the author, rather completed as part of the LISTEN project. The goal of the selection was to obtain a minimal sized subset in which there would be the largest number of subjects having at least one HRTF in the subset judged as *excellent* in Listening Test 1. A procedure developed by Katz and Parseihian (2012) was used that systematically reduced the size of a subset of HRTFs using the subjects' *excellent* judgements as the elimination criteria.

The procedure began with a subset of HRTFs that satisfied all subjects. This subset was selected by choosing the most frequently judged HRTFs down to some percentage value until all subjects had at least one HRTF in the subset judged as *excellent*. From this subset, each of the HRTFs were removed one at a time and the number of subjects that no longer had any HRTFs in the subset judged as *excellent* was tallied. The HRTF whose removal caused the smallest number in the tally was then removed from the subset, and the process was repeated until some threshold of the number of subjects being satisfied in the subset was reached; in the first instance this was 100%. Thus to further reduce the size of the subset, as the calculation proceeded, a compromise had to be made between creating a small subset size and having a large percentage of subjects covered by the subset.

It is important to note that as the elimination criteria was limited in its specificity; the removal of one HRTF or another might lead to the same number of subjects being satisfied and thus create two equally effective branches of subsets. In this sense, there were multiple solutions to reducing the subset and the optimal subset was found by allowing the iteration to continue for all possible solutions 50,000 times. At the end of these iterations, the best performing subset was chosen, using seven LISTEN database HRTFs, which were labelled A, B, C, D, E, F, G, and H.

It was hoped, using this methodology, to find a somewhat perceptually orthogonal set of HRFTs that could be used to cover the majority of subjects in terms of their *excellent* judgements. Despite the fact that there was no single solution to the problem of finding the best performing subset, it should be highlighted as a point of interest that the most often selected HRTFs by the subjects were not necessarily included in the optimal subset chosen via the analysis.

A verification analysis was performed on the selected subset of HRTFs in order to have a clear picture of which subjects judged which HRTFs. The results showed that whilst there were HRTFs judged by multiple subjects in the subset, there were HRTFs from the subset that were singularly chosen by one or two subjects. It was thus decided that for the purpose of this test to bring the subset of HRTFs to only the first five (A, B, C, D, and E). This subset was still able cover a large majority of subjects in the database. These five HRTFs were used in Listening Test 2.3 for two subjects, including the author, with three replicates. Both subjects were males and had extensive experience judging audio quality. The subjects, as for the previous versions of the listening test, presented positive results for a pure tone audiometry test in accordance with ISO/IEC IS 8253-1:1989 standard.

7.7.2 *Listening Test 2.3 results*

The results for Listening Test 2.3, for the two subjects 1063 and 1088, of which the first was the author, are shown in figure 20. The plots display judgements for three replicates, in which only five different HRTFs were judged. The error variance across the HRTFs judged, representing the average spread of the responses across replicates, is displayed on each plot. There was a significant decrease in the error variance for responses when comparing Listening Test 2 and Listening Test 2.3 for subject 1063 (the only subject that took part in both versions of the listening test), with values of 1.09 and 0.83 respectively. Noting that the width of confidence intervals are proportional to the square root of the error variance, or the root-mean-square-error, the observed reduction in the variance across replicates from Listening Test 2 to Listening Test 2.3 represents an approximate 12% decrease in root-mean-square-error. The second subject showed a similar degree of variance with a value of 0.80. The results suggest that indeed a drastic reduction in the number of stimuli judged, from 51 to 5, can produce better reproducibility and more reliable responses from subjects. Despite the improvement, within subject variance was still deemed high (i.e. still almost equivalent to the standard deviation within one replicate, and having the same HRTF visibly judged as near best and near worst across replicates), and it was hoped that a further reduction could be achieved by accounting for some other possible biases in the listening test design such as the stimulus and response mapping being used.

7.8 LISTENING TEST 3

The results from Listening Test 2.3 demonstrated that more reliable judgements of HRTFs might be obtained by keeping the number of HRTFs to a reasonably small number such as five. The next phase of the study was to try and address any other biases that may be affecting the subjects' perceptual judgements related to: the stimuli used, the experience of the subjects themselves, and the response mapping using the GUI. These issues were addressed in an iterative process until the most effective methodology was achieved, producing the most reliable results.

7.8.1 *Listening Test 3 design*

The same two subjects that took part in Listening Test 2.3 performed this version of the test, subsequently known as Listening Test 3. Each subject performed three replicates of the test, and all audition of stimuli took place in the same sound dampened chamber as used in Listening Test 2.3.

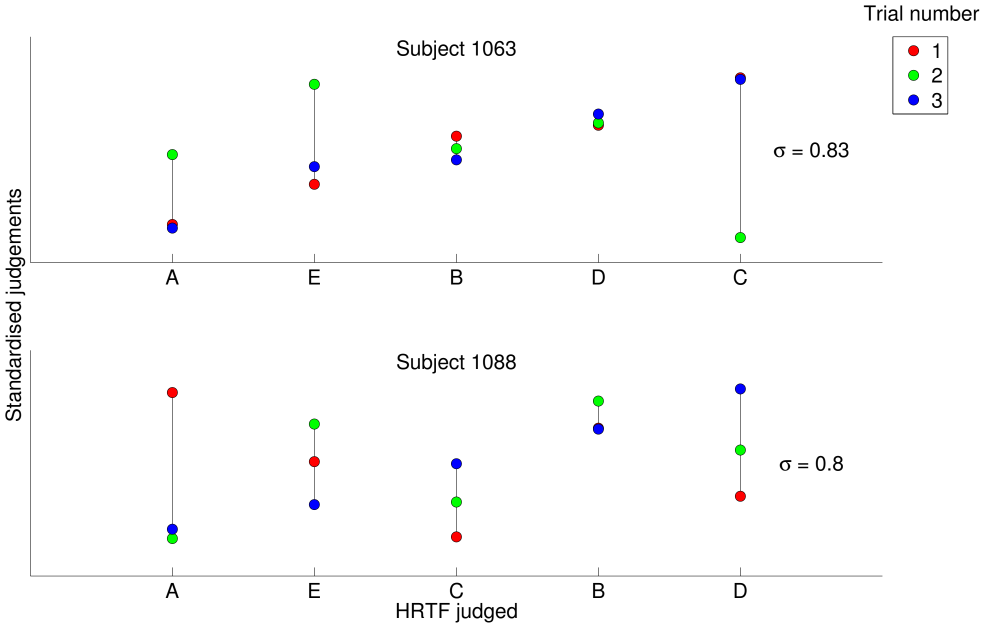


Figure 20: Three replicates of two subjects' responses for five different HRTFs using Listening Test 2.3. The standardised judgements are represented on the vertical axis increasing from bottom to top, corresponding to the endpoint descriptors *incoherent* and *perfect* respectively. The judged HRTFs are represented on the horizontal axis and are ordered in increasing positive judgement from left to right according to the third replicate. The responses for each HRTF are joined by a line in order to represent the degree of variation between trials.

For Listening Test 3, the HRTFs judged, headphones, and other hardware used, along with the procedures used for creating the renderings in VAS, were identical to those used in Listening Test 2.3. The significant difference between the two versions of the listening test was in terms of the criteria being evaluated; subjects in Listening Test 3 were asked to judge the different HRTFs in terms of three attributes separately. The choice of attributes was based on techniques that exist for generating the correct descriptive language for listening tests. The selection of the attributes used in this study was taken from studies that used a descriptive analysis methodology (Lorho, 2005a; Zacharov and Koivuniemi, 2001). In these studies, a systematic approach was taken in order to produce a descriptive language for the evaluation of spatial audio. For both studies, a panel of at least 12 listeners was chosen, trained, and tested, and then asked to produce language that consisted of attributes that corresponded to different spatial and timbral aspects of a large number of presented stimuli. Over the course of a number of weeks a discussion phase was used to create a reduced set of common descriptive language and associated rating scales, followed by a training and testing by the subjects of stimuli using this reduced number of attributes. Finally, a statistical analysis using an ANOVA and a Principal Component Analysis (PCA) was performed, to get an overview of the how these attributes are related.

In the current study, each attribute had negative and positive end-point descriptors *worst* and *best* respectively, at either end of a continuous scale. The three attributes, based on the mentioned studies and adapted for the purposes of this study, were:

- *Sense of direction*. This attribute described how well the direction of the sound source could be defined, i.e. clearly discerned and distinct. This attribute is in essence related to whether the position of the source was diffuse (*not definable*) or unambiguously at a specific position in space (*well definable*).

- *Sense of distance*. This attribute described how strongly the sensation of distance was perceived, or how ambiguous the sensation of distance was. This attribute did not relate to a perceived metric distance of the sound source, but rather how well the listener could perceive it as coming from a certain distance. Judgments would be influenced by whether or not the sound source was well externalised. A judgement near end-point descriptor *not definable* meant that the distance of the sound source was ambiguous.

- *Front image quality*. This attribute was related to the general localisation of the frontal sound trajectory. It described how well the percept of the sound coming from in front of the listener could be defined. It was specified that for a sound source that was diffuse, or perceived at a location not in front of the listener, a judgement be made at the *not definable* end-point of the scale. For a sound source that was perceived clearly in front of the listener, it was specified that a judgement be made at the *well definable* end-point of the scale.

The use of three attributes was seen as the upper limit for a listening test of this type given that each replicate was not to take longer than 30 minutes with a pause of at least one hour between replicates. The attributes *sense of direction* and *sense of distance* and their end-point descriptors, were derived from the two mentioned studies using the descriptive analysis methodology and were in line with other studies using a similar methodology and stimuli (Rumsey and Berg, 2001). *Sense of distance* was referred to as *sense of depth* in one of the studies. These attributes were shown to be descriptive in their nature; they were not attitudinal, relating to preference or natural experiences, which as described previously is important as affective judgements are prone to within subject variance.

The third attribute *front image quality* is based on ITU-R BS.1116-1:1997 recommendation for listening tests used to evaluate multichannel arrangements of speakers. This attribute was used by Rumsey (1999) for subjective assessment of surround sound processing algorithms. It was shown that conclusive results, in terms of reproducibility, could be obtained with this attribute but not for a more generalized spatial attribute termed *spatial impression*. It was also evident from the same study that conclusive results could not be obtained for hedonic judgements, i.e. judgements pertaining to whether the subject liked or disliked the sound. This finding is in line with arguments by Zielinski et al. (2008) that highlight the lack of reliability for hedonic judgements. Finally, this attribute was seen as particularly relevant to judgements of binaural synthesis quality as it is often observed that virtual sound sources in front of the listener are poorly defined (as was shown in the previous chapter for localisation tasks). These attributes were chosen with commercial applications in mind, focusing on the quality of a rendering in VAS rather than localisation accuracy. The use of three attributes was seen as a suitable trade-off between experimental efficiency, potential user confusion, and the need for relevant assessment information.

The same trajectories were used in Listening Test 3 as were used for Listening Test 2.3 (see section 7.4.1 for an explanation of why these trajectories were chosen), yet instead of presenting the two trajectories in sequence, one of the two trajectories was assigned to each specific attribute. The first two attributes used the vertical arc from Listening Test 2. This third attribute used the horizontal circle from Listening Test 2 as it described positions in space directly in front of the listener.

The use of descriptive attributes was key to generating a more robust listening test, which was easier for the listener as it reduced stimulus evaluation ambiguity. Previous versions of Listening Test 2 were seen as problematic due to the fact that there were multiple ways in which each rendering could be judged; subjects were asked to evaluate realism, externalisation, and sound source position with respect to ideal trajectories. For example, a rendering in VAS might have described a perfectly horizontal circle around the listener, but felt as if it was very close to the listener (i.e. poorly externalised; see section 3.5.3). At the same time the sound source might be perceived as diffuse in the frontal region. Subjects in this case would need to make an internal evaluation about how they weigh the different aspects of the rendering: externalisation, accurate elevation and azimuth angle, and coherence, against each other. This internal weighting is subject to change between trials and results in within subject variance. In summary, Listening Test 3 used more specific aspects of the quality of the binaural synthesis as opposed to previous versions, which used a global judgement of renderings in VAS.

Judgements in all versions of Listening Test 2 were not only problematic due to the fact that they required a weighted average across a number of locations in space; the criterion used might have been contributing to within subject variance if not sufficiently descriptive. For example, even if subjects were asked to only give perceptual judgements based on the distance between the perceived and prescribed location of the virtual sound source over the whole trajectory, there still might be a degree of variability in the results as humans are much better at judging the azimuth and elevation of a sound source than its distance (Coleman, 1962). Thus, one aspect of an attribute might dominate or mask another for a particular sensory evaluation of audio stimuli if it is not using accurate descriptive language, as suggested by Berge (2006). The same author has proposed that a measure to counteract these effects would be to exhaust all perceived attributes of the stimulus under consideration by the listener, in an effort to draw out the more subtle characteristics that might be relevant to the particular stimulus being tested. Once a large number of attributes has been generated, a process of selection can be applied in order to find the attributes to be used. In the current listening test, the attributes have been chosen to try and cover the most significant aspects of the binaural synthesis without having an excessive number of attributes, which would make the test too time consuming.

7.8.2 *Listening Test 3 interface*

Subjects made responses on three separate continuous scales, for each of the three attributes, via a GUI created by the author using numerical programming environment Matlab shown in figure 21. The end-point descriptors were labelled on each scale. Subjects selected the HRTF to be judged and depending on the attribute being evaluated, listened to either Trajectory 1 or Trajectory 2 (see figure 14). The judged HRTFs that were closest to the negative and positive end-point descriptors, labelled *Worst* and *Best* respectively, were displayed at all times on the scale for each attribute. These two best and worst HRTFs were also available to the subject to audition at any time. This allowed for a reference when judging the different HRTFs. The scale was automatically adjusted to fit these two HRTFs at the labelled end-point descriptors. Subjects were however able to zoom and pan across the scale so that they would be able to place an HRTF anywhere they wished. If a newly judged HRTF was closer to either the negative or positive end-point descriptors, it was then displayed instead and updated as one of the best or worst references to be auditioned. No other judged HRTFs were displayed on the scale. Once a subject had made a judgement for all three attributes for a particular HRTF the responses were registered by clicking on the button *Register*

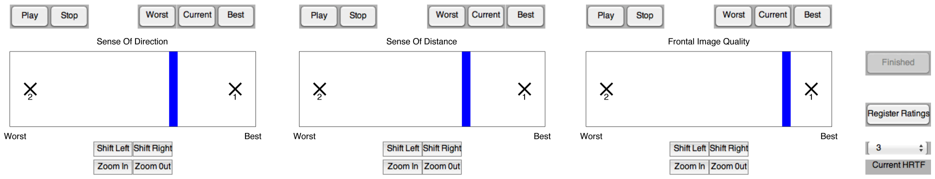


Figure 21: Graphical user interface used in the Listening Test 3.

Ratings. It was possible to audition the HRTFs in any order and return to a previously judged HRTF and change the response. The same visual aid, in the form of a video of the prescribed sound source location, was provided to subjects as used in Listening Test 2.

7.8.3 *Listening Test 3 results*

The standardised responses for the two subjects are shown in figure 22, 23, and 24, for the three attributes *sense of direction*, *sense of distance*, and *front image quality* respectively. Judgements for three replicates, for subjects 1063 and 1088, are presented. Responses for the two subjects are presented for each of the three attributes along with the error variance across the HRTFs judged. The results show that globally there was a reduction in the variance across replicates for the three attributes when compared to Listening Test 2.3 in which the same subjects, HRTFs judged, and conditions were used. The error variance across replicates for some of the attributes was smaller than for Listening Test 2.3 suggesting that better reproducibility was obtained. Subject 1063 (the author) showed a global judgement mean of 0.83 for Listening Test 2.3 (see figure 20), compared to 0.36 and 0.58 for the attributes *sense of direction* and *front image quality* respectively, in Listening Test 3. Subject 1088 showed a global judgement mean of 0.80 for Listening Test 2.3, compared to 0.69 and 0.26 for the attributes *sense of direction* and *sense of distance* respectively, in Listening Test 3 (see figure 20). The attributes highlighted are obviously different for the two subjects and chosen for their improved repeatability. This improvement could be accounted for by the fact that the attributes used were more precise and descriptive in nature leading to the use of specific strategies by the subject. The differences between the subjects might be explained by the fact that different strategies have varying effectiveness, along with the obvious differences in expertise and sensibility to the stimuli.

Further to the mentioned differences in error variance (σ), the results also demonstrated that particular HRTFs were judged more consistently than others. Interestingly, different HRTFs showed reduced variance for different subjects. For example, HRTF A showed low response variance for the attributes *sense of direction* and *front image quality* for subject 1063, and HRTF C showed low response variance for the same attributes as well as for *sense of distance* for subject 1088. This highlights the differences across subjects and how different HRTFs can be perceived as more stable for particular attributes than others.

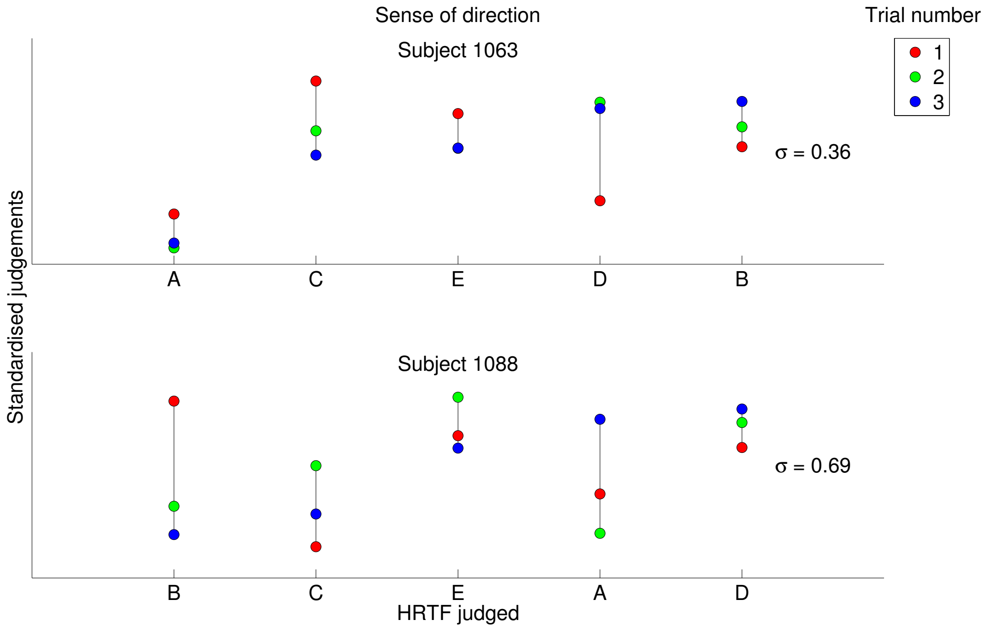


Figure 22: Three replicates of two subjects' responses for five different HRTFs using Listening Test 3 for attribute *sense of direction*. The standardised judgements are represented on the vertical axis increasing from bottom to top, corresponding to the end-point descriptors *incoherent* and *perfect* respectively. The judged HRTFs are represented on the horizontal axis and are arranged in ascending standardised judgement according to the third trial. The responses for each HRTF are joined by a line in order to represent the degree of variation between trials.

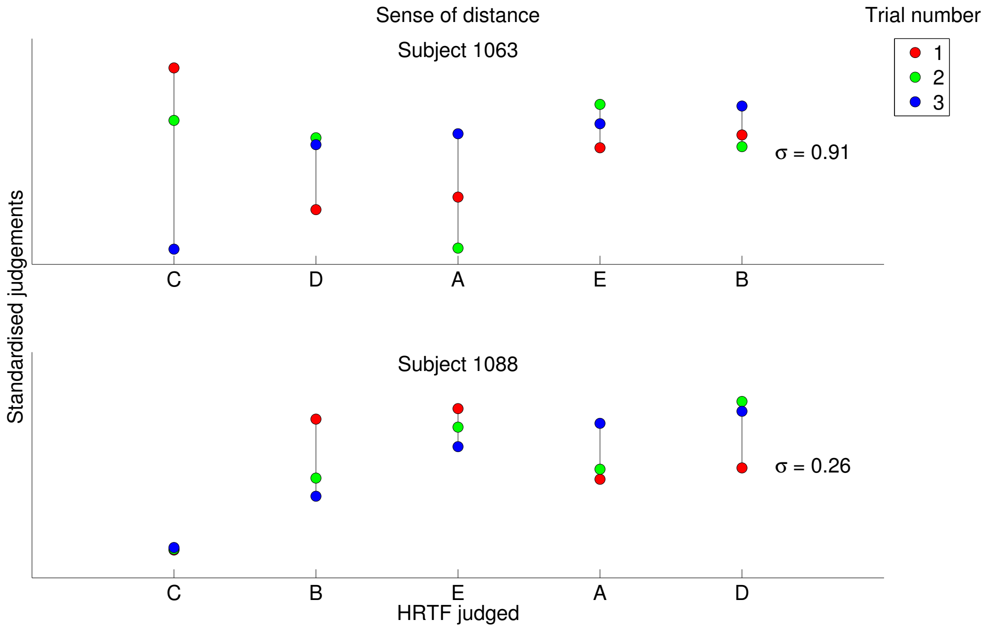


Figure 23: Same figure as 22 for the attribute *sense of distance* using Listening Test 3.

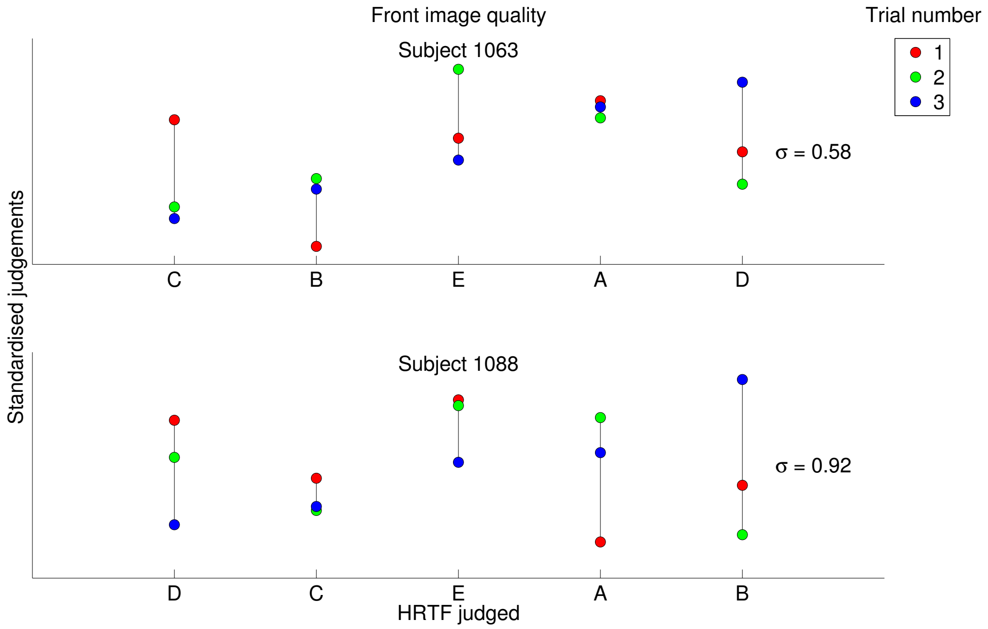


Figure 24: Same figure as 22 for the attribute *front image quality* using Listening Test 3.

7.9 listening test 3.1

Despite the results suggesting that the reproducibility of subject responses was improving using Listening Test 3, the number of replicates was still seen as small for a thorough analysis. The variance was therefore assessed for an increased number of replicates from three in Listening Test 3 to four in Listening Test 3.1.

7.9.1 *Listening Test 3.1 procedure*

The same experimental procedure was followed as in Listening Test 3, except that four replicates were performed instead of three, and a varied set of HRTFs were used. In addition, the second subject taking part was no longer the same, with the new subject having extensive experience assessing audio quality. Four attributes were assessed for each HRTF in Listening Test 3.1 as opposed to three in Listening Test 3, with the addition of a global attribute *overall rating*, in which subjects were asked to make a one additional judgement combining the other three attributes. This fourth attribute was added in an effort to establish whether judgements based on global measures of the quality of the binaural syntheses showed greater variance, as was the case for Listening Test 1 and all versions of Listening Test 2.

It was of interest to include two HRTFs that were judged by a large number of subjects as *bad*. This, it was hoped, would increase the differences across the HRTFs in the subset and lead to a perceptually simplified task with respect to Listening Test 3. Two HRTFs were replaced relative to the previous version producing the set: C, D, E, *H*, and *I*. To make the new subset of HRTFs, the two HRTFs to replace was first selected. The HRTFs A and B were chosen as they caused the lowest reduction in the amount of subjects covered by at least one *excellent* judgement when removed. Then the HRTF I was chosen as one of the replacements as it had the highest number of subjects that judged it as *bad*, whilst maintaining at least five subjects that judged it as *excellent*. This was important as to avoid HRTFs that were judged as *bad* by all subjects resulting from a possible measurement error or abnormal subject geometry. Following this, HRTF H was selected as the second replacement, as it was judged by the second largest amount of subjects as *bad*. HRTF H was also chosen because it had the largest number of distinct subjects that judged it as *bad*, i.e. sharing the smallest number of subjects that also judged I as *bad*. HRTF H had five subjects that judged it as *excellent* ensuring that it was not poorly judged globally.

7.9.2 *Listening Test 3.1 results*

As for the results of Listening Test 3, subject responses for the two subjects are displayed in figures 25, 26, and 27, for the each attribute across all replicates. In Listening Test 3.1, four attributes were used for judgements of the HRTFs and four replicates were performed. Comparisons between Listening Test 3 and Listening Test 3.1 are difficult to make given the different subset of HRTFs judged and the replacement of one of the subjects. However, it is worth noting that the smallest error variance values were observed for this version of the listening test when comparing to all other versions. Subject 1063 (the author) demonstrated an error variance of 0.28 for the attribute *sense of distance*, and subject *cd* demonstrated a significantly reduced value of 0.11 for the attribute *front image quality* (subject *cd* was not part of the public LISTEN database). These reductions were observed despite the fact that variance has a tendency to increase as the sample size (i.e. number of replicates) is increased, and then plateau for large sample sizes, for a normal distribution of values. Again, it is assumed that effective strategies were being used by the subjects to reproduce their responses across replicates.

The fourth attribute used in this version of the listening test (*overall rating*) had an error variance that was higher than the best performing attributes but not relatively higher than the worst performing attributes. This result would suggest that given the listening test required judgements of three different attributes, a global judgement might involve some form of internal average across the attributes leading to a high degree of variance given their distinct nature. A consistent strategy for weighting the three attributes across replicates may have been a challenging task, which is reflected in the error variance of the results. The mean correlation of HRTF judgements between attributes showed that, depending on the subject, the overall rating was more strongly correlated with some attributes than others. For subject 1063, the mean correlation coefficient between the attributes *overall rating* and *sense of direction* was high at 0.79, compared to the other attributes *sense of distance* (0.00) and *front image quality* (0.38). This result suggests that the *overall rating* was mostly dependent on how the subject perceived the attribute *sense of*

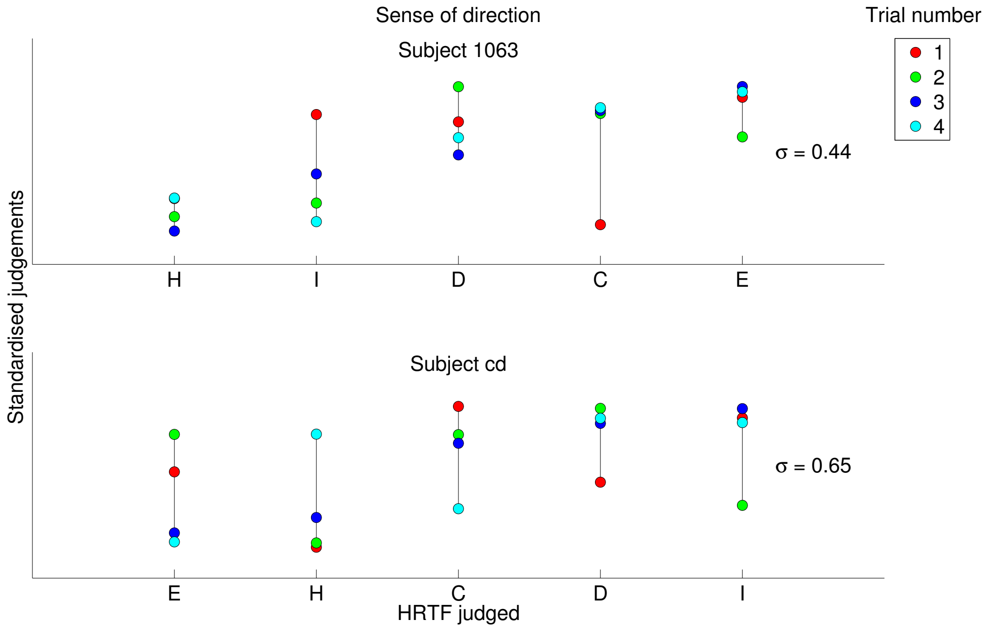


Figure 25: Four replicates of two subjects' responses for five different HRTFs using Listening Test 3.1 for attribute *sense of direction*. The standardised judgements are represented on the vertical axis increasing from bottom to top, corresponding to the end-point descriptors *incoherent* and *perfect* respectively. The judged HRTFs are represented on the horizontal axis and are arranged in increasing standardised judgement according to the third trial. The responses for each HRTF are joined by a line in order to represent the degree of variation between trials.

direction. For subject *cd*, the correlation was less pronounced with a mean maximum correlation coefficient of 0.50 between the attributes *overall rating* and *front image quality*. The other attributes *sense of direction* and *sense of distance* had mean correlation coefficients of 0.22 and 0.17 respectively, which suggests that the *overall rating* was less dependent on any one attribute and possibly more a weighted average of all attributes. It should be noted that these conclusions were based on results from only two subjects and would need more subjects in order to confirm the trend.

7.10 LISTENING TEST 3.2

Despite the improvements in response reproducibility that Listening Test 3 and Listening Test 3.1 provided, it was thought that there were still biases in the VAS renderings, descriptive language, and listening test design used. Listening Test 3.2 aimed to address these biases and was the final version in this series of iterations starting from Listening Test 1. With this final version, within subject variance was assessed, along with the effect of subject expertise.

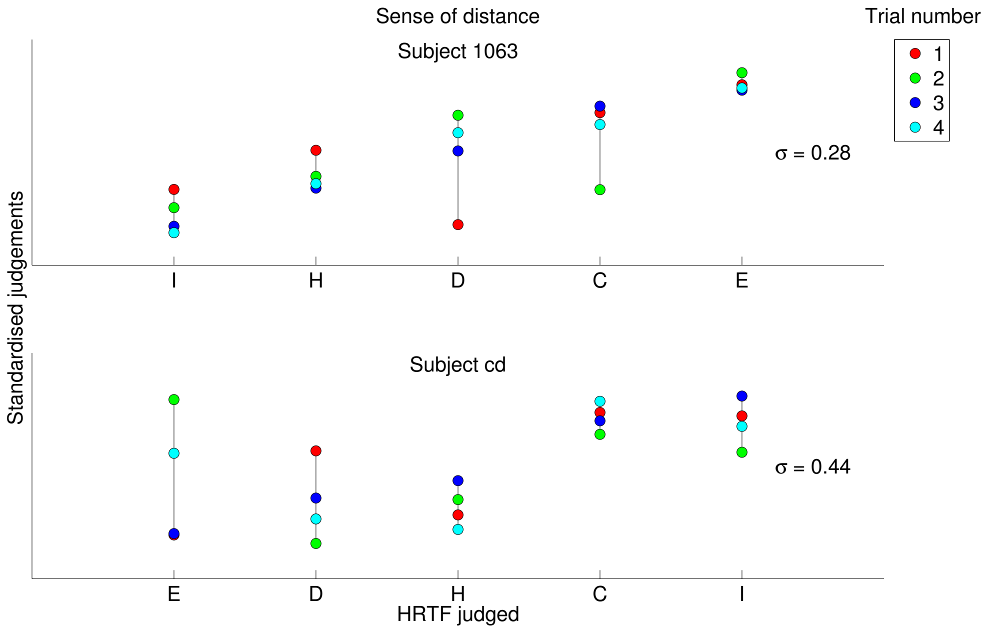


Figure 26: Same figure as 25 for the attribute *sense of distance* using Listening Test 3.1.

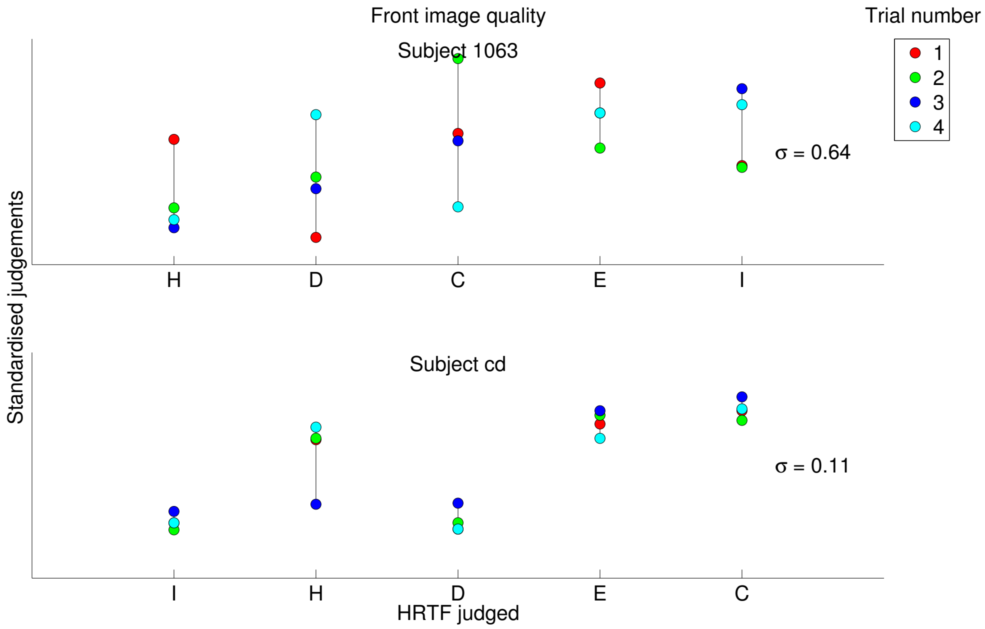


Figure 27: Same figure as 25 for the attribute *front image quality* using Listening Test 3.1.

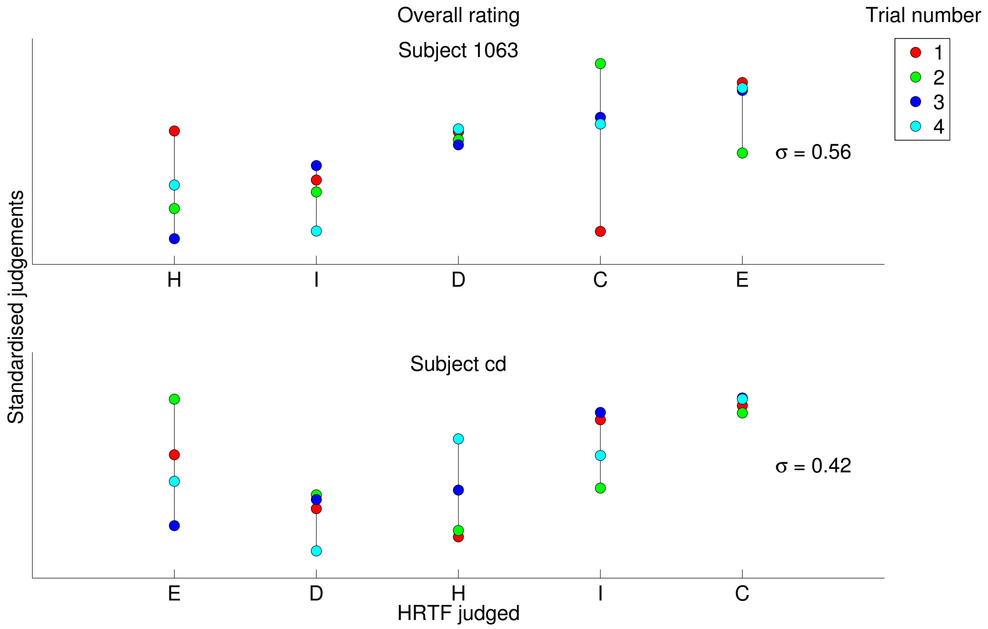


Figure 28: Same figure as 25 for the attribute *overall rating* using Listening Test 3.1.

7.10.1 *Listening Test 3.2 subject categories*

A total of six subjects (one female and five male), including the author, participated in this version of the listening test. Subjects were categorised in terms of their level of expertise as an assessor so that the relationship between expertise and repeatability could be analysed. The categorisation was based on the subjects' responses to an HTML web-based questionnaire resembling that of Mattila and Zacharov (2001). The questionnaire assessed aspects of the subjects' capabilities as an assessor of audio renderings, such as:

- Whether the subject was a musician and at what level of experience.
- Whether the subject had experience listening to binaural syntheses.
- Whether the subject had experience with listening tests using binaural syntheses.

Other questions relating to subjective judgements of the test, the interface, and strategies used, included:

- Whether the subject found the test difficult or not.
- Whether the subject used the zoom function.
- Whether the subject recognised that the same HRTFs were used in each replicate.
- Whether the subject recognised that one of the judged HRTFs was the reference HRTF (see below).

Based on the subjects' responses to the first three criteria, they were classified according to definitions taken from the ISO standard 8586-2 (ISO/IEC IS 8586-2:1994), applied by the food industry and recommended for adoption in the field of audio by Zacharov and Lorho (2006). The mentioned definitions are given in table 5. Standardized definitions were employed due to the inconsistency in the audio literature with respect to the terms untrained, naïve, experienced, and expert (Zacharov and Bech, 2006). A subject was categorised as an *initiated assessor* if they had no experience with binaural syntheses, were not musicians, and had only performed the training for the current listening test. Subjects that had prior exposure to binaural syntheses and listening tests using binaural syntheses, had performed the training, but were not musicians were categorised as *selected assessor*. Subjects with the same level of expertise as the *initiated assessor* or *selected assessor* but were in fact musicians were categorised as *experts*. This distinction between musicians and non-musicians was made in light of studies that have shown musicians to have highly specialized auditory skills (Pantev et al., 1998), and notably enhanced auditory cortical representations for specific aspects of audition such as timbre (Pantev et al., 2001). The level of musical experience was also used in a questionnaire aimed to determine the subject expertise by Bech (1992). The author of this body of work was assigned the category *expert assessor* due to the extensive experience in making judgements specific to this listening test, which was comparable to no other subject. The author also had a low level of experience as a musician. It is acknowledged that the classification of the subjects using the mentioned criteria, and the use of such a questionnaire, was not a precise procedure, yet for the purpose of this study was sufficient to gauge the expertise and suitability of the subjects for the listening test.

All subjects presented positive results for a pure tone audiometry test in accordance with ISO/IEC IS 8253-1:1989 standard, were paid volunteers, and had ages between 26 and 48.

7.10.2 *Listening Test 3.2 procedure*

The headphones used in Listening Test 3.2 were different to those used in the other versions. Intra-aural headphones designed specifically for research purposes, calibrated to produce an approximately flat frequency response within 3 dB at the level of the tymphanic membrane over the frequency range from 200 Hz to 16 kHz, were used. Accordingly, no ear-canal resonance synthesis was included in this study, or any previous study. These headphones (*ER*·2 by manufacturer Etymōtic Research) were the same used for the study in the previous chapter for localisation tasks. The change of headphone was made due to the fact that a higher fidelity rendering in VAS would be provided; the *ER*·2 headphones were shown to be the most effective headphones in terms of localisation accuracy from the eight different types tested (Schönstein et al., 2008). The *ER*·2 headphones have also been validated in terms of accuracy for localisation tasks in a number of studies (Best et al., 2005; Jin et al., 2004; Pralong and Carlile, 1994).

ASSESSOR CATEGORY	LABEL	DEFINITION
Assessor	A	Any person taking part in a sensory test
Naïve assessor	NA	A person who does not meet any particular criterion
Initiated assessor	IA	A person who has already participated in a sensory test
Selected assessor	SA	Assessor chosen for his/her ability to carry out a sensory test
Expert	E	In the general sense, a person who through knowledge or experience has competence to give an opinion in the field about which he/she is consulted (Please note that the term expert does not provide any indication regarding the qualification or suitability of the individual to perform listening tests)
Expert assessor	EA	Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products
Specialised expert assessor	SEA	Expert assessor who has additional experience as a specialist in the product and/or process and/or marketing, and who is able to perform sensory analysis of the product and evaluate or predict effects of variations relating to raw materials, recipes, processing, storage, ageing, and so on

Table 5: Summary of assessor categories employed in sensory analysis, as defined in ISO/IEC IS 8586-2:1994 standard, applied to the food industry and recommended for adoption in the field of audio. Table taken from (Zacharov and Lorho, 2006).

The test stimulus used in Listening Test 3.2 was changed slightly with respect to the older versions. In this version of the listening test, each noise burst used was freshly generated, as opposed to the same noise burst repeated continuously as in previous versions. This modification of the stimulus was performed in order to avoid any learning effects of the signal. The noise signal had a duration of 200 ms and was ramped by applying a raised cosine to the first and last 10 ms. Despite having a longer duration than noise signals used in similar studies the stimulus length was not seen to have an effect on perceptual judgements due to the fact that localisation performance has been shown to peak and plateau at approximately 80 ms for a noise stimulus (Vliegen and Van Opstal, 2004; Hofman and Van Opstal, 1998, see also section 6.2.2). The stimulus level was adjusted to 45 dB sensation level (again, see section 6.2.2 for a definition of sensation level).

The two types of renderings presented in the listening test were generated by convolving the noise signal with the Directional Transfer Function (DTF) for two specified short trajectory sequences; Trajectory 1 and Trajectory 2 (see figure 29):

1. Three positions of elevation $-15°$, $0°$ and $15°$ and azimuth $135°$ (hoop coordinates; Leong and Carlile, 1998, , see also section 2.1) starting at the position with elevation $-15°$ and ending at elevation $+15°$.

2. Three positions of azimuth $0°$, $15°$ and $30°$ and elevation $0°$ starting at the position with azimuth $0°$ and ending at azimuth $30°$.

Three positions in space, as opposed to all recorded positions for a particular angle of elevation or azimuth, or in the extreme case all 187 positions, were used in order to reduce the spatial and timbral variations in the stimulus over time. The spectrum of HRTFs vary across space and the perceived differences for non-individualized HRTFs can be substantial as demonstrated by the spatial distribution of localisation errors in the study by Wenzel et al. (1993). If variations over time are large, then listeners may find it hard to average the quality over time, which can lead to an increase in random errors (Zielinski et al., 2008). Since the purpose of the current study was to establish whether reproducible judgements of HRTFs could be obtained, a small region of space and subsequently a small number of positions was utilized as a basis, with the possibility of increasing the number of positions in future studies. The effect of reducing the number of positions being judged, along with use of attribute judgements as opposed to a global localisation judgement, were informally tested and found to reduce variance across a number of replicates. Positions behind the listener were chosen for Trajectory 1 due to the fact that these locations are often better externalised and localised; locations at the front are often perceived as coming from behind (see Wenzel et al., 1993, for example). For the attributes that aimed to assess the quality of the auditory image in terms of how coherent its location was, it was preferred to separate issues relating to the localisation of frontal sound sources. The locations in Trajectory 2 were specifically chosen as coming from in front of the listener in order to evaluate these known issues in

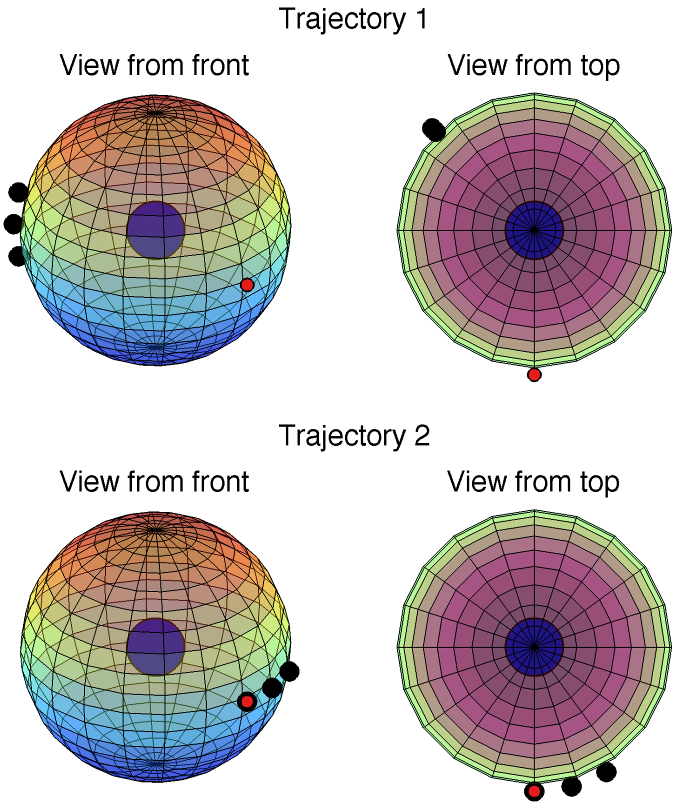


Figure 29: Trajectories used for the two test stimuli. Black circles represent rendered positions. Red circles represent the position directly in front of the listener, at azimuth 0° and elevation 0°, included in the figure as a reference. The larger blue circle at the centre of the sphere represents the position of the listener's head. Trajectory 1 corresponds to the two attributes *sense of direction* and *sense of distance*, and Trajectory 2 corresponds to the trajectory *front image quality*.

VAS renderings. Positions in Trajectory 1 did not vary in azimuth in order to evaluate the perceived positions in terms of elevation.

The use of three clearly defined positions in space, rather than using a dynamic rendering as in previous versions of the listening test, meant that interpolated HRIRs between measured positions were no longer used. This was a significant change in the experimental procedure as only measured HRTFs were used in the current listening test, assuring that the quality of the HRTF interpolation technique was not a confounding factor for subject judgements. The binaural synthesis was therefore static compared to the real-time dynamic presentation of the virtual sound source in versions beginning with Listening Test 2. Despite the fact that a dynamic binaural synthesis can offer additional cues to sound source location (Wightman and Kistler, 1999), it was decided to use a static presentation in order to avoid the use of interpolated HRTFs.

Renderings using each of the two trajectories were created for the common set of five HRTFs and for the subjects' own HRTFs. A separate stimuli set was generated for each subject taking into account the subjects' Interaural Time Differences (ITDs). The ITD processing procedure was also performed on the subjects' own HRTFs to avoid any difference in the manner in which the binaural syntheses were created for all HRTFs being judged.

Five subject HRTFs (D, E, J, F, and G) were used in this listening test corresponding to a slightly different subset than for Listening Test 3 or 3.1. This selection was made in a similar manner as the selection used for previous versions; the selection was based on the results from Listening Test 1 in terms of how many times a particular HRTF was judged as *excellent*. All possible combinations of five HRTFs were evaluated in terms of the mentioned perceptual judgements and the subset of HRTFs that satisfied the largest number of subjects (42 out of 45), in terms of *excellent* judgements, was selected. The method for selecting the subset was changed due to the fact that the number of HRTFs (five) was predetermined due to the observations on error variance from previous versions of the listening test; the previous analysis for subset selection was an algorithm that tried to find the smallest set of HRTFs of any size. In order to ensure that the HRTFs selected in the subset were not poorly rated globally, potentially relating to possible measurement artifacts, the subset with the largest minimum number of *excellent* ratings (a value of nine) for individual HRTFs was chosen. In addition to the five mentioned subject HRTFs, each of the subject's own HRTFs were included.

All HRTFs were individually converted to DTFs (see Middlebrooks and Green, 1990) by dividing out the spatially weighted average of all 187 measurement positions (see section 3.2). The use of DTFs which involves a diffuse field equalisation differs from previous versions of the listening test, which used raw HRTF recordings with no equalisation. It was necessary to use an equalisation due to the fact that the subjects' own HRTFs, which were included as a reference, represented HRTF recordings that used a different setup (namely different speakers and microphones) to those recorded as part of the original LISTEN database. Since differences in the way HRTFs are recorded can have significant effects on perceived renderings in VAS, it was crucial to perform an equalisation so that all the HRTFs used in the listening test were comparable. An equalisation of this sort removes the transfer function of the recording system as DTFs do not include non-directional filtering effects, such as the frequency response of the speaker used. The subjects' individual HRTFs, used for the binaural synthesis in this study, were recorded using one speaker out of a possible three, and the same locations in space as for the LISTEN database.

The DTFs were also normalised in root-mean-square across all positions for the left and right ear for all subjects. The DTFs were decomposed into the minimum phase and all-pass components. The all-pass component was replaced with a pure delay that represented the ITD of each subject. The ITD was calculated using the Maximum Inter-Aural Cross-Correlation (MaxIACC) of the energy envelopes of each subject's individual HRTFs. The stimuli sets, corresponding to each HRTF judged, as each subject used identical ITD values for each respective position, removing ITD as a test variable across HRTFs for each subject.

The type of listening test used in Listening Test 3.2 was based on the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method (see for example Macpherson and Middlebrooks, 2000), standardised in ITU-R BS.1534-1:2003 recommendation. This type of listening test was developed in order to evaluate audio systems exhibiting intermediate levels of audio quality as previously mentioned, and is recommended that no more than 15 items be included in any one trial. As the name suggests, there is a hidden reference included as one of the renderings being assessed, along with a signified reference and anchor (i.e. labelled on the interface for the subject). The reference was in this case a binaural synthesis using the subject's own HRTFs. The anchor was a degraded version of the reference; 3.5 kHz low-pass filtered using a 200-point FIR filter in compliance with the MUSHRA standard. This listening test method incorporates some additional features with respect to previous versions, particularly in terms of the reference and anchor that aids the mapping of judgements. The labelled reference and anchor can be considered as determining the end-points of the continuous scale, and calibrating it as they set permanent "yardsticks", to use the language of Guilford (1936). Subjects were told that the positive descriptor end-point corresponded to the reference and that the anchor should correspond to a judgement near, but not necessarily at, the negative end-point descriptor. The reference and anchor help to reduce a concentration of responses on the scale, known as a contraction bias, which can lead to a poor resolution and inaccurate judgements (Lipshitz and Vanderkooy, 1980), as they allow for listeners to use the entire range. Finally, it is known that the use of a reference and anchor can help to design experiments that will achieve a high degree of repeatability as it reduces a bias in which subjects do not use the whole scale for responses, known as the range equalising bias (see Marston and Mason, 2005, for an example using the same MUSHRA listening test type as in this study).

Each HRTF was presented to the listener along the two previously described trajectories, Trajectory 1 and Trajectory 2 (see figure 29). Trajectory

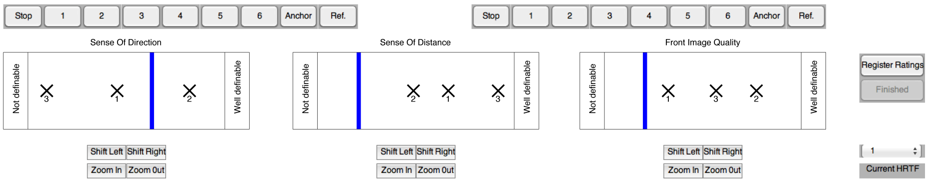


Figure 30: Graphical user interface used in the Listening Test 3.2.

1 (positions varying in elevation) was presented for judgements of the first two attributes (*sense of direction* and *sense of distance*), while Trajectory 2 (positions varying in azimuth) was presented for judgements of the third attribute (*front image quality*). The GUI, shown in figure 30 was designed so that judgements for all six HRTFs would be made on the same continuous scale for each attribute, rather then best and worst judged HRTFs. The end-point descriptors *Best* and *Worst* from Listening Test 3.1 were replaced by the more descriptive terms *Well definable* and *Not definable*. The change of end-point descriptors was part of modifications performed across previous versions of the listening test to remove language that might invoke hedonic judgements from the listener as it is known that this leads to increased response variance, as detailed in section 7.4.1. Therefore, instead of having available the best and worst judged HRTFs for audition at all times as in Listening Test 3.1, the reference and anchor were presented instead.

Subjects selected an HRTF to judge from a drop-down menu and once it had been judged for all three attributes the responses were registered by the subject. After registering the judgements for a particular HRTF, its corresponding number (an integer between 1 and 6) was displayed on the continuous scale. Playback of any of the six HRTFs, presented as buttons numbered from 1 to 6, was permitted at any time during the listening test. Subjects could only play one trajectory for one of the HRTFs at a time. The numbered buttons on the left of the GUI were for the Trajectory 1 and the buttons on the right were for Trajectory 2. Along with the six different HRTFs, there was a button for the specified reference and anchor. Each profile corresponded to a continuous loop of the signal noise convolved with the specified DTF for the given trajectory, as was the case for the reference and anchor (except that the anchor was a degraded version of the reference). Subjects could make judgements of the HRTFs in any order and return to a previously judged HRTF to change a response. The same zoom and pan functions were available to the subjects as for previous versions of Listening Test 3.

Each subject was individually familiarised with the equipment and the test procedure. The trajectories were displayed to the subject, prior to making any judgements as they are presented in figure 29. Particular attention was paid to the definitions of the attributes to be used; the definitions of each attribute as detailed in section 7.8.1 were presented to the subjects during training using the exact same wording, yet in French rather than English (see Appendix A). It was highlighted that judgements were not to be made with respect to the position of the sound source for the three attributes, but rather with respect to whether the attribute was *well definable* or *not definable*, as specified in section 7.8.1.

Given the known importance of listener training, due to its ability to increase listeners' sensitivity to auditory characteristics and ability to reliably judge their perceptions (Bech, 1992; Gabrielsson et al., 1974; Watson, 1980; Neher et al., 2002; Olive, 2003), all subjects participated in a training session before starting the five replicates. Subjects were given as long as they needed to familiarise themselves with the interface. Subjects were told to continue listening to the six HRTFs until they were able to discern the differences between them for each attribute. No objective evaluation of the training, in terms of whether the subjects could discern the differences among the HRTFs, was performed. No visual aid was given to the subjects as part of the GUI used in Listening Test 3.2 due to the small number of positions used; subjects were shown the desired positions, using a three-dimensional image (figure 29), for the two trajectories as part of the training.

7.10.3 *Listening Test 3.2 results*

An analysis of subject responses was framed around the question of whether HRTFs could be reliably judged using the given optimal listening test design. Subjective reports, perceptual responses, assessment times across replicates, along with subjects' level of expertise, were all assessed in an effort to distinguish the degree of reproducibility of judgements made, and the possible factors that influence reproducibility.

7.10.4 *Listening Test 3.2 subjective reports*

Subjects responded unanimously via a questionnaire following the listening test that they found the task difficult. Subjects reported that the attribute *sense of distance* was the most challenging to evaluate due to difficulty in judging the distance of the sound source, with some subjects admitting that judgements were almost random. This finding is in line with previous studies that have shown a fairly large degree of variability in distance judgements for a fixed sound source distance in VAS (Zahorik, 2002). Two of the subjects, including the author of this body of work, reported to having been aware that one of the HRTFs being judged was a hidden version of the reference. None of the subjects, except the author of this body of work, reported to have been aware that the HRTFs used in each trial were the same. The author of this body of work was included in the analysis of the results despite the biases stated due to the interest of having an *expert assessor*. Most of the HRTFs judged were not perceived by the subjects as coming from in front for the horizontal trajectory (Trajectory 2). Varying strategies were reported by the subjects for the evaluation of the renderings with respect to the three attributes, supporting the notion that this could be a source of both inter and within subject variation.

7.10.5 *Listening Test 3.2 reproducibility of responses*

The ISO/IEC IS 5497:1982: standard defines repeatability, or reproducibility, as the closeness in agreement between mutually independent test results obtained under conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time. This definition of repeatability has been strictly adhered to except with respect to the laboratory used, as the listening tests were conducted at two different locations due to the fact that the subjects were based at different locations. However the setting of both were for all intents and purposes identical; they were both acoustically dampened and isolated chambers. The measure of repeatability was understood to be at the level of the listener for this study.

Globally, the variation in perceptual judgements of the different HRTFs across the five randomised replicates was significant. Scaling differences, which exist when listeners use different amounts of the scale for judgements, were observed between subjects. As for the previous versions of the listening test, responses were scaled so that they could be compared across subjects. More precisely, the judgement data for each subject was scaled to unit variance and zero mean. Figure 31 shows the HRTF judgements using the standardised data for each attribute. Judgements for the attributes (a) *sense of direction*, (b) *sense of distance*, and (c) *front image quality* are displayed from top to bottom, and presented as box plots. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points as a true representation of the spread of responses was preferred. The subjects are represented on the horizontal axis and the judged HRTFs are marked by colour. The subjects' own HRTF is labelled *Subject* in the legend. Subjects are ordered from left to right from most variance to least, calculated across all attributes. The standardised judgements are represented on the vertical axis with the labels *Well* and *Not* corresponding to the end-point descriptors *well definable* and *not definable* respectively.

From these results it can be seen that the spread of judgements for a particular HRTF varies significantly between subjects. For example, HRTF 1058 (blue in the figure) was reliably judged by subjects 1089 and 1064 (represented by a small box), and relatively unreliably judged by subject 1079 (represented by a large box). This shows that judgement reproducibility was subject specific and not related to the HRTF being judged. The same pattern was observed for Listening Test 3 and discussed in section 7.8.3.

Also evident is the correlation between judgements for attributes *sense of direction* and *sense of distance*; the mean correlation coefficient across all subjects was 0.74. The correlation between the other pairs of attributes was less pronounced with a mean value of 0.46. The attribute that exhibited the least variance in judgements between replicates for all subjects was that of *front image quality*. This attribute also showed the lowest error variance in Listening Test 3.1 of all the attributes and subjects tested. The attributes *sense of*

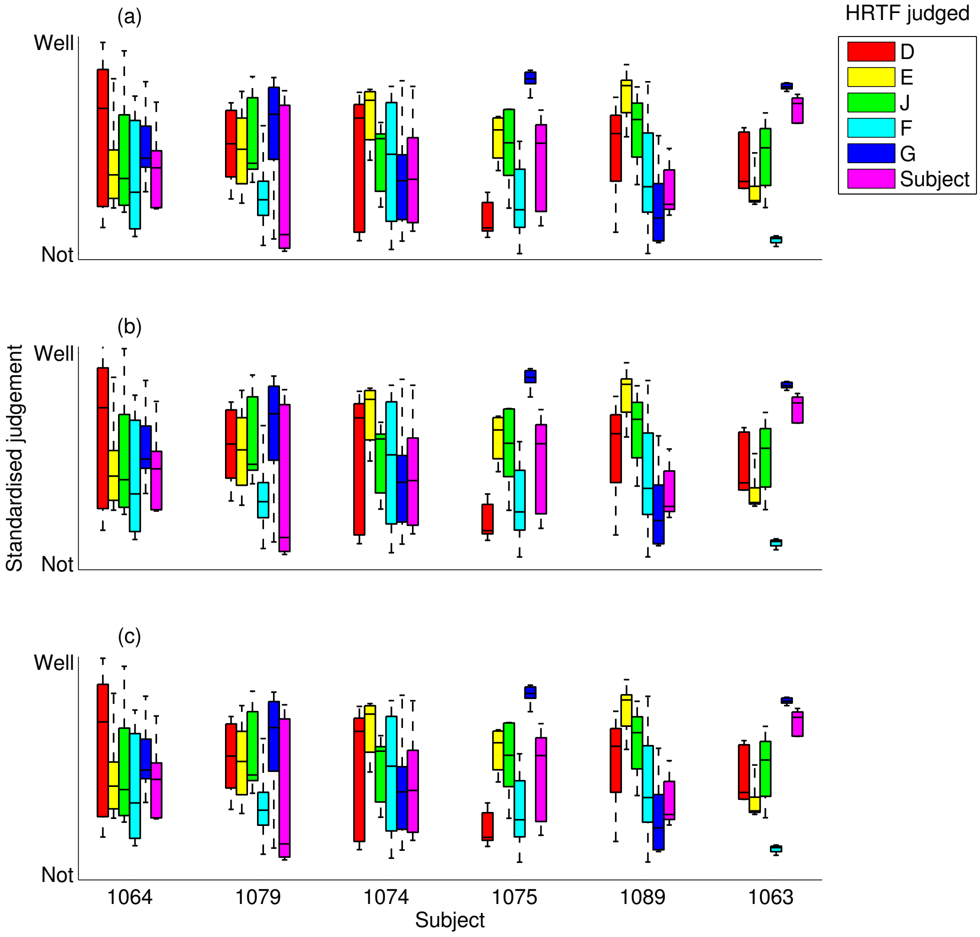


Figure 31: Box plots representing the spread of judgements for HRTFs for the attributes (a) *sense of direction*, (b) *sense of distance*, and (c) *front image quality*. Subjects are represented on the horizontal axis and the HRTF judged is determined by colour and displayed on the legend. Judgements are represented on the vertical axis between the labels *Well* and *Not* corresponding to the descriptor end-points *well defined* and *not definable* respectively.

direction and *sense of distance* had similar amounts of variance. These findings, taken together with the results from the questionnaire, would suggest that the attribute *sense of distance* is not a particularly reliable one and most probably redundant when also judging *sense of direction*. Also noteworthy is the fact that, for subjects whom the variance was low, the subject's own HRTFs (the hidden reference) were not always judged as being the most *well definable*; see for example responses for the attribute *sense of direction* for subjects 1075 and 1089. This type of result has also been shown in a study by Usher and Martens (2007) where the subjects' own HRTFs were not judged as being perceived as the most natural sounding. This result has a direct impact on methods that aim to provide a listener with individualized HRTFs (Guillon, 2009; Katz, 2001; Middlebrooks, 1999b), since it is possible that an HRTF other than the subject's own will be judged better. It is noted again that the current study employed a minimum-phase DTF representation of the measured HRTFs, which may have influenced how the subjects' own HRTFs were perceived. What characteristics of a particular HRTF lead to a favorable judgement is still not entirely clear in this context. Most studies that seek to single out the important aspects of the HRTF have assessed only accuracy in localisation tasks (Jin et al., 2004; Kulkarni and Colburn, 1998; Langendijk and Bronkhorst, 2002; Martens, 1987).

7.10.6 *Listening Test 3.2 subject expertise*

Given the assignment of assessor categories from table 5, the correlation between level of expertise and variance was evaluated. The correlation coefficients for the attributes *sense of direction*, *sense of distance* and *front image quality* were 0.73, 0.69 and 0.77 respectively. Whilst the correlations were not statistically significant at the 95% confidence level, the findings tend to support the notion that subject training and experience can produce significant benefits including improved reproducibility (Bech, 1992; Watson, 1980; Neher et al., 2002; Olive, 2003). Particularly relevant is a study by Gabrielsson et al. (1974) that showed differences in response reliability between subjects categorised as either "listeners in general", "musicians", or "hi-fi subjects".

It is also noteworthy that in the current study some subjects showed the same amount of variance across all three attributes and other subjects had large differences as shown in table 6. For example, subject 1074 showed very low variance for the attribute *front image quality* (0.03) and high error variance for the attributes *sense of direction* and *sense of distance* (0.89 and 0.94 respectively). Subject 1063 (the author) had a low error variance for all attributes: error variance of 0.14 for *sense of direction*, 0.09 for *sense of distance*, and 0.1 for *front image quality*. The results demonstrate that reproducibility is not subject specific, rather related to subject expertise in conjunction with how difficult the subject found judgements for a particular attribute and whether a consistent strategy was developed. In addition, it supports the idea that subjects might perceive some HRTFs as being better for some attributes of the rendering than others. A similar finding has been shown by Boren and Ro-

subject	1064	1079	1074	1075	1089	1063
Sense of direction	0.96	0.83	0.89	0.33	0.54	0.14
Sense of distance	0.92	0.81	0.94	0.3	0.54	0.09
Front image quality	0.97	0.88	0.03	0.71	0.23	0.1

Table 6: Error variance across replicates for each subject and attributes *sense of direction*, *sense of distance*, and *front image quality*.

ginska (2011) in which judgements varied across the different criteria such as externalisation, elevation, and front/back discrimination.

To further investigate the role of subject expertise on reproducibility, the error variance of HRTF judgements for a particular attribute across all five replicates was calculated from the standardised data for each subject. A one-way ANOVA was performed on the variance values for each attribute to test the effect of subject on repeatability. For all three attributes, subject was a statistically significant factor ($p < 0.01$). Subject error variance across all judged HRTFs is shown in figure 32 along with the results from section 7.10.1 corresponding to listener's level of expertise. The error variance is represented on the horizontal axis and the assessor category on the vertical axis. The labels on the vertical axis correspond to the labels for assessor categories from initiated assessor to expert assessor (labels IA to EA respectively, taken from table 5). The colour of the circles along with the legend, determines what attribute the error variances correspond to. The points in the figure are also labelled in terms of the subject making the judgement. Each circle on the figure thus represents the variance across all trials for one attribute for a particular subject.

The results show that while there were some differences in the observed error variance across attributes for assessor category *experts* (labelled E), the broad categories used were effective in predicting the reproducibility of the subjects in this experiment. The mean error variance was high for categories *initiated assessor* (labelled IA) and *selected assessor* (labelled SA) with values 0.84 and 0.95 respectively, lower for the category *experts* (labelled E) with a value of 0.5, and lowest for category *expert assessor* with a value of 0.11.

7.10.7 *Listening Test 3.2 analysis of judgement time*

The number of auditions for each replicate, measured as the total number of times the subject listened to the HRTFs including repeated auditions of the same HRTF, was calculated and is shown in figure 33(a). The time taken to complete each replicate for all subjects was also measured and is shown in Figure 33(b). The mean time was approximately 16 minutes (standard deviation of 4 minutes) and the mean number of auditions across all replicates and subjects was approximately 8 (standard deviation of 3).

A one-way ANOVA showed that replicate number was the only statistically significant factor ($p < 0.01$) with respect to the time taken to judge all HRTFs and the number of auditions across all HRTFs. The results of the ANOVA

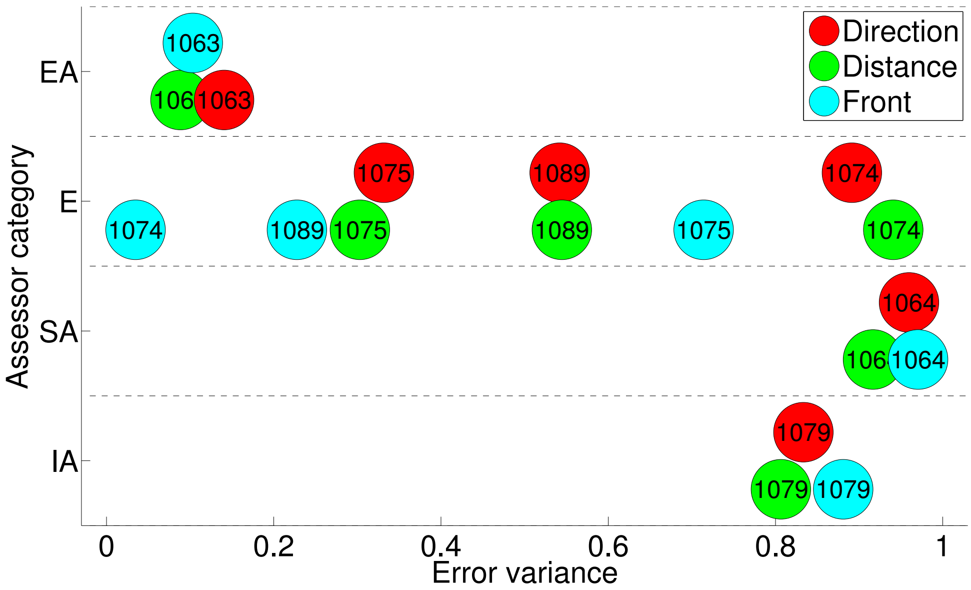


Figure 32: Subject error variances across replicates, represented on the horizontal axis, for each attribute judged as a function of assessor category (see table 5), represented on the vertical axis. Attributes judged correspond to the colour of the circles and are labelled in the legend. Each circle is labelled with a number that corresponds to the subject that made the judgements. Circles vary slightly in terms of their vertical position within a particular assessor category for ease of view.

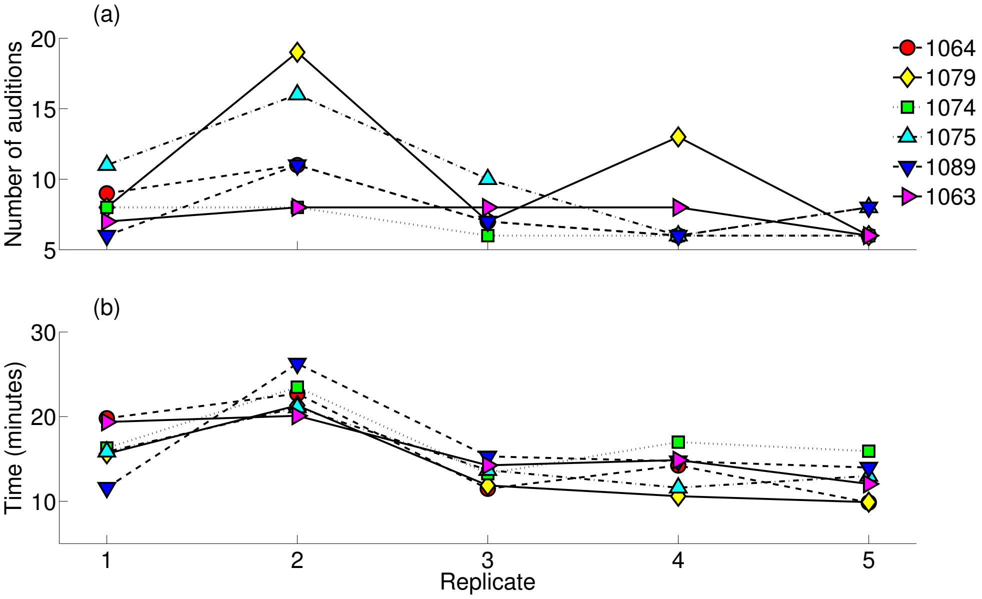


Figure 33: (a) Number of auditions and (b) time taken for each replicate performed. The time taken in minutes and the number of auditions is presented on the vertical axis and the replicate number on the horizontal axis. The different marker shapes correspond to the different subjects.

showed that neither the HRTF judged nor the subject judging was a significant factor on either of these measures. In terms of the time taken to assess all HRTFs, most subjects showed a longer time for first and second replicates and then a decline in the time taken. This trend most probably highlights a growing familiarity with the test design and task, which made it faster to complete the listening test.

In order to assess how the variability in responses changed across replicates, the rank of judged HRTFs was analysed; a rank of one being the highest and six the lowest. Figure 34 shows how much each individually judged HRTF changed in rank from one replicate to the next for the attributes (a) *sense of direction*, (b) *sense of distance*, and (c) *front image quality*. The vertical axis shows the cumulative mean in change of rank, calculated as the cumulative sum of the change in rank for all the HRTFs judged by one subject divided by the total number of replicates. The change in rank from one replicate to the next for a subject was calculated by taking the difference in rank for each HRTF and summing the difference values. For example, if in replicate number one a particular HRTF was ranked as number one (the best) by a subject and then ranked as number six (the worst) in the following replicate the magnitude of the change in rank was taken as five, i.e. six minus one, for that HRTF. The change in rank was summed across all HRTFs judged for each replicate and each subject. The horizontal axis represents the number of replicates so that a value of two corresponds to the change in rank from replicate one to two, and so on.

The results in figure 34 for the change in rank show that despite the changes in time and number of auditions shown in figure 33 (i.e. the rise in magnitude for replicate number two) the variability in judgements for most subjects did not change, and more specifically did not improve (represented by a decrease in values over replicate number), across replicates in terms of subjects' rank of the different HRTFs. This would suggest that no learning effects took place across the replicates and that the error variance represented in figure 32 across all replicates is a valid indicator of the subjects' ability to provide reproducible perceptual judgements of different HRTFs. The largest change in variability across replicates was observed for the attribute *sense of direction*; an overall decrease in values occurred for subject 1074 (an improvement in reproducibility), and an increase of six occurred for subject 1089 (a decline in reproducibility). This might suggest that these subjects either developed a strategy for judging HRTFs for a particular attribute in the case of an improvement, or became increasingly unsure with how to judge the HRTFs in the case of a decline. These trends were however somewhat isolated and not observed across all attributes for any particular subject as can be seen in table 7.

7.11 DISCUSSION

The progression from Listening Test 1 through to Listening Test 3.2 was driven by the need for reliable perceptual judgements of renderings in VAS using different HRTFs. As shown in figure 35 the error variance across sub-

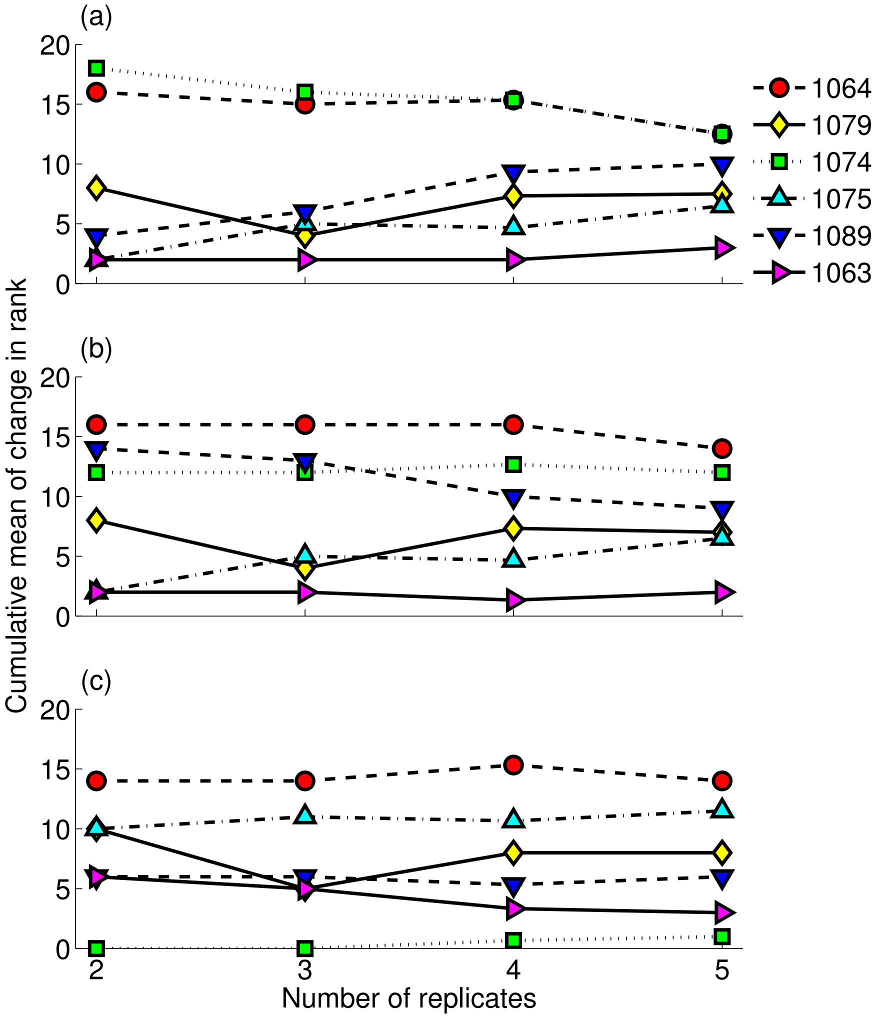


Figure 34: Change in rank from one replicate to another for judged HRTFs, for each subject, for the attributes (a) *sense of direction*, (b) *sense of distance*, and (c) *front image quality*. The vertical axis is represented by the cumulative sum of change in rank across all HRTFs divided by the total number of replicates. The horizontal axis corresponds to the number of replicates performed so that the value of two represents the change in rank from replicate one to two, and so on.

SUBJECT	1064	1079	1074	1075	1089	1063
Sense of direction	-1.02	0.18	-1.72	1.32	2.13	0.3
Sense of distance	-0.6	0.03	0.07	1.32	-1.8	-0.07
Front image quality	0.13	-0.3	0.37	0.42	-0.07	-1.07

Table 7: Table of slope values for a linear regression of the cumulative mean of change in rank values shown in figure 34. Slope values are shown for each subject by attribute across replicates.

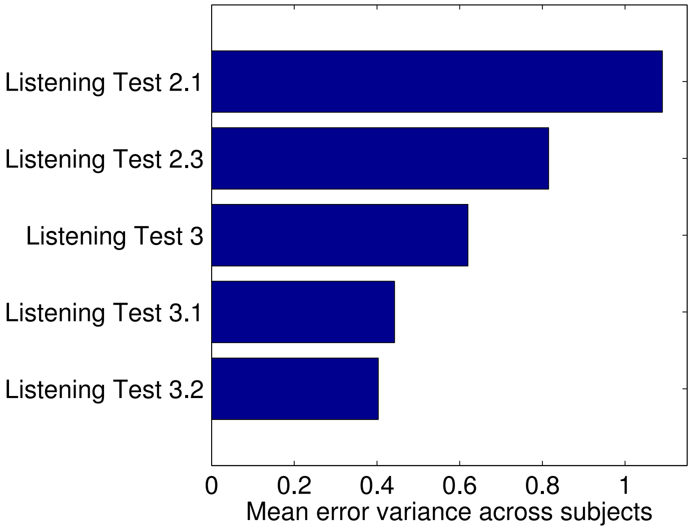


Figure 35: Mean error variance across subjects for the different versions of the listening test. Note for Listening Test 3.2, only subjects categorised as either an *expert* or *expert assessor* were included.

jects was reduced at each iteration. Note that for Listening Test 3.2, only subjects categorised as either an *expert* or an *expert assessor* were included in order to have a somewhat consistent level of expertise across the different versions of the listening tests; subjects performing the listening test from Listening Test 2.1 to Listening Test 3.1 were all familiar with judging audio quality for renderings in VAS. Whilst the comparison across the different versions of the listening tests has limitations due to fact that there were different subjects participating in the experiments, it does allow for a guide of how the different iterations helped to improve repeatability.

With respect to the individual iterations themselves, the switch from category type judgements in Listening Test 1 to a continuous scale in Listening Test 2 allowed for the use of more powerful statistical and a meaningful analysis across subjects. The continuous scale meant that responses could be scaled and translated for further analysis, which was desirable in order to rank all judged HRTFs for each subject, yet there were limitations to this design. Listening Test 2 was a significantly more demanding task than Listening Test 1 considering subjects only had to choose from three broad classes for the judged renderings, corresponding to judgements of *excellent*, *fair*, or *bad*, in Listening Test 1, as opposed to a number of classes that was effectively the same number as there were HRTFs to judge (i.e. the continuous scale). This greatly increased the number of comparisons to be auditioned, and consequently made the task very demanding.

The different versions of Listening Test 2 aimed at testing the reproducibility of subject responses by assessing error variance across multiple replicates for the same HRTFs judged. Error variance was shown to somewhat reduce as the number of HRTFs to be judged was reduced, leading to a less demanding task. It is known that subjects rely heavily on comparisons in descriptive sensory analysis given that as humans we are very poor absolute measuring instruments; in comparison we are very good at comparing stimuli (Lawless and Heymann, 1999). The reduction in the number of HRTFs to judge meant that there was a very large reduction in comparisons between VAS renderings needed to be auditioned in order to place each HRTF on the continuous scale. The final number of HRTFs was reduced to five from 51, which represented a number of stimuli that was well below to the recommended maximum of 15 items for multistimulus listening tests (ITU-R recommendation BS. 1543-1).

Efforts were also made in the different versions of Listening Test 2 to progressively use more descriptive terms in the labelling of end-point descriptors on the continuous scale (Lorho, 2005b). Language used for the categories in Listening Test 1 such as *excellent*, *fair*, or *bad*, which have an affective component to them were avoided due to the known biases involved with their use (see for a review Zielinski et al., 2008).

Listening Test 3 and subsequent versions were shown to provide a further improvement in the reproducibility of subject responses by using carefully selected descriptive attributes, from a technique that used a descriptive analysis methodology, to be judged rather than a more global measure. The more descriptive nature of the attributes being judged may have made it easier for subjects to develop a strategy for evaluating renderings, which would translate into better reproducibility of responses across replicates.

A prominent feature of later versions of Listening Test 3 was the addition of signaled references, in the form of specified renderings available for audition, in order to aid the comparison of the different HRTFs. The references, either the subject's best and worst judged HRTF or their own HRTF and a degraded version of it (anchor), were designed to present the subject with a what could be termed as "yardsticks" (Guilford, 1936) or *goal posts*. These stimuli set the range of the continuous scale for each judged HRTF, and provided the same set of stimuli to use for comparisons across judged HRTFs, subsequently making responses more reliable across replicates. One version of Listening Test 3 (Listening Test 3.2) also had a largely reduced number of locations in space to judge for each trajectory. This reduced the variation in the auditioned stimulus over time (i.e. only three virtual sound sources), caused by the perceived changes in the spectrum for non-individualised HRTFs at different locations, making it easier for listeners to judge the fidelity of the VAS rendering. It should be noted that by only using three locations in space for the listening test, the judgements made were not representative of all the HRTFs across the whole measured space. The limited number of locations was used as a basis to establishing whether at this reduced representation of space reproducible responses could be obtained. Having shown that it is possible to produce reliable judgements, future work should aim to increase the representation of space whilst maintaining the observed levels of error variance.

The optimal listening test design, which had aimed to reduce known biases, was used to test the effect of listener experience on judgement reproducibility and to some extent the effect of listener training. Subjects were loosely classified based on their musical experience and exposure to listening tests using binaural syntheses, and these classifications were shown to correlate with response repeatability. This result is in line with early studies, such as that performed by Gabrielsson et al. (1974) looking at judgements of sound quality by subjects classified as either "listeners in general", musicians, or "hi-fi subjects", that showed differences in the reproducibility of results for the subject categories. More recent studies, assessing the effects of training subjects for the evaluation of spatial sound reproduction, similar to tasks used in the current study, have demonstrated that reproducibility is worse for naïve listeners when compared to an experienced listening panel, and that these differences can be reduced with training (Neher et al., 2002). The effect of listener training was not strictly analysed in this study, however it was suggested that little learning took place as the change is HRTF rank was shown not to improve across replicates.

In terms of the optimal listening test (Listening Test 3.2) results, given the large magnitude in error variance observed for listeners with little expertise it would be difficult to determine an optimum HRTF in a listening test setting. It is therefore recommended that only experts be used for the evaluation of processing or selection methods. A questionnaire as used in this version of the listening test (see section 7.10.1) could be used as a pre-selection for subjects in order to determine their suitability for the type of listening test being used, with a post-selection performed based on subject response reproducibility, as suggested by Zacharov and Lorho (2006). A cutoff value could be used in the post-selection phase in order to determine which subjects were reliably judging the different HRTFs as proposed by Gabrielsson (1979). Gabrielsson suggested that a measure termed *reliability of mean ratings*, calculated as simply one minus the F-ratio (the ratio of explained over unexplained variance) of the ANOVA across replicates, be used as such. It was concluded by the author, based on experience, that a value of 0.5 should be considered as a lower limit for acceptable reliability of mean ratings. A similar methodology could be applied to results from Listening Test 3.2 in order to determine which subjects were reliable and which were not in terms of their judgements of the different HRTFs. In the current study, a cutoff value of 0.5 would result in only 7 of the 18 sets of replicates (i.e. one attribute and one subject) being suitable for analysis.

This proposed measure of reliability based on mean error variance is, of course, not relevant to HRTF selection applications for consumer markets in which the user needs to make an evaluation and HRTF choice. The key findings drawn from the results of Listening Test 3.2 apply to perceptual tests and it is recommended that subject reproducibility, along with expertise, be taken into account using replicates of the test if conclusions are to be drawn from perceptual judgements. It is also recommended that the test design characteristics detailed for this version of the listening test be considered in order to reduce the effects of judgement variability and listening test biases.

Finally, the results from Listening Test 3.2 showed that learning effects can be avoided if adequate subject training is provided, in terms of the variability in how subjects ranked the different HRTFs, which is applicable to perceptual tasks of any kind and supports studies showing the benefits of subject training.

7.12 CONCLUSION

The current study was an iterative process that explored whether reliable perceptual judgements of binaural syntheses using different HRFTs was possible. In addition, knowledge about the effectiveness of a VAS rendering via listening tests and the corresponding HRTFs used can lead to insights into what components of these spectral cues are the most perceptually relevant by comparing judged HRTFs with the listener's own HRTFs, as will be seen in the following chapter and chapter 9. Repeatability was assessed as a function of error variance across replicates for the different versions of the listening test. Each iteration, and subsequent improvement of the listening test design, aimed to reduce known biases and lead to a reduction in the measured variance, until an optimal version was decided upon. To this end, the current study was successful in developing a procedure that enables low variability in subject responses, given that the subject has some expertise. Furthermore, the described methodology quantifies the level of subject response repeatability so that the experimenter can rely on the validity of the obtained results using measured error variances.

The final listening test, Listening Test 3.2, allowed for a thorough analysis of the variability in perceptual judgements of HRTFs and the effect that subject expertise had on this variability. The results showed that the variability across replicates was significant for all subjects, with a small number of the more experienced subjects showing reliable responses for some attributes. Given the responses from subjective reports and an analysis of the repeatability of the subjects' judgements, the attributes *sense of direction* and *front image quality* were suggested as the most useful. These two attributes were not strongly correlated suggesting that they represented different aspects of the quality of a rendering in VAS. In addition, it was found that for some subjects the mentioned attributes correlated most with the *overall rating* suggesting their importance in the perceived quality of the binaural synthesis. The effect of subject expertise on variability was also analysed and a correlation was found; the more experienced the listener, lower the variance in judgements across trials.

Taken together, through the use of attribute evaluations and assessor selection, this study offers a concise methodology for obtaining consistent evaluations of HRTFs geared towards commercial applications of binaural syntheses. Reliable perceptual judgements of renderings in VAS for different HRTFs is a crucial ingredient to being able to evaluate HRTF individualisation methods given the observed variability in the results for the first versions of the listening test. Furthermore, the methodology used in this study is not limited to the evaluation of binaural syntheses and can be applied to the evaluations of audio quality in general.

8 SALIENT SPECTRAL CUES FOR BINAURAL SYNTHESIS

The purpose of this chapter was to use the LISTEN database of subject Head-Related Transfer Functions (HRTFs) and morphology (see section 6.2.3) to determine, via various forms of analysis, how best to describe inter-subject differences. Various statistical methods and aspects of subject HRTFs or morphology were used to calculate these inter-subject differences. The different methods were validated using a criterion based on the perceptual judgements from a listening test described in the previous chapter. The results were used to shed light on what components of the HRTF were the most perceptually salient with respect to an effective binaural synthesis.

8.1 BACKGROUND

As described in chapter 5, for a realistic illusion of virtual sound sources via a binaural synthesis, a high-fidelity rendering of the auditory scene, known as the Virtual Auditory Space (VAS), is preferable, as described in chapter 5. The use of HRTFs that are perceptually equivalent to the acoustic filters of the listener is one way of enabling a high fidelity rendering. Various techniques exist in the literature for providing listeners with these acoustic filters without the need for the laborious task of measuring a large number of HRTFs, as detailed in section 5.2.2. These studies support the notion that HRTF selection from a database might be a viable customisation technique. The lack of a conclusive perceptual validation for such a technique was one of the main motivations for the current study.

Another major motivation for the work presented in this chapter was the prospect of addressing one of the most challenging questions related to how humans perceive auditory objects in their environment: what cues embedded in the HRTF are the most perceptually significant? Studies relating to the individualisation or personalisation of HRTFs, such as those detailed in section 5.2.2, provide insights into what components of the HRTF are most perceptually salient by drawing on the effectiveness of customisation methods in terms of localisation accuracy or listening test results, and the relation to the particular aspect of the HRTF used in the analysis.

Taken together, previous studies have explored a variety of different methods for describing what is considered significant in the HRTF for humans to effectively interpret sound sources in their environment, yet there are few studies that use perceptual validation to support any findings. The current study aimed to assess inter-subject differences using a large set of subjects and their corresponding HRTF and morphology data, along with a comprehensive perceptual validation. A variety of the most common descriptive techniques were compared, in terms of their ability to explain the subjects' perceptual judgements from a listening test (Listening Test 1 from the previous chapter; see section 7.3).

8.2 METHOD

8.2.1 *Database analysis*

This study is centered on two parts of a general method:

1. an analysis of a database of HRTFs and morphological measurements, which established a number of different techniques for measuring inter-subject differences, and

2. a validation of these techniques using perceptual judgements of the mentioned HRTFs from a listening test.

The validation was used as a vehicle to help discern what aspects of the HRTF were the most perceptually relevant; this involved finding how much one can reduce the spectral information in an HRTF, and testing the most salient spectral components and frequency ranges.

The HRTF and morphological data used in the study were manipulated in six different ways in order to establish the most effective measure of inter-subject differences. The different methods used were chosen so that they would represent some of the most common and effective analysis techniques of HRTFs found in the literature. This involved reducing the data according to different criteria in an effort to distill the most relevant spectral or morphological information. The first two methods (section 8.2.3.1 and 8.2.3.2) used a Principal Component Analysis (PCA) to remove redundant information within the HRTFs. The third method (section 8.2.3.3) focused on the optimal alignment of spectral features in the HRTFs between subjects. The fourth method (section 8.2.3.4) focused on the most frequently cited spectral feature: the frequency notch. The fifth method (section 8.2.3.5) uses peaks, as opposed to notches, in the HRTF data. The final method (section 8.2.3.6) used a PCA of the subject morphology data.

The results of the analysis methods were used to determine subject similarity, producing different metrics to describe the differences between each pair of subjects. These metrics were used to represent the subjects in a high dimensional space, in which the Euclidean distance between the subjects represented the inter-subject differences in terms of their HRTFs. The distances between the subjects were then assessed according to each subject's perceptual judgements from the listening test previously described. This enabled a common methodology for validating the six different analysis methods and their varying techniques for distilling the relevant information in subject HRTFs. By comparing the effectiveness of the distances between subjects to describe the listening test results, some insights into which components of the HRTF were the most perceptually salient could be developed. This was made possible given that each analysis method focused on different aspects of the HRTF, most often drawing from the work of studies that have showed promising results in the past. The study performed in the current chapter is unique in the sense that it presents a number of the most common HRTF descriptive techniques on a single dataset and using the same validation criteria, thereby enabling meaningful comparisons.

8.2.2 *HRTF and morphology database*

The database of HRTF recordings and corresponding listening test results have been used and described in studies from previous chapters (see section 6.2.3 and 7.3 respectively). The current study combines the results from the first listening test in chapter 7 (Listening Test 1, which was not created, nor performed by the author) and an analysis of HRTFs of the very same subjects that participated. Despite the fact that judgements from Listening Test 1 were categorised (i.e. judgements of either *excellent*, *fair*, or *bad*), which restricted the scaling that could be applied to the responses for further analysis of how responses compared across subjects, the results were appropriate for an analysis paired with HRTF data; statistical methods were used to draw on the large number of subject pairwise comparisons in order to produce robust results.

Morphology measurements, included in the LISTEN database, were also used in this study. For each subject 22 morphological parameters were available, corresponding to a subset of measurements acquired from a similar online public database (the CIPIC database; see Algazi et al., 2001b). The parameters x_{13}, x_{14}, x_{15}, d_8 and θ_2 were not available in the current LISTEN database. The full set of CIPIC morphological parameters are shown in figure 36. These parameters, whilst chosen to be descriptive of the human body and ear by the CIPIC database creators, should be recognised as a limited representation of subject morphology and not describing all its form.

The procedure for recording most of the morphological parameters involved the use of a graphical user interface and software to manually measure distances using photos of the subjects. Despite repeating the described procedure to ensure its accuracy, it can be assumed that the method has an associated degree of error for the morphological database. Some parameters such as the *cavum concha depth*, *pinna rotation angle*, and *head circumference*, were measured by hand at the time of the HRTF recordings. Distance parameters were measured in meters and angular parameters in degrees. There was a small number of missing measurements for some of the subjects. This occurred for eight subjects with never more than five missing values for a particular parameter, for a total of 16 missing values. For one subject a measured parameter was deemed an outlier as it is was an isolated value and more than three standard deviations above the mean. This value was removed from the database and counted as a missing value. No morphological data was available for one of the 46 subjects, yet HRTF measurements and listening test responses were available. Methods used in the analyses to deal with missing values are described in the appropriate section. In addition, listening test results were missing for one subject.

x_1 head width
x_2 head height
x_3 head depth
x_4 pinna offset down
x_5 pinna offset back
x_6 neck width
x_7 neck height
x_8 neck depth
x_9 torso top width
x_{10} torso top height
x_{11} torso top depth
x_{12} shoulder width
x_{13} head offset forward
x_{14} height
x_{15} seated height
x_{16} head circumference
x_{17} shoulder circumference

d_1 cavum concha height
d_2 cymba concha height
d_3 cavum concha width
d_4 fossa height
d_5 pinna height
d_6 pinna width
d_7 intertragal incisure width
d_8 cavum concha depth
θ_1 pinna rotation angle
θ_2 pinna are angle

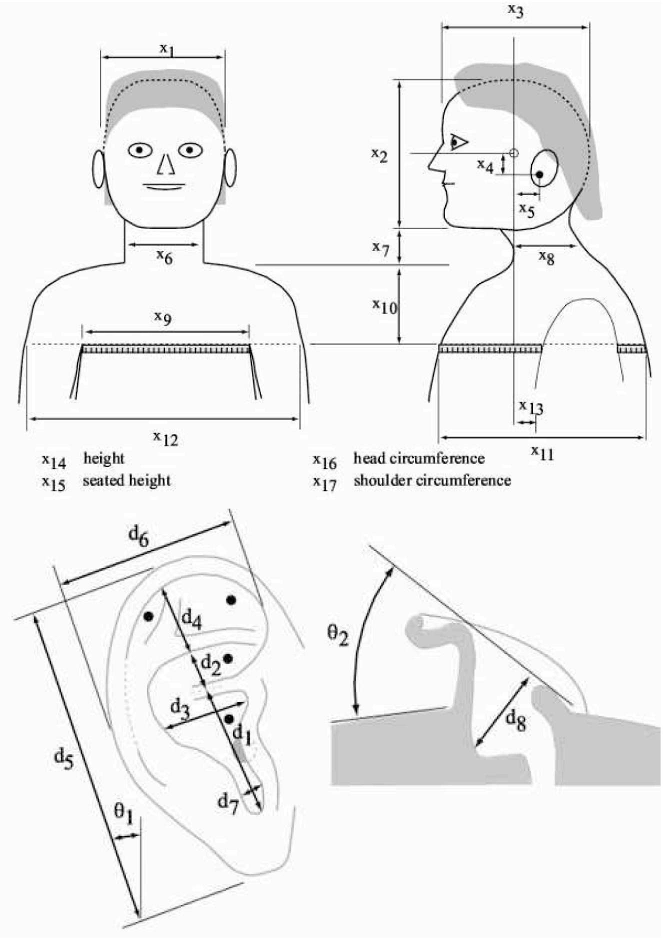


Figure 36: Morphological parameters taken from the CIPIC database of which 22 were used in this study. Figure taken from Algazi et al. (2001b).

8.2.3 *Subjects represented in multidimensional spaces*

As previously mentioned, the overall analytic approach was to map each subject into a high dimensional space in which the Euclidean distance between subjects represented HRTF similarity; a pair of subjects that were relatively close together resembled each other and subjects that were further apart were dissimilar. A number of different techniques were explored for creating these Multidimensional Spaces (MSs), which are described in detail in the following sections. In brief, the first five methods used the Directional Transfer Function (DTF) data, aiming to reduce the quantity of information extracted from the HRTFs in order to create each MS. The final technique used only the morphological data to create a MS.

The effectiveness of each MS was evaluated by comparing how the distances between subjects correlated with the subjects' perceptual judgements of the different sets of HRTFs from Listening Test 1 (section 7.3). This was done by grouping distances between pairs of subjects in a particular MS as a function of the judgement (i.e. *excellent*, *fair*, or *bad*) given by one subject for the HRTFs of the other subjects. The distance values for these three judgement categories were then analysed across all subjects (detailed in section 8.2.4). The following sections describe the methods used to create each of the six MSs.

8.2.3.1 *DTF magnitude*

As previously described, the first analysis method used a PCA to compress the subject DTF magnitude data and create a MS. This type of data compression is particularly well suited to HRTFs since there is a large amount of redundant information across measurements. Despite the existence of large perceptual differences in HRTFs for different listeners, there does exist a large degree of similarity in their general form, which is exploited by a PCA. PCA is the technique that is often used in the literature to reduce HRTF data into meaningful representations (Middlebrooks and Green, 1992; Kistler and Wightman, 1992; Chen et al., 1995; Bai and Ou, 2005; Sodnik et al., 2006; Hwang and Park, 2007; Shin and Park, 2008; Hwang and Park, 2008; Wang et al., 2008; Martens, 1987), and was thus used in this study for its suitability and compatibility with past research. As a precursor to the PCA, the DTF magnitudes were logarithmically transformed, or converted to the decibel scale, in order to simulate loudness transduction in the auditory pathway (see Martens, 1987, for the earliest application of PCA using HRTFs).

As part of a post-processing performed for the current study, the DTFs from the online database were normalised in the time domain using the root-mean-square of the signal and spectrally smoothed using a critical band filter in an effort to mimic the effects of the cochlear filter on neural encoding (see Carlile and Pralong, 1994). The smoothing of detail in subject DTFs eliminated variance in the magnitude data that was perceptually irrelevant, which may be considered as noise in the signal, thereby improving the analysis. It should be noted however that the smoothed DTFs were still not equivalent to having been passed through the cochlear filter due to the linear sampling in

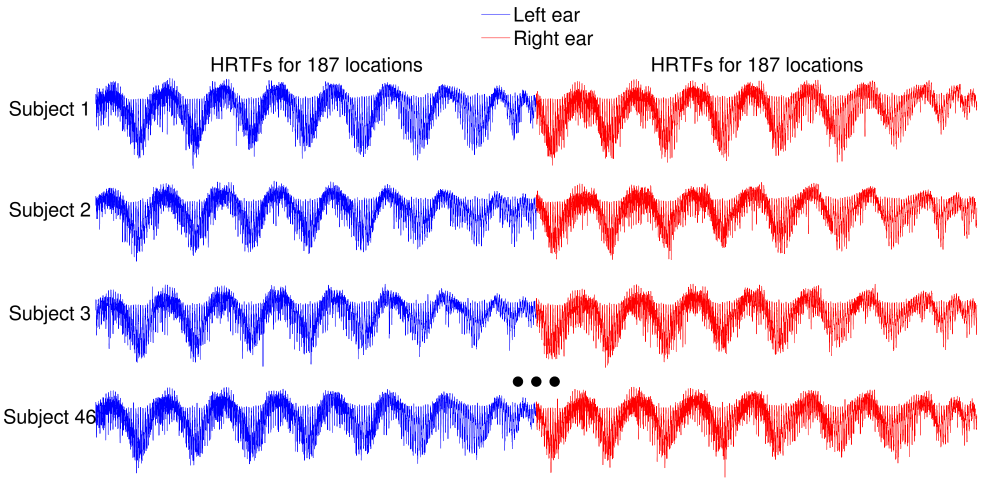


Figure 37: Visualisation of the input matrix for the PCA.

frequency, which is the focus of the following section. The PCA was then performed on the spectral magnitude of the DTFs by organising the data into a matrix where each row represented a subject and each column a magnitude value in decibels for a specific frequency. More specifically, for each subject all DTFs for all positions in space were concatenated; the left ear was concatenated first, then the right, so that there was a single vector of magnitude values, as shown in figure 37. This concatenation technique was chosen in order to represent each subject as a row in the input matrix, which is essential for the analysis. This procedure has been used previously in similar studies by Jin et al. (2000) and Zeng et al. (2010).

The PCA, by way of a singular value decomposition, exploited redundancies in the DTF matrix and generated a new set of orthogonal axes that minimised deviations (using a sum squared error). This new coordinate system is termed the Principal Components (PCs) of the data (see Middlebrooks and Green, 1992, for an example). The PCs are organised in a hierarchical manner; the first PC describing the most variance in the DTF data with subsequent PCs describing less and less variance. For each of the PCs a set of corresponding weights is produced for each subject so that each vector of concatenated DTF magnitudes could be reconstructed if all the PCs and corresponding weights were used. This reconstruction is possible due to the fact that a PCA is a linear transformation. Previous studies have shown that for a perceptually effective reconstruction of DTFs, only a small number of PCs have been found to be necessary (Leung and Carlile, 2009; Kistler and Wightman, 1992), which leads to a reduction in dimensionality of the original data. The calculated weights from the PCA represent the original DTF data projected onto the new set of axes (the PCs). In the current study, the approach was then to use the set of weights for each subject as coordinates in a MS to describe inter-subject differences, as performed first by Usher and Martens (2007).

The weights were ordered according to the hierarchy of PCs for each subject. This meant that if for example the subjects were mapped into a one-dimensional space they would be using the weights of the first PC for each subject as coordinates; the first PC described the most variance in the HRTF data and the weights represented deviations from the redundant forms in the DTFs. In this way the weights themselves are a valuable low dimensional metric for describing how subjects differ with respect to their DTFs. The intention of this study is to use these weights in an attempt to explain or describe the listening test results. As the number of weights, and thus dimensions in the MS increases, more and more variance in the data would be described. The amount of variance described by each additional PC and corresponding weights decreases as more are included.

8.2.3.2 *Fractional octave-band filtered DTF*

The human auditory system has a higher sensitivity for frequencies in the upper regions of the audible frequency range; it samples frequencies on a scale that is almost logarithmic. For this reason, the DTFs of each subject were processed with a fractional octave-band filter in order to obtain a more perceptually relevant representation of the DTFs in terms of the cochlea's frequency resolution. It was hypothesised that a MS that represented subject DTFs as they are interpreted by the auditory system after the cochlear filter might better describe the perceptual judgements from the listening test. The method employed by Middlebrooks (1999b) was implemented in this study in which 85 bandpass filters were used at equal intervals on an octave scale from 3 to 16 kHz. The order of the filters was 8 with a stop band attenuation of approximately 75 dB, in line with the standard specifications (ANSI S1.11-2004, American National Standards Inst., 2004). Based on the standard, the center frequency f_c and bandwidth Δf_c of an perceptually motivated bandpass filter is defined by:

$$f_c^{(N)}(j) = 1000(2^{j/N})\ \text{Hz} \ \text{ and } \ \Delta f_c^{(N)}(j) = f_c^{(N)}(j)\frac{2^{1/N}-1}{2^{1/2N}}\ \text{Hz} \tag{4}$$

where $f_c^{(N)}(j)$ is the center frequency of the j^{th} bandpass filter expressed in Hz, j is an integer when $j = 0$, $f_c^{(N)}(0) = 1$ kHz, which is the reference frequency for the audio range, and $(1/N)$ is the bandwidth designator where N equals 35 for the current study, i.e. 35 bands per octave, which equates to 85 bandpass filters.

The filtered DTFs were concatenated and arranged into a matrix, in the same manner as the unfiltered DTFs as mentioned in the previous section, and a PCA was performed.

8.2.3.3 *Frequency scale factor*

In the study by Middlebrooks (1999b) it was suggested that DTFs "differ systematically among subjects in regard to the position of spectral features along a frequency axis". Subsequent findings showed that virtual localisation for a particular subject can be improved by aligning the peaks and notches present in non-individualised HRTFs with those of the subject's own

HRTFs (Middlebrooks, 1999a). Given the success of this so-called frequency scaling technique, the procedure was replicated in this study and used as a measure of inter-subject differences. It was hypothesised that this measure would correlate well with the listening test results as it embodied a mixture of some of the most perceptually salient spectral cues in its calculation.

As previously detailed in section 5.2.2.2, the fractional octave-band filtered DTF data, mentioned in the previous section, was used to calculate an optimal Frequency Scale Factor (FSF) between every pair of subjects that best aligned their spectral features. The process, described in detail by Middlebrooks (1999b), involved calculating the disparity between two DTFs from a pair of subjects by scaling them both upward and downward in the frequency domain in steps of 0.0286 of an octave. This scaling corresponded to the space between the center frequencies of the fractional octave-band filters. Each shift of 0.0286 of an octave equates to a linear factor of 1.02. A frequency range between 3.7 to 12.9 kHz, determined by Middlebrooks to be the most significant region of the spectrum, was used for comparing the DTFs, with each shift passing frequency magnitude values in at one end of this range and out the other (see figure 39 for an example of a set of scaled DTFs).

For each FSF, the 64 filter-bank components that were in the determined frequency range were subtracted from one another and the variance of the resulting difference spectrum was calculated. This metric value, referred to as the Inter-Subject Spectral Difference (ISSD), was computed for all possible FSFs. A vector of ISSD values was calculated for every spatial position, or every pair of subject DTFs for each position. A mean ISSD vector was then calculated across all ISSD vectors, which was equivalent to calculating the mean ISSD for each FSF across all positions. The scale factor corresponding to the minimum of the mean ISSD was taken as the global optimal FSF for each pair of subjects. Figure 38 shows the ISSD vectors for each position (gray lines) as a function of FSF, along with the mean ISSD values (represented by the black dots), for a pair of subjects used in the study. Using this frequency scale factor of 1.37 for the same pair of subjects, figure 39 displays a selection of subject fractional octave-band filtered DTFs scaled in the frequency domain to better match the other subject's DTFs. DTFs for positions along the midline, i.e. at 0° azimuth, are presented for varying elevations. The elevation angle for each pair of DTFs is displayed on the plot expressed for ease of presentation from 45° to 315°, corresponding to elevation angles that begin at 0° directly below the listener to in front at 90° and behind at 270°. The plot on the left represents pairs of unscaled DTFs and the plot on the right scaled DTFs. The vertical grid lines on the plots correspond to the range of frequencies (3.7 to 12.9 kHz) taken into account when calculating the ISSD.

For the scaled DTFs, subject 1025 DTFs were scaled by a factor of 1.17 (8 steps) to the left and subject 1050 DTFs were scaled by 1.17 (8 steps) to the right, for a combined optimal scale factor of 1.37. It can be seen that the alignment of spectral features such as peaks and notches is greatly improved for some of the scaled DTFs in the right plot of figure (e.g. location with elevation 225°), with some positions still displaying differences between scaled DTFs

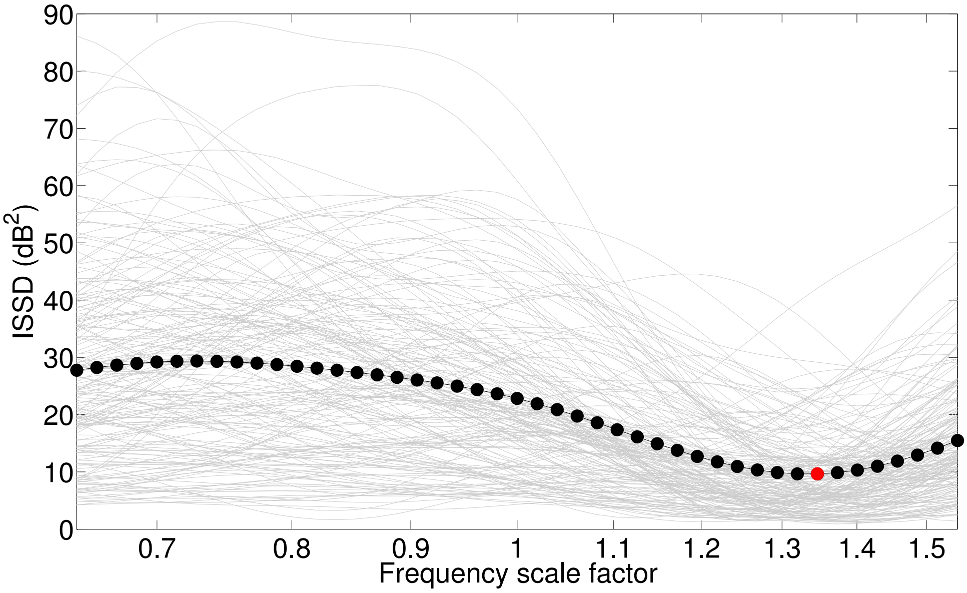


Figure 38: Mean inter-subject spectral difference across all positions, represented as black dots, as a function of frequency scale factor, for one pair subjects in the LISTEN database (subject 1025 and 1050). The red point displays the minimum mean frequency scale factor for this pair of subjects. Gray lines represent the calculated inter-subject spectral difference for every position in space (i.e. each pair of measured subject DTFs), from which the mean is calculated.

(e.g. location with elevation 45°). The fact that some positions show less of an improvement is due to the use of a global FSF calculated from the mean of ISSD values across all positions in space as was shown in figure 38. If individual FSFs had been used for each location in space (see figure 39), each shifted pair of DTFs would have their spectral features optimally aligned. The use of a global FSF was chosen in order to represent subjects in a MS, previous studies have shown that the mean ISSD values between scaled DTFs across all positions to decrease relative to unscaled DTFs for a large number of subjects (Middlebrooks, 1999a), and that localisation accuracy could be improved by using a global FSF and non-individualised HRTFs (Middlebrooks, 1999b).

Since the FSFs were used as a measure of the disparity between a pair of subjects, the direction of the scaling (i.e. whether subject A was scaled towards subject B or vice versa) was not considered relevant in this study. As such, all FSFs were considered as upscaling, and any scale factor that was less than 1.0 was taken as its reciprocal so that all FSFs were greater than 1.0. This was particularly relevant given that FSFs were to be used as distances between subjects in a MS, and a scale factor less than 1.0 for a pair of subject DTFs would translate to a closer resemblance when in fact a FSF value of 1.0 indicates the highest degree of similarity (i.e. no scaling needed). FSFs were calculated for both the left and right ear separately with the ear offering the most effective MS, in terms of the listening test results, being chosen for further analysis (see section 8.2.4). The justification for using data from only one ear has been detailed by Middlebrooks (1999b). It was of interest to use

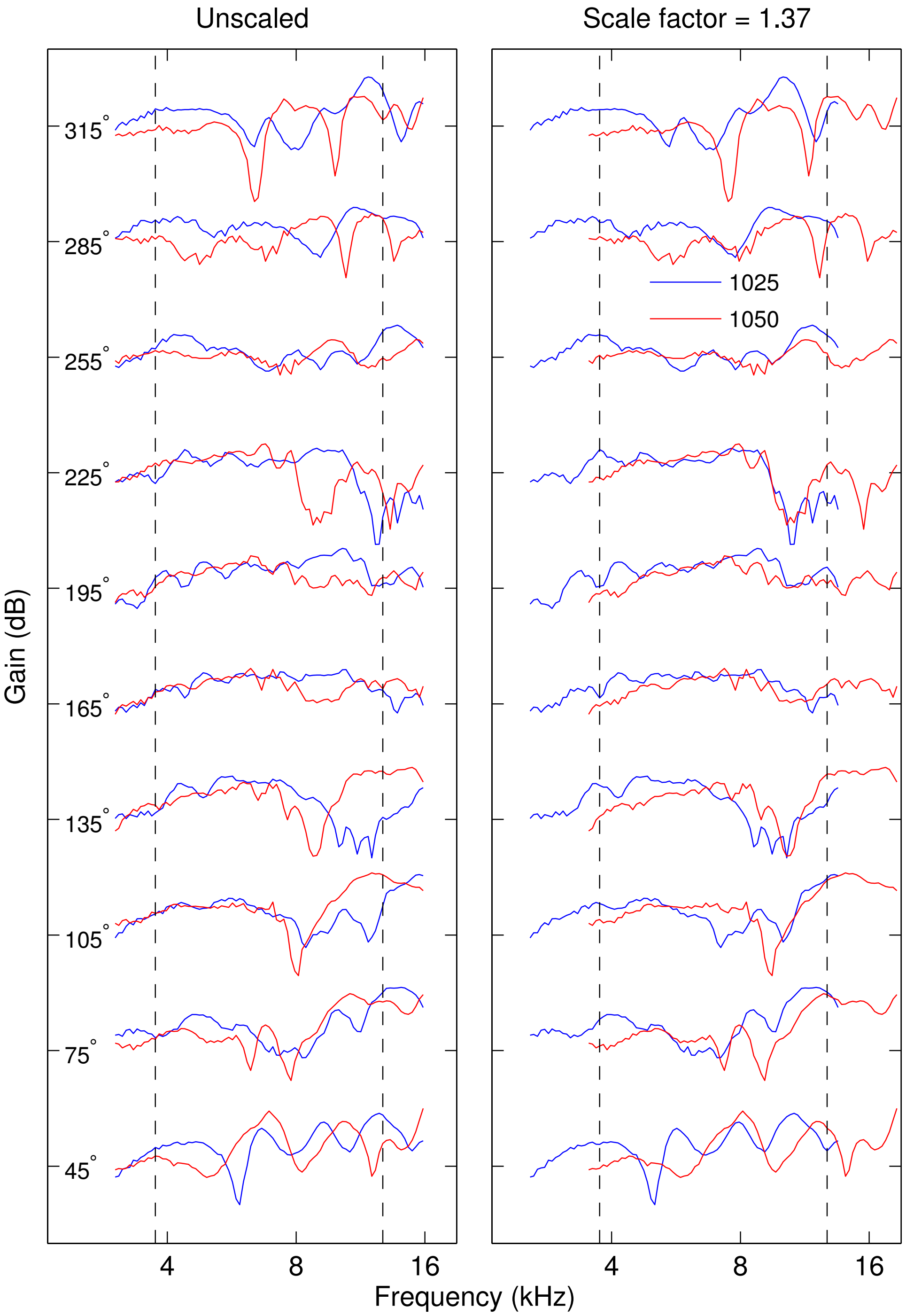


Figure 39: DTFs for directions along the midline at elevations ranging from in front to behind the listener (see text). The fractional octave-band filtered DTFs are presented for a pair of subjects without (unscaled) and with an optimal frequency scaling. For the scaled DTFs, subject 1025 DTFs were scaled by a factor of 1.17 to the left and subject 1050 DTFs were scaled by 1.17 to the right, for a combined optimal scale factor of 1.37. The vertical grid lines correspond to the frequency range used when calculating ISSDs.

the same methodology in order to ensure the compatibility of the results for comparisons with previous works. The correlation coefficient between the base-2 logarithms of the optimal scale factors of the left and right ears was found to be 0.85 in the current study and 0.95 in the Middlebrooks (1999b) study. A notable difference between these two studies is the resolution of the measured HRTFs in space. Middlebrooks (1999b) used 400 measured sound source locations and an interpolation to 393 HRTF positions in space to ensure an even distribution around the listener, whereas the current study used a set of 187 sound source locations, not entirely evenly spaced around the listener, and with no interpolation.

One additional step was incorporated into the final calculation of the FSFs, which is not part of the procedure used by Middlebrooks, due to an observed bias for some calculated FSFs. This step involved varying which subject DTF (from the pair of subject DTFs used to calculate each FSF) was first shifted in the frequency domain (either upward or downward depending on the subject DTF). The use of either subject DTF as the first to be shifted, or scaled by one step of 0.0286 of an octave, was significant as it had an effect on the calculated ISSD and by consequence the optimal FSF. The observed differences in the calculated ISSD values were due to the fact that for every pair of DTFs being used one DTF was being shifted so that magnitude values at high frequencies passed out the specified range (3.7 to 12.9 kHz) whilst the other DTF was being shifted so that magnitude values at low frequencies passed out the specified frequency range (see scaled DTFs in figure 39). The DTF that has high frequency magnitude values passing out will be introducing low frequency magnitude values, which typically result in less added variance and a lower ISSD value (relative to the other DTF) due to the fact that the spectrum of DTFs at low frequencies are rather flat in comparison to high frequencies. Since the ISSD values are relatively lower for the shift of one DTF, a bias is generated for each corresponding scale factor depending on whether this DTF was shifted first or second, which is an arbitrary choice. By taking the mean ISSD value for both scenarios (one DTF scaled first and then the other DTF scaled first), this bias was avoided. Figure 40 shows the calculated ISSD values for one position in space, as in figure 38, for when either subject 1050 or subject 1025 were shifted first. The bias is apparent for every second FSF that has a relatively lower ISSD value than the adjacent FSF for when either of the subjects is shifted first. The mean difference between the corrected and uncorrected ISSD values was typically in the order of approximately 0.2 dB^2, with a maximum difference of approximately 1.5 dB^2.

The FSFs for all pairs of subjects was used as a distance vector for a classical multidimensional scaling which projected the subjects into a MS configuration in which the Euclidean distances between subjects best reflected the scaling factor values. Multidimensional scaling creates a high order space in which an optimal configuration of the subjects is achieved so that the distances between each pair of subjects is as close as possible to the calculated FSF for that pair of subject DTFs. The output of the multidimensional scaling was a selection of approximately 20 dimensions out of a possible 46, determined automatically using a Classical Multidimensional Scaling algorithm

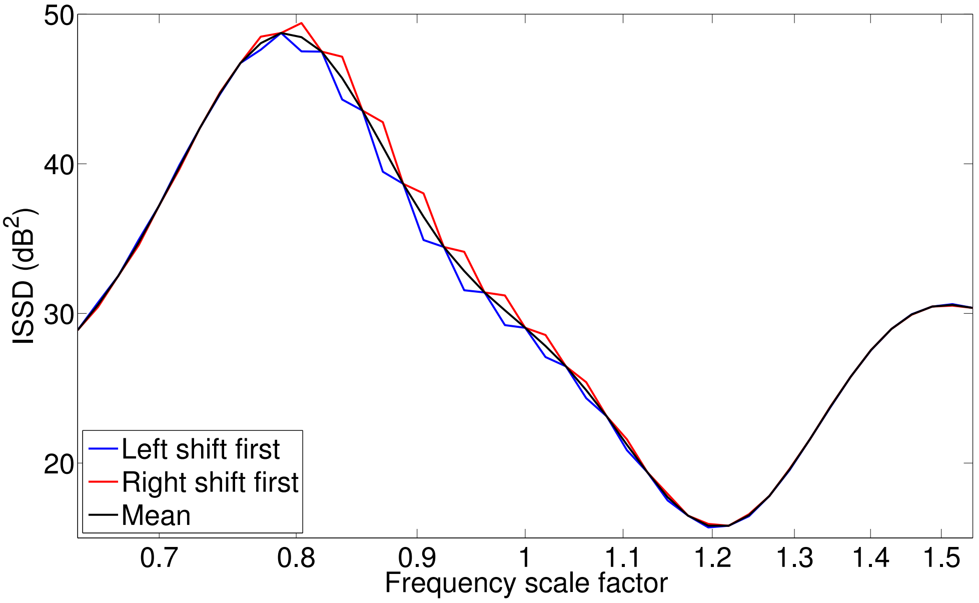


Figure 40: Calculated ISSD values for one position (−45° azimuth and −45° elevation) as a function of FSF for when subject 1050 and 1025 are shifted first (blue and red respectively), along with the mean calculated for each FSF (black). The bias for every second FSF is apparent for when either subject 1050 or 1025 is shifted first.

in an effort to best represent the FSFs between subjects (see Seber (1984) for details).

A separate analysis of the constructed MS was performed in order to compare the distances between pairs of subjects and the FSFs that were calculated for the same pair of subject DTFs. The differences between the value of the Euclidean distances in the MS and calculated FSFs were considered distance errors. These distance errors were then normalised by dividing by the maximum distance between any two subjects in the MS. The mean distance error was calculated to be approximately 0.14, or 14% of the largest distance between subjects, with a standard deviation of approximately 0.07. This signifies the level of accuracy for the forced configuration of subjects in the MS with respect to the calculated FSFs between subjects.

8.2.3.4 *Notch frequencies*

Work by Middlebrooks (1999b,a) demonstrated that there are frequency dependent components of the HRTF that are perceptually important, particularly for localisation. One component that has been discussed in a number of studies as being significant are the sharp notches in the spectrum, known as spectral notches (Greff and Katz, 2007; Iida et al., 2007; Rodriduez and Ramirez, 2005). These notches have been shown to be important cues for vertical localisation, i.e. indicating whether a sound source is located above or below the listener. Spectral notches can be thought of as a more overt spectral feature as compared to FSFs that are more covert in the sense that their significance emerges only after analysing all HRTFs for all positions in space (Carlile et al., 2005).

A detailed explanation of the method used to extract the frequencies from the DTFs for each position in space is described in a study by Raykar et al. (2005). For each DTF, a number of the most significant notches were identified for each position within the frequency range from 3 to 16 kHz, as determined by the algorithm used by Raykar et al. (2005). For each subject the frequencies of these notches were computed for the left and right ear and stored as a matrix where each row represented a position and each column a notch in decreasing order of significance. The significance of each notch was judged automatically using Raykar's algorithm and was related to the depth of the notch after a number of signal processing operations. A two-dimensional correlation coefficient, similar to the spherical correlation coefficient used in chapter 6 (see section 6.2.9), was used to calculate the disparity in notch frequencies between each pair of subjects. This correlation coefficient was analogous to the calculated FSFs in section 8.2.3.3 in the sense that it was a global measure reflecting the similarity, across all measured DTFs, between a pair of subjects. Correlation coefficients were computed for both the left and right ear data separately. The correlation coefficients were used, as in the previous method, as a distance matrix for a classical multidimensional scaling producing two MSs (one for the left ear data and one for right ear data) that reflected subject dissimilarity. As with the previous method, the most effective MS of the two was retained using a criterion detailed in section 8.2.4. In this way, by maintaining the same procedure, comparisons between the results for methods using FSFs and frequency notches could be made.

The calculated correlation coefficients were low, with a maximum of approximately 0.25 between any pair of subjects for both the left and right ear calculations. The similarity between the calculated correlation coefficients for the left and right ears was as high as for the FSFs; the correlation coefficient between the left and right ear correlation coefficients was only 0.08 compared to 0.85 for the FSFs. This result may suggest that the method using subject notch frequencies was not as descriptive as the FSFs, given that one would expect a higher degree of correlation due to the similarity between HRTFs for a subject's left and right ear (for locations that are symmetrical about the midline).

8.2.3.5 *Covert peak*

As described in the previous introductory chapter (see section 4.1), early experiments performed by Blauert (1969) with respect to narrow-band stimuli localisation in the free-field have shown that performance is attributed to centre frequency of the stimulus rather than its location. The author described that each frequency of the so-called directivity bands were associated with a direction in space. These results were replicated in studies by (Butler and Helwig, 1983) in which inter-subject differences were observed, making this feature of human audition a possible candidate for this study. Butler (1987) and Butler et al. (1990) introduced the notion of covert peaks in HRTFs in order to further explore the concept of directivity bands for broadband stimuli.

The name was chosen due to the fact that the maximum magnitude, or peak, of an HRTF for a certain frequency can only be ascertained with knowledge of all the recorded HRTF over the entire space.

Thus, for this analysis method each subject's DTFs were ordered according to frequency, and the position in space for which there was a maximum magnitude across all frequencies was used as the location of the covert peak. In order to effectively sample frequencies on a perceptually relevant scale, the magnitude values from the fractional octave-band filtered DTFs were used (see section 8.2.3.2). As with previous methods, only frequencies between 3 and 16 kHz were used. A position in space was calculated for each frequency producing two vectors of azimuth and elevation angles for each subject in the database. This process was carried out for the left and right ear separately. As for the previous method, a two-dimensional correlation coefficient was used to calculate the disparity in covert peak positions between each pair of subjects. The correlation coefficients were then used, as for the method using FSFs and the previous section, as a distance vector, representing pairwise distances between subjects in space, for a classical multidimensional scaling. The process produced a coordinate for each subject in the two MSs (one for the left and right ear). As with the previous method, the most effective MS of the two was retained (see section 8.2.4 for details). In this way, by maintaining the same procedure, comparisons between the results for the previous three methods could be performed.

8.2.3.6 *Morphological parameters*

The final analysis method, or MS generation method, used subject morphological data in an attempt to establish a direct relationship between it and subjects' perceptual judgements. This technique bypassed acoustic measurements entirely, instead focusing on the physical shapes of the subjects' morphology that influence the HRTF most (the focus of the next chapter).

The morphological data was organised into a matrix where each row represented a subject and each column one of the 22 morphological parameters previously detailed in section 8.2.1. Despite the importance of asymmetries between the left and right ear in perceiving sound sources (Searle et al., 1975; Brookes and Treble, 2005), measurements for only one ear's pinna morphological parameters were retained in order to avoid data redundancy in the PCA analysis. The choice of left or right ear parameters was made based on the most effective of two MSs (left or right pinna parameters combined with body parameters). The choice to use parameters of only one ear, as opposed to calculating a mean, was justified given the high correlation between left and right ear measurements in the database used (correlation coefficient of 0.98). The effectiveness of a MS was judged according to the correlation between the results from the listening test and the distances between pairs of subjects in the chosen MS as will be described in the following section. A PCA was performed on the morphology matrix and the weights were used to generate a set of coordinates for each subject in a MS, as was performed for the first two methods detailed in section 8.2.3.1 and section 8.2.3.2.

8.2.4 *Validation of multidimensional spaces*

The six different methods described above used either subject HRTF or morphology data to create six different MSs. The extent to which these spaces effectively described subject dissimilarity was assessed using the subjects' own perceptual judgements of the HRTFs via a listening test (Listening Test 1 from the previous chapter). By using qualitative responses from the listening test, rather than a pure localisation task, the hypothesis of this study was that the perceptual judgements involved would better assess the overall effectiveness of the VAS rendering. Perceptual judgements were seen as more relevant to commercial applications of HRTF customisation, given that they have less of a focus on pure localisation accuracy, which is only one component of generating a compelling VAS for the listener (Martens, 2003).

The HRTF recordings and database, methods used in the binaural synthesis of the stimuli, and procedure of the listening test, have been described in detail previously in section 6.2.3 and 6.2.4. For each of the 46 HRTFs from the database, subjects had to make a judgement, selecting either *excellent*, *fair*, or *bad*, based on whether the virtual sound source accurately followed the predetermined trajectory.

8.3 RESULTS

8.3.1 *Validation of principal components*

As a precursor to the analysis of the effectiveness of the different MSs (for the first two methods using a PCA of subject DTFs), a validation of the calculated PCs was performed. Figure 41 shows the amount of variance in the original DTF data explained by the increasing number of PCs (see section 8.2.3.1 for details of the PCA technique used). The results demonstrate a gradual increase, from about 15% for the first PC, in the amount of variability accounted for by the accumulation of PCs used to describe the data. These results are in line with findings from similar studies, in which subject DTFs were concatenated for a PCA of a dataset of 36 different subjects (Jin et al., 2000). The PCA in the study by Jin et al. (2000) however was performed on a compressed representation of subject DTFs; instead of concatenating all left and right ear DTFs for the input matrix as was the case in the current study. For this compression method, a pre-processing PCA was performed by first concatenating left and right ear DTFs for the 393 positions recorded for each subject. The resulting 800 point frequency magnitude vectors were then combined across all subjects into one data matrix containing a total of 14,148 vectors (i.e. 393 vectors for each of the 36 subjects). The weights for each subject, from the PCA of the entire dataset, were then concatenated to create 36 compressed representations of the subject DTFs. These 36 vectors will be correlated given that the PCA was performed across all subjects. A PCA was then performed, as in the current study, on the concatenated vectors so that a set of weights is calculated for each subject.

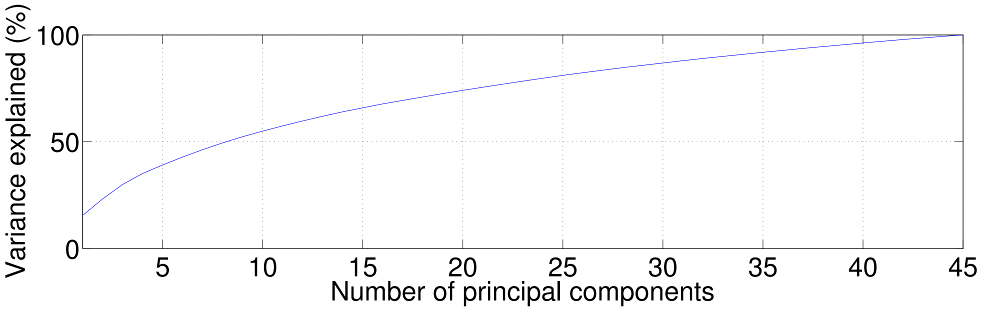


Figure 41: Cumulative percentage of variance explained as a function of PC for a PCA of all concatenated subject DTFs.

This process equates to performing a PCA twice; once on the HRTFs of all subjects, producing a set of weights, and then again on a matrix of weights in which each row was a concatenation of each subject's PC weights. This technique reduces the amount of variation in the input matrix as well as the size of the input matrix, and is reflected in the fact that only some nine PCs were required to describe 90% of the variance in the input matrix as opposed to 35 PCs in the current study (see figure 41). Whilst this type of compression is useful for reducing the computational load of the analysis of HRTFs, it does discard a significant amount of the inter-subject differences for the generation of a MS.

Studies that have not concatenated subject DTFs for PCA, having each row in the input matrix correspond to only one position in space, have shown that the percentage of variance explained in the first PC can be as large as 80%, and the need for between only five and eight PCs to describe 90% of the variability in the input matrix (Zeng et al., 2010; Kistler and Wightman, 1992). This number of PCs appears to correlate well with the number needed for a perceptually effective reconstruction of subject DTFs. Leung and Carlile (2009) have shown that localisation accuracy using HRTFs reconstructed from only 10 PCs is comparable to that using the subjects' own uncompressed HRTFs in VAS. Kistler and Wightman (1992) showed that accurate localisation was achieved with as little as five PCs. These results highlight the large degree of redundancy in subject DTFs and the appropriateness of PCA as a technique for compressing the data and drawing out the most significant aspects of variability between subjects in the form of the PCs. The differences in the percentage of variance explained in each PC between the mentioned studies and the current study can be attributed to the fact that the former used datasets with a largely reduced number of columns (due to the fact that DTFs were not concatenated), represented by frequency-magnitude values of the spectrum, which lead to a decrease in the amount of variance between rows.

In an effort to evaluate the effectiveness of the calculated PCs to describe the DTFs of each subject in the database, an analysis of reconstructed DTFs using an increasing number of PCs was performed in the current study. For this analysis, each subject was removed from the subject dataset for the PCA, which equated to the removal of the corresponding row of subject DTFs from the input matrix. A PCA was performed on this reduced input matrix and weights were calculated for each subject based on a projection of that subject's removed vector of concatenated DTFs onto the calculated PCs. The subject DTFs were then reconstructed using an increasing number of PCs and corresponding weights. As more weights were used, the reconstruction of the subject DTFs approached the original subject DTF. The reconstruction was never a perfect match, even after all the weights and corresponding PCs were used, as the original subject DTFs were not part of the dataset used in the PCA.

Two metrics were used to measure the effectiveness of the reconstructed DTFs: the cosine of the angle between the reconstructed and original subject concatenated DTF magnitude vectors, and the percentage mean square error. The first metric is a common measure of the correspondence of the two vectors, equivalent to the dot product of two vectors normalised by the Euclidean distance between the points represented by the vectors. As the angle between the two vectors, in a space with as many dimensions as there are weights, becomes smaller, the ratio approaches a value of one. The second metric is the mean square error between the original and reconstructed HRTF divided by the maximum range within the data. This metric has been used in previous studies comparing the compression rates for a PCA of different representations of HRTFs on logarithmic and linear scales (Leung and Carlile, 2009). Figure 42 shows the two metrics as a function of the number of PCs used to reconstruct the subjects' DTFs. Each line on the plot represents one of the subjects in the database. The results show that there are only three subjects that stood out as having a slightly poorer reconstruction of their DTFs for the metric using percentage mean square error (right plot in figure 42). This shows that the calculated PCs were generally descriptive of the population of HRTFs as each subject was reconstructed from a PCA of concatenated DTFs that did not include their own HRTFs. Neither of the metrics show a perfect reconstruction of the DTFs as the subject's own HRTFs were never included in the PCA. Leung and Carlile (2009) have shown that the percentage mean square error for reconstructed subject HRTFs can fall below 10%, as opposed to approximately 30% in the current study, with as little as 10 PCs, yet in their study HRTFs were not concatenated and the subject's own HRTFs were included in the database being analysed resulting in a much more efficient reconstruction.

8.3.2 *Statistical analysis*

The effectiveness of each MS to predict the results from the listening test was analysed by comparing distances between subjects. For each subject, the Euclidean distances to the 44 other subjects in the MS were computed and these distance values were divided into three judgement groups based on whether the corresponding HRTFs were judged as either *excellent*, *fair*, or *bad* by the subject. This process was repeated across all subjects, and all distance values were grouped according to judgement. For each MS, the distribution of the distance values for the three judgement groups and all subjects was as-

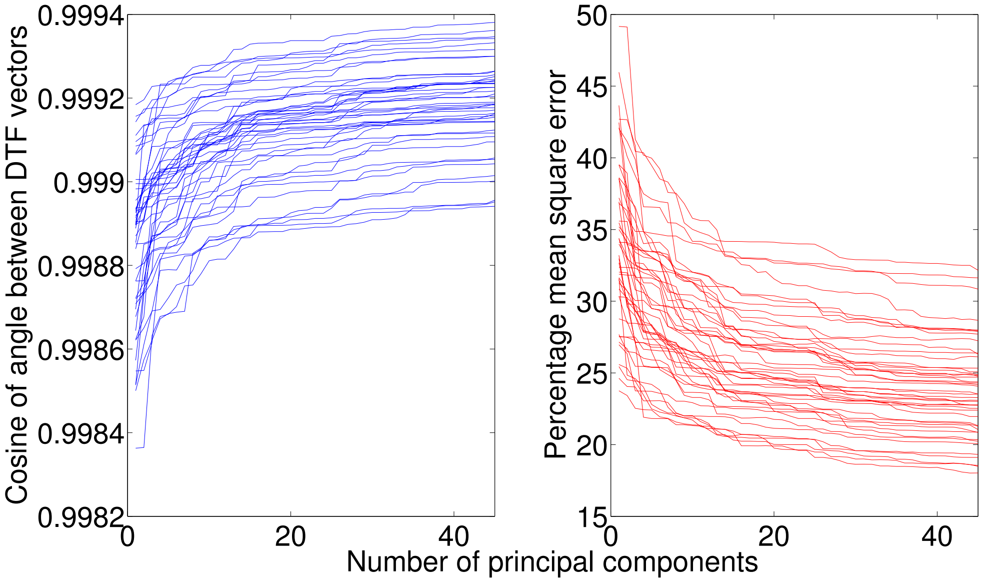


Figure 42: Two metrics showing the reconstruction of each subject's DTFs using PCs calculated from an input matrix of all subject DTFs except those of the subject in question. The left plot shows the cosine of the reconstructed and original concatenated subject DTFs, and the right plot shows the percentage mean squared error.

sessed and a one-way ANalysis Of VAriance (ANOVA) for comparing means was used. The F-ratio value, calculated as the amount of variance between the judgement groups divided by the within group variance whilst taking into account the degrees of freedom, was used as a metric to judge the effectiveness of a MS. The F-ratio has previously been used in studies assessing differences in the distribution of data; for example, testing the variability in loudspeaker quality judgements (Bech, 1992).

Ideally, for an effective MS, the distance values for HRTFs judged as *excellent* would be distributed towards lower values (i.e. being closer in the MS and more similar), HRTFs judged as *fair* distributed over the midrange of the distance values, and HRTFs judged as *bad* distributed towards higher values. The F-ratio value of the ANOVA described the extent to which the distance values separated into the described ideal distributions and was thus used as a metric to evaluate each MS. The higher the F-ratio, the more effective the MS was in predicting subject perceptual evaluations of the HRTFs. For each of the six analysis methods the F-ratio value was determined and used to rank their effectiveness.

The F-ratio is used to tell whether the judgement group had an effect on the distances between pairs of HRTFs in the MSs. It does not need to be necessarily related to a confidence level for statistical significance and corresponding p-value; if the F-ratio increases then this indicates that the effect of judgement group is stronger. The F-ratio is robust to one judgement group (say that of *excellent* judgements) having more of an effect than another. The F-ratio will reduce if one of the three groups is less effective; the between group variance can change while the within variance remains the same. One drawback to using the F-ratio value as a metric of effectiveness is that measured differences among the judgement groups have no sign; the *bad* and the *excellent* judgement group could both have distance values that are small (it is desirable that only the *excellent* group have small distance values), and this would give a high F-ratio. An inspection of the distribution of distance values in each judgement category can ensure that the F-ratio is indeed measuring the effectiveness of the MSs.

Figure 43 shows the distributions for the different MSs for each of the six methods. For each of the judgement groups: *excellent*, *fair*, and *bad*, corresponding to the labels E, F, and B in the figure, a mirrored histogram is shown with the width of each bar corresponding to the number of data points (or distance values between subjects) in each bin. The normalised distance values between subjects are represented on the vertical axis. Histograms for HRTFs judged as *excellent* are narrower as fewer HRTFs were judged in the category as compared to the other categories. The total number of data points is given by n; figure 43(f) has less data points as there were subjects in the database for which no morphological parameters were measured, and thus is not directly comparable with the other methods.

The maximum F-ratio along with the corresponding p-value, for a statistical test of how likely the null hypothesis that the three distributions were drawn from the same population, is shown for each MS. Distributions for a MS were considered statistically significant if the p-value was smaller than 0.05 and the null hypothesis rejected. Results in figure 43 are ordered from (a) to (f) in decreasing statistical significance (decreasing F-ratio), which ranks the different methods from most effective to least effective according to this metric. The most effective being the analysis using the frequency scale factors and the least effective method being the analysis using morphological parameters. As the effectiveness of the MSs decreases, the distributions become less distinct, and the median distances for each group no longer increase with perceptual judgement. The results show that the alignment of peaks and notches in the subject HRTFs using FSFs is the most effective analysis method for describing inter-subject differences given the validation using listening test results. The alignment of frequency notches over peaks most probably dominated the calculation of the FSFs given that the MS created using frequency notches was more effective than that using covert peaks, yet the two methods are not directly comparable as they used overt and covert features of the HRTF respectively. The fractional octave-band filtered DTFs were more effective than the unfiltered DTFs most probably due to the fact that they were a more perceptually relevant representation of subject HRTFs, mimicking how sounds are interpreted at the beginning of the auditory pathway at the level of the basilar membrane. The finding that the MS created using subject morphological parameters was not effective in describing the listening test results, with a distribution that was not statistically significant (p-value of 0.16), is consistent with findings from studies showing the prediction of subject HRTFs based solely on morphology did not improve localisation accuracy with respect to non-individualised HRTFs (Zotkin et al., 2003).

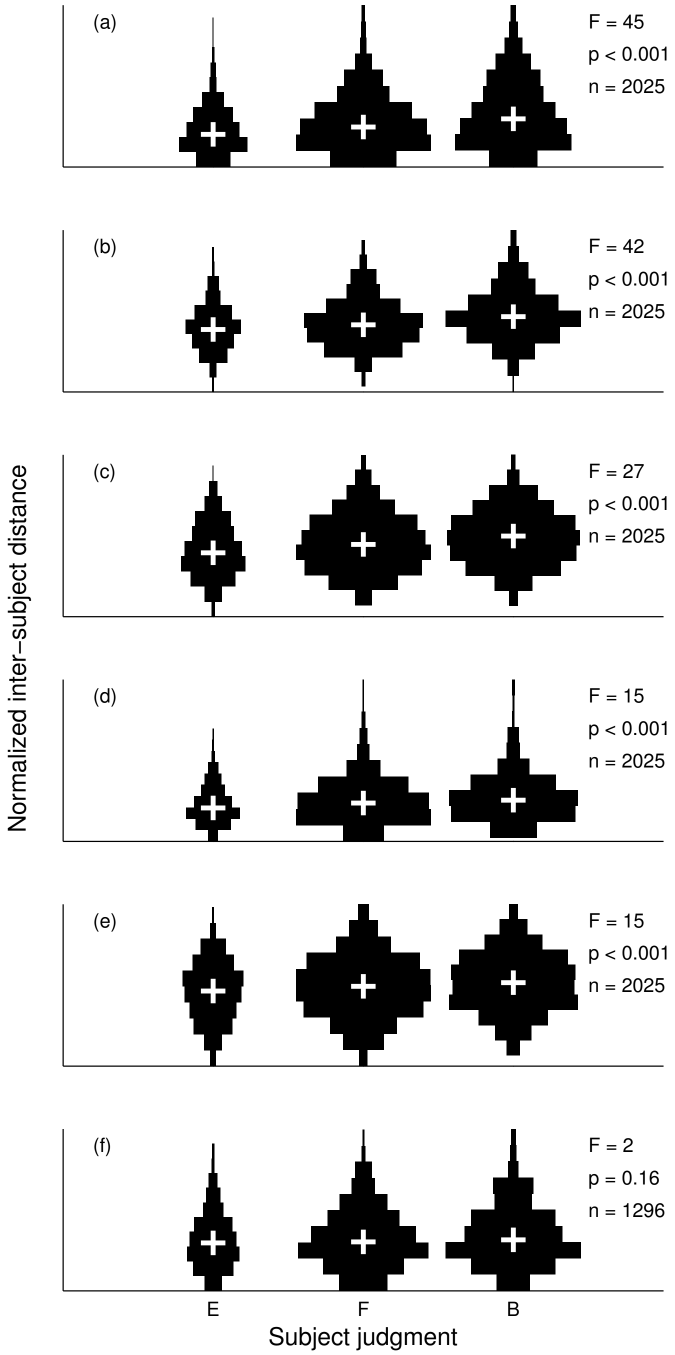


Figure 43: Histograms of distributions, across all subjects, of normalised distances between subjects and HRTFs judged as either *excellent* (E), *fair* (F), or *bad* (B), for the six MSs using either: (a) frequency scale factors, (b) frequency notches, (c) fractional octave-band filtered DTFs, (d) covert peaks, (e) unfiltered DTFs, or (f) morphology. The white + indicate distribution medians. All dimensions were used in each MS.

8.3.3 *Optimal frequency range and dimensions*

When considering the different analysis methods that used a PCA of DTF data, it is important to note that not all the spectral information was necessarily relevant in terms of the subjects' perceptual judgements. The simplest example of this is the human audible range of frequencies, which is between approximately 20 Hz and 20 kHz. An even more narrow range was explored in this study in order to establish the most perceptually relevant window of the spectrum and achieve a more effective MS. A large range of frequencies was used along with a statistical analysis based on the listening test results to find the optimal frequency range.

In addition to the optimal frequency range, it is possible that not all the dimensions in the created MSs would be perceptually relevant in terms of the listening test results or for the generation of robust MSs. There may be some dimensions, represented by their associated PC weights, that highlight specific features of the HRTF that are more significant than others. This is possible despite the fact that PCs are linearly dependent. An analysis was thus performed to compute the optimal selection of dimensions for techniques that used a PCA. For the other four analysis methods, a selection of optimal dimensions was also performed along with the use of the optimal frequency ranges.

The optimal frequency range was calculated for the fractional octave-band filtered and non-filtered DTF data by performing the previously mentioned statistical analysis described in section 8.3.2 and obtaining an F-ratio for a number of different frequency ranges. For each set of upper and lower frequency cutoff values, starting with a lower limit of 0 Hz and finishing with an upper limit of 20 kHz in increments of 250 Hz, a new matrix of concatenated subject DTFs was created and a PCA was performed. The MS created using the weights from the PCA was then validated for each particular frequency range. This analysis used all possible dimensions in its calculation and served as a global indicator to where the optimal frequency range might exist without taking into account the optimal selection of dimensions, which was performed as a second analysis and described below. Figure 44 shows the F-ratio values for each frequency range, with the lower limit plotted on the horizontal axis and the upper limit on the vertical axis. The results for the fractional octave-band filtered and non-filtered DTF data are shown in figure 44(a) and (b) respectively. The F-ratios are indicated on the plot contour lines and the regions on the plot are shaded according to the values; a lighter shade represents a higher F-ratio. The triangular regions at the bottom right corner represent frequency ranges for which the lower limit was larger than the upper limit and hence it was not possible to calculate an F-ratio.

For this analysis, the calculated optimal frequency ranges of the spectrum were considered the most perceptually relevant regions of subject DTFs due to the fact that they were calculated taking into account perceptual judgements via the listening test. It follows that these are the regions of the HRTF that are the most significant in terms of effectively interpreting sound sources in VAS with non-individualized HRTFs (see section 8.2.4 for details

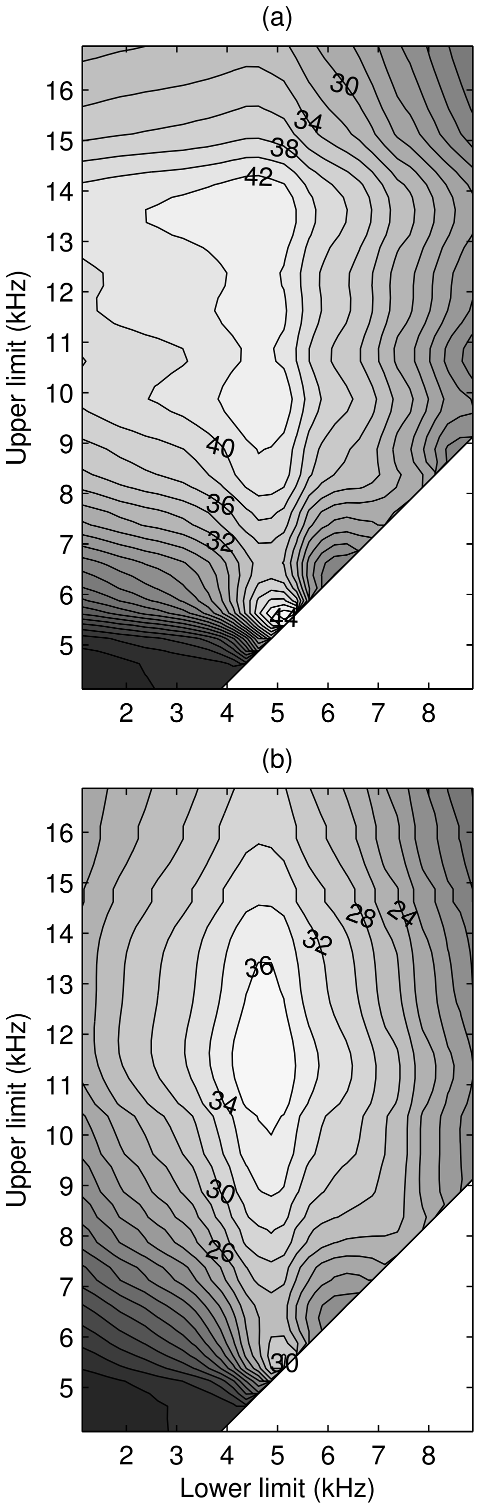


Figure 44: Contour plots of F-ratio values for different frequency ranges using all dimensions. F-ratio values are shown for MSs created via a PCA using two variants of the HRTF data: (a) non-filtered DTFs, and (b) fractional octave-band filtered DTFs.

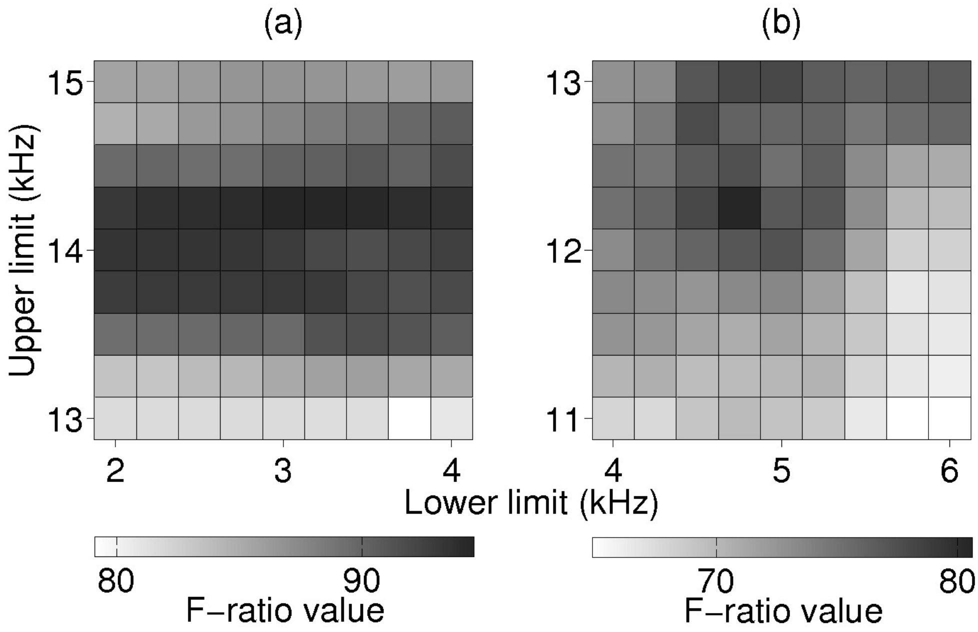


Figure 45: F-ratio values for different frequency ranges for a narrow region of frequencies using an optimal selection of dimensions for (a) non-filtered and (b) filtered DTFs. Each cell represents a frequency range. The shade of the cell corresponds to the effectiveness of the MS used; a darker shade translates to a more effective MS.

of the listening test). For the non-filtered DTF data, displayed in figure 44(a), the results showed that the optimal frequency ranges were those with lower limits between approximately 2 and 6 kHz and upper limits between approximately 9 and 14 kHz. There was in fact another region in which frequency ranges had a higher F-ratio value than the mentioned ranges, corresponding to a lower limit of 5.25 kHz and an upper limit of 5.75 kHz, however this frequency range was deemed a statistical anomaly rather than a region of the spectrum that contained the most perceptually relevant detail given the narrow window of frequencies it contained. The results demonstrate that if a lower limit of approximately 5 kHz is chosen, effective MSs can be created by using upper limits between approximately 9 and 13.5 kHz. It appears that using upper limits outside of these frequencies and in particular lower limits above approximately 5.5 kHz lead to poorer performing MSs.

The results for the filtered DTFs displayed a slightly different pattern, shown in figure 44(b), with the optimal frequency ranges having lower limits, between approximately 4 and 6 kHz, and upper limits between 10 and 14 kHz. The F-ratio values were slightly lower for the filtered DTFs than for the non-filtered DTFs. The differences in the results between the filtered and non-filtered DTFs are due to the significant smoothing of spectral detail in the DTFs, and the lower sampling in higher frequencies, when using a fractional octave-band filter.

With the global calculation of the optimal frequency ranges, it was then possible to compute, within these regions that showed higher F-ratio values when using all dimensions, what the optimal selection of dimensions were for both filtered and non-filtered DTFs. The chosen regions were assessed in sections that were 2 by 2 kHz, in increments of 250 Hz giving a total of 81 different frequency ranges. The regions that demonstrated the highest F-ratio using a selection of optimal dimensions were those with lower limits between 2 and 4 kHz and upper limits between 13 and 15 kHz for the non-filtered DTFs, and with lower limits between 4 and 6 kHz and upper limits between 11 and 13 kHz for the filtered DTFs. A selection of frequency ranges within the mentioned regions was used rather than all frequency ranges so that information about the significance of the dimensions was restricted to frequency ranges that were effective (i.e. a high F-ratio); using frequency ranges with a low F-ratio introduced noise into the calculation of the most significant dimensions. In addition, the use of a restricted number of frequency ranges reduced the calculation time of what was a computationally demanding analysis; the calculation time was reduced from a few days to a few hours.

For the calculation of an optimal selection of dimensions, a forward sequential feature selection method was used (see Guyon, 2008), for each frequency range within the selected region, by adding one dimension at a time of subject coordinates (or PC weights) and generating a loss function to test the combination of dimensions. The loss function chosen was the F-ratio described in section 8.3.2, calculated using the MS created from the chosen subset of dimensions. The most effective dimensions were added one at a time up until the loss function no longer improved. Thus, for each frequency range the most significant dimensions were obtained along with the order in which they were added, with the first being the most significant and the last the least significant. In this way, only the most perceptually relevant PC weights were retained, due to the fact that the loss function purely assessed how well a configuration of dimensions could describe the perceptual judgements of the subjects. For each of the 81 frequency ranges analysed, a MS was created and an optimal selection of dimensions was chosen and the F-ratio for the combination of frequency range and dimensions was recorded. The same sequential feature selection was used to generate a selection of optimal dimensions for the other analysis techniques not using a PCA, and also for the analysis method using only morphological parameters. For the analysis method using morphological parameters, the input data matrix was also reduced by using an optimal selection of columns (or morphological parameters) chosen using the forward sequential feature selection detailed in the next chapter.

The plots (a) and (b) in figure 45 show the F-ratio for the chosen frequencies within the region that showed the highest F-ratio for an optimal selection of dimensions. The optimal frequency ranges were calculated as 3.25 to 14.25 kHz and 4.75 to 12.25 kHz for the non-filtered and fractional octave-band filtered DTFs respectively. The highest F-ratio was obtained for the non-filtered DTFs with its selection of optimal dimensions and was calculated as the most effective MS in this study. Figure 46(a) and (b) show the dimensions used for each frequency range tested. Each of the frequency ranges tested are represented on the vertical axis with only the lower limit frequencies displayed; the upper limit frequencies range from 13 to 15 kHz and 11 to 13 kHz for the non-filtered and filtered DTFs respectively, in increments of 250 Hz between the displayed lower limits in increasing order from bottom to top. The dimension number is represented on the horizontal axis and displayed in descending order of importance (from left to right) as determined by the variance each PC describes in the original data. A cell in the plot is shaded if it was used in a MS for a given frequency range, and the shade of the cell represents the order in which it was added using the forward sequential feature selection. The shade of each cell therefore represents the importance of the dimension with darker shades being more important and lighter shades less.

It is clear from these results that overall there were some dimensions that were very effective in describing the perceptual judgements of the subjects and others that were not. These dimensions represented the subject weights for corresponding PCs that were effective at constructing a MS in which distances between subjects correlated well with the perceptual judgements from the listening test. This translates to dimensions, or weights, that were successful in distilling meaningful inter-subject differences. The dimensions that were not selected represented weights of PCs that, despite capturing significant degrees of the variance in the DTFs, were not perceptually relevant across all subjects.

The results show that only a small number of the ordered dimensions were needed in the effective MSs; over half the dimensions were not used if one considers dimensions that were used for at least 50% of the frequency ranges tested. Of the first 10 dimensions, which together described over 50% of the variance in the subject filtered and non-filtered concatenated DTF data, there were only three dimensions being used for at least 80% of the frequency ranges. The remaining seven dimensions from the first 10 were used for a maximum of just over 20% of the frequency ranges, demonstrating a clear selectivity for the most effective dimensions. Moreover, the three dimensions number 1, 2, and 6, for both the non-filtered and filtered DTFs, were almost always selected in this ascending order.

8.3.4 *Inspection of principal components*

An inspection of the PCs, calculated as part of the analysis of the optimal selection of dimensions in the MSs, was performed in an effort to determine if there were any differences between those that were selected as effective and those that were not. Remembering that the PCA returned PCs as concatenated values for all positions for the left and right ear, it was possible to isolate PCs for each of the measured DTF locations in the database. Each PC is therefore representing the different components of variance that existed between subjects for each position. The upper plot of figure 47 shows the magnitude of the PCs number 1, 2, and 3 as a function of frequency for the position directly in front of the listener (0° azimuth and 0° elevation). Only magnitudes for frequencies in the optimal frequency range (detailed in section 8.3.3) were considered. The lower plot of figure 47 displays all subject

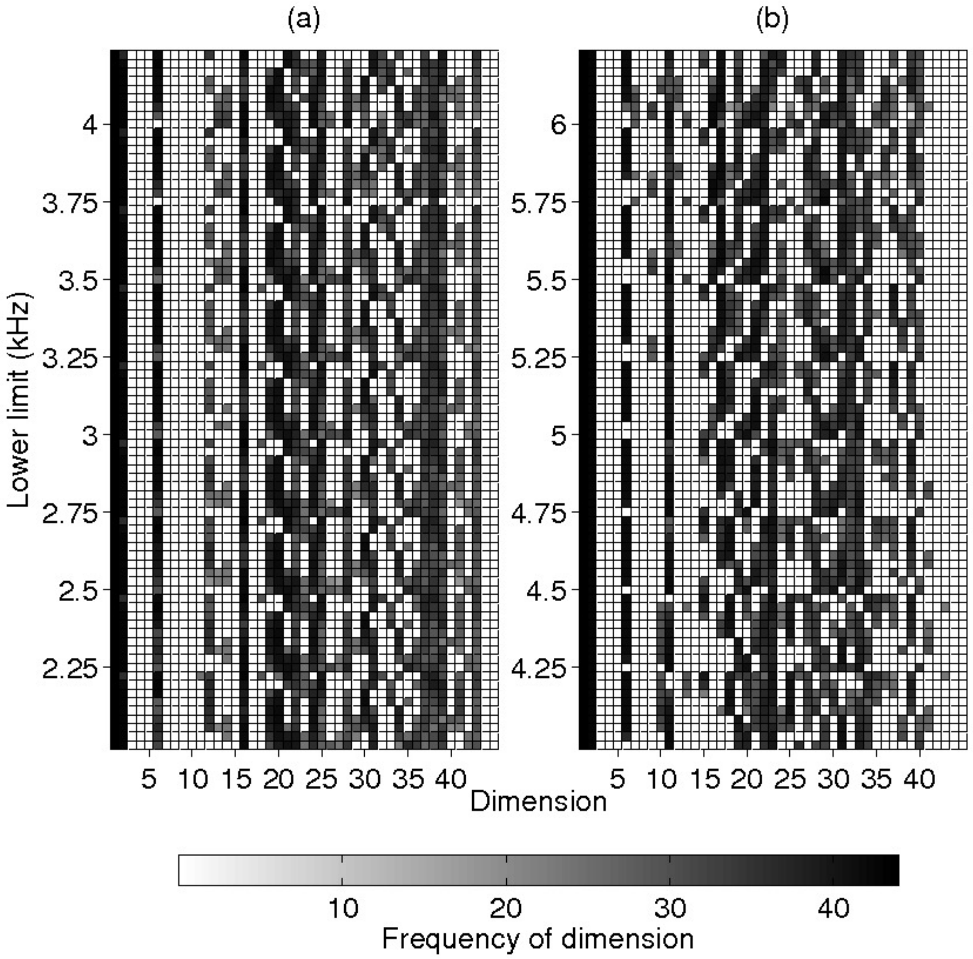


Figure 46: Dimensions used for each frequency range for (a) non-filtered and (b) fractional octave-band filtered DTFs respectively. The shade of the cells corresponds to the significance of the dimension; a black cell represents a dimension that was selected first for a MS and a white cell a dimension that was not selected.

DTFs for the same corresponding position in space, colour-coded green and red representing DTFs for which the first PC weight was positive and negative respectively. The colour-coding was performed in order to get some crude separation of the PCs in order to highlight how PC weights might be used to classify subject HRTFs.

With respect to the upper plot, the first two PCs 1 and 2 were always selected as the most significant, as can be seen in figure 46, and are shown in bold. The third PC was almost never selected. The results show that PC number 1, the PC that explains the largest degree of variance in the data, describes changes in the DTF spectrum at higher frequencies since magnitudes for lower frequencies are near zero, which will make reconstructed DTF magnitudes near zero no matter the weight values. If this is compared with the subject DTFs in the lower plot, it can be seen that there is a large degree of variance across subjects in the same frequency region that PC number 1 has a deviation in its magnitude. Note that the deviation seen in PC number 1 in the higher frequency region is not related to any notches or peaks in the subject DTFs, rather the degree of observed variance across subjects.

The lower plot shows that the subject DTFs are separable based on whether they had a positive or negative weight for the same higher frequency region (approximately 10 to 13 kHz) in which PC number 1 had the largest deviation from zero. For every frequency magnitude bin between 10,250 and 12,750 Hz, an ANOVA showed that the DTFs were statistically separable, tested at a level of significance of 0.05.

The result demonstrates the effectiveness of the weights, calculated from a PCA, to delineate the subjects according to their DTFs. This is an obvious simplification of how the weights can be used to distinguish the subject DTFs given that their magnitudes have not been taken into account and only their sign considered, yet it demonstrates the point clearly. If the weights for PC number 1 are significant in terms of describing the results from the listening test, which was suggested by the fact that their corresponding weights were shown to be significant (see figure 46), then it may be inferred that the mentioned region of higher frequencies was a perceptually relevant aspect of the HRTF. Conversely, PC number 3, which was rarely included in any of the optimal MSs, might not have been perceptually relevant given that it is describing variance in the lower frequencies.

It must be noted however that the figure is only presenting the PCs for one position in space among many used in the PCA, making it difficult to draw conclusions from the weights, which are single values for all positions in space. In addition, PCs number 2 and 3 are somewhat difficult to interpret given the fact that PCs are linearly dependent. Other studies that have inspected PCs from a PCA of subject DTFs have been cautious to draw conclusions about their perceptual relevance (Kistler and Wightman, 1992). Nevertheless, the representation of the PCs and subject DTFs provides an insight into how certain dimensions, or PC weights, in the MSs might be more effective than others at describing the results from the listening test. Future work might be able to quantify the differences between PCs, and relate these differences to the observed weights.

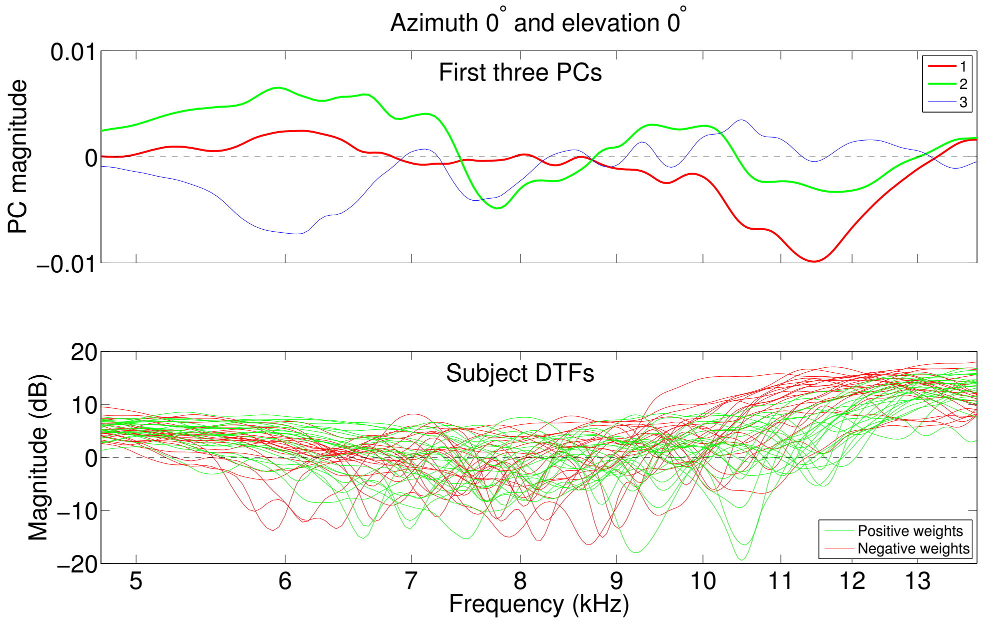


Figure 47: The first three PCs for the position 0° azimuth and 0° elevation are shown in the upper plot. The red and green lines in bold represent the PCs that were the most effective, and the blue thin line represents a PC that was almost never selected. The lower plot of the figure displays subject DTFs for the same position in space, colour-coded according to the sign of the weights of the first PC.

8.3.5 *Validation of optimised multidimensional spaces*

Each of the MSs using the six different methods was recalculated using an optimal selection of dimensions, and in the case of the methods using a PCA of subject HRTFs, using an optimal frequency range. Figure 48 shows the distribution of inter-subject distances as a function of the three judgement categories: *excellent* (E), *fair* (F), and *bad* (B). The results show an improvement in the effectiveness of these MSs to describe the perceptual judgements from the listening test results with respect to the MSs created using all dimensions (figure 43). The notable difference in the results was the proportionally larger improvement of the methods that used an optimal frequency range; the methods using a PCA of subject DTFs were now the most effective, whereas with the analysis using all dimensions the method using FSFs was the most effective.

8.3.6 *Comparison of different multidimensional spaces*

Further to the statistical analysis of the optimised MSs shown in figure 48, a calculation of whether the inter-subject differences that were described using the different methods resembled each other or not was performed. In other words, how well did the different MSs correlated with one another in terms of the Euclidean distances between subjects. Each of the six optimal MSs shown in figure 48 were included for comparison, making a total

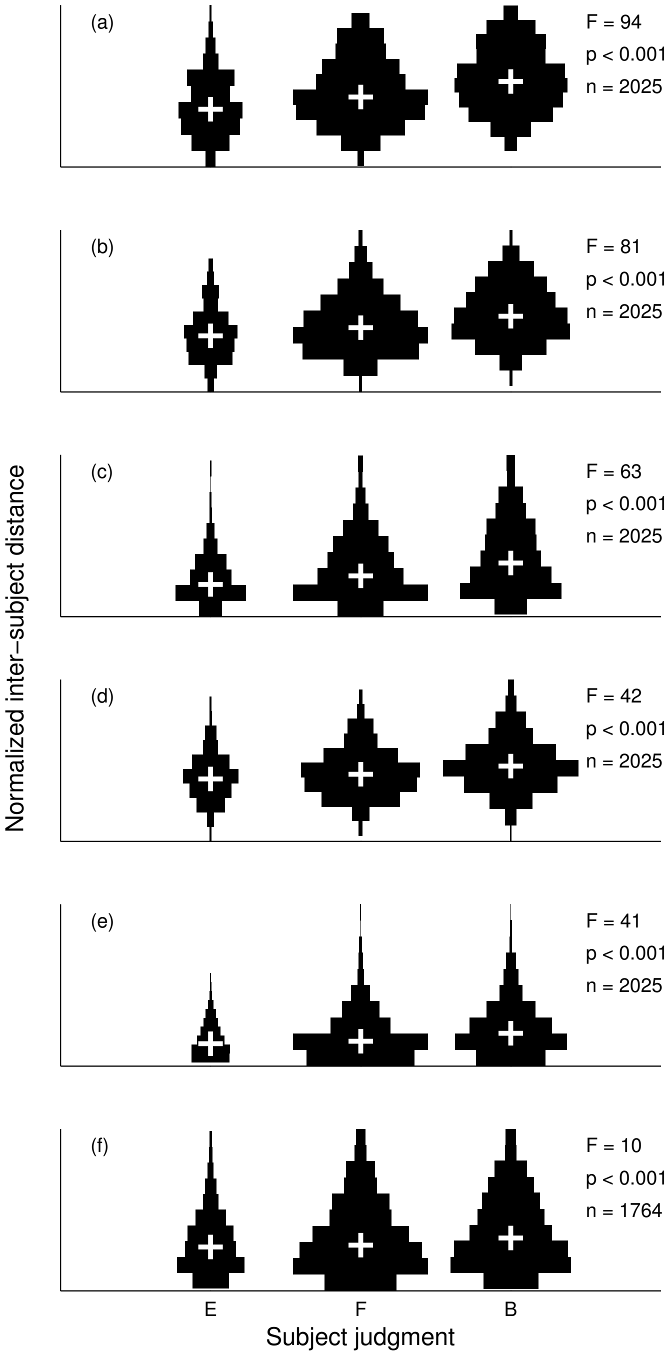


Figure 48: Histograms of distributions, across all subjects, of normalised distances between subjects and HRTFs judged as either *excellent* (E), *fair* (F), or *bad* (B), for the six MSs using either: (a) non-filtered DTFs, (b) fractional octave-band filtered DTFs, (c) FSFs, (d) frequency notches, (e) covert peaks, or (f) morphology. The white + indicate distribution medians. The optimal set of dimensions were used for each MS.

of 15 pairwise comparisons. For each MS, a vector of pairwise distances between subjects was calculated. This corresponded to the distance between each subject as represented in the MS using the various analysis methods from the PCA of concatenated DTFs to the multidimensional scaling of optimal frequency scale factors. Not all subjects that took part in the listening test were used for this comparison; three subjects were removed from all the MSs, corresponding to those that did not have measured values for the MS using only a selection of optimal subject morphological parameters.

The calculated distances were standardised so that they had a mean of zero and a variance of one. This allowed for a normalised comparison of subject pairwise distances between the different MSs. The standardisation was necessary as the different analysis methods had varying types of data that used different units of measure, for example decibel magnitude for the HRTF data and scale factor for the multidimensional scaling method. Correlation coefficients were then evaluated between the different vectors of inter-subject distances for each pair of MSs being compared. The correlation coefficients were taken as a measure of how the MSs correlated with each other and are provided in table 8. A relatively strong correlation existed between the analysis methods using: non-filtered DTFs, fractional octave-band filtered DTFs, and FSF data. These are the same three methods that had the most effective MSs with respect to the measured F-ratios (see figure. 48).

The results highlight, interestingly, a similarity between the MSs using DTFs (non-filtered and fractional octave-band filtered DTFs) and FSFs, with correlation coefficients of 0.69 and 0.68 respectively. The high correlation coefficient of 0.8 for the comparison of the MSs using non-filtered and filtered DTFs on the other hand was expected given the high degree of similarity in the data used in the PCA. These results indicate that the different analysis methods were generating the same inter-subject distances and hence comparable spaces, albeit it on different scales, despite the fact that they incorporated significantly different aspects of the subject HRTFs. It is noteworthy that the higher correlation coefficients correspond to comparisons of the most effective MSs in terms of the listening test results. This suggests that the lower correlation coefficients for the other pairs of MSs was due to the fact that the orientation of the subjects in space, and the distances between subjects, were not generated using methods that effectively distilled the most perceptually salient components of the HRTFs, or that the data being analysed was not relevant to the quality of the rendering in VAS.

8.4 DISCUSSION

The current study aimed at globally assessing what components of the HRTF are perceptually relevant in terms of how humans perceive their acoustic environment. To this end, the study highlighted the extent to which the information in an HRTF can be reduced via the different analysis methods by representing subjects in multidimensional spaces with inter-subject differences in DTFs corresponding to Euclidean distances between subjects. By comparing how these different MSs were able to describe the results from

	FILT. dtf!s	FREQ. SCAL.	NOTCH	PEAK	MORPH.
UNFILT. DTF	0.8	0.69	0.31	0.29	0.04
FILT. DTFS		0.68	0.28	0.32	-0.01
FREQ. SCAL.			0.32	0.52	0.06
NOTCH				0.33	0.09
PEAK					0.13

Table 8: Correlation coefficients comparing vectors of pairwise distances between subjects in the six different MSs.

the listening test it was possible to gauge how effective the different analysis methods were at distilling the relevant spectral information embedded in the HRTF. The proposed method of validating the different MSs is novel and statistically powerful due to the fact that it incorporated a significant number of subjects, leading to a large number of pairwise comparisons or instances in the dataset, which made it somewhat resistant to noise in the results from the listening test. This is important due to the fact that a previous study by the authors (section 7.10) showed a large degree of variance in perceptual judgements for a significantly smaller number of HRTFs tested via a listening test (six as opposed to 45 in this study; see Schönstein and Katz, 2011). Despite clear differences in the judgement criteria used in the listening test in the current analysis and in the mentioned study, the results suggested that a significant number of judgements from the current study might be somewhat interchangeable if the listening test was repeated many times, especially for non-expert subjects. It is seen as crucial that when using perceptual judgements of HRTFs to validate any type of model that describes their role in perceiving sound sources, that a statistical technique robust to noise be used (as in the current analysis) or many repetitions of the listening test be performed (as in the mentioned study).

The non-filtered DTFs were more effective than the fractional octave-band filtered DTFs in predicting the subjects' perceptual judgements. This is a somewhat unexpected result given that the filtered DTFs represented the HRTFs on a more perceptually relevant scale in the frequency domain. One possible explanation might relate to findings that have shown the compression efficiency of a PCA is greater for HRTF data in a linear frequency domain than a logarithmically spaced frequency domain (Leung and Carlile, 2009). Since a subset of PC weights were used to represent the subjects in a MS, this could have had an effect on how inter-subject differences were described. Overall, the results from these PCA methods add to a large body of existing work demonstrating their suitability in describing differences in HRTFs between subjects.

The FSFs, despite incorporating very different aspects of subject DTFs and drastically reducing the volume of DTF information, generated MSs that were comparable, in terms of the distances between subjects in the space (see section 8.3.6) and their effectiveness in describing subject perceptual judgements (i.e. comparable F-ratio values in figure 48), to the MSs created using the PCA of DTFs. In comparison, the MS created using notch frequencies was not comparable in terms of the correlation coefficients and was not as effective. This finding suggests the importance of covert features in HRTFs, such as the alignment of peaks and notches across all positions, as opposed to more overt features such as the correlation between notch frequencies for individual positions, in describing relevant inter-subject differences. This hypothesis is supported by the fact that the results from the comparison of the different MSs suggest that FSFs are mostly a function of the alignment of covert peaks between subject DTFs due to the fact that the correlation coefficient was higher between the method using FSFs and covert peaks than FSFs and notch frequencies (0.52 and 0.32 respectively; see table 8). This is contrary to some studies showing the more prominent role of notches over peaks for localisation tasks (Greff and Katz, 2007; Iida et al., 2007), yet these studies were not explicitly analysing convert peaks, rather the frequency of peaks in each individual HRTF. There are studies that have highlighted the importance of covert peaks by showing that narrow-band centre frequency can be an effective predictor to sound source location cannot necessarily be translated to broadband localisation, as suggested by Middlebrooks (1992), or in the case of the current study, to judgements of broadband virtual sound sources.

The analysis methods using PCA further allowed the exploration of optimal frequency ranges. These were calculated to be between 3.25 and 14.25 kHz for the non-filtered DTFs and 4.75 and 12.25 kHz for the filtered DTFs, which is in line with numerous studies suggesting that this is the region in which the most significant spectral cues lie (Hebrank and Wright, 1974; King and Oldfield, 1997). This frequency range is essentially the region of the spectrum that is influenced by the pinna, given the wavelength of these frequencies and the scale of the outer ear. This finding does not imply, however, that all the information in this window was relevant for perceiving sound sources. In fact, the results suggested more localised spectral regions of importance. From the results of the analysis using all dimensions in the MSs, significant upper limits were found between 9 and 14 kHz and 10 and 13 kHz for the non-filtered and filtered DTFs respectively. Lower limits were consistently most effective at 5 kHz, showing a sharp decline in performance as they approached 6 kHz. A more narrow frequency region of the spectrum between 5 and 6 kHz was also found to be significant in the current study. These regions correspond well with previous works using localisation tasks in VAS, instead of listening tests, by Langendijk and Bronkhorst (2002). It was found that cues in the 6 to 12 kHz band tested were important for elevation and the 8 to 16 kHz band tested was important for determining whether a sound source was in front or behind the listener. It was also suggested that due to the interaction between up-down and front-back localisation errors that there may have been an overlap of the different cues in the bands tested. This overlap may further support the more distinct regions found in the current study; the analysis method employed enabled a much higher resolution of frequency bands for validation. Another study by Bronkhorst (1995) also demonstrated the role of font-back cues in the region between 7 and

16 kHz and suggested spectral cues at 6 kHz were used to determine the up-down location of sound sources. Langendijk and Bronkhorst (2002) proposed that spectral notches, defined as narrow frequency features, were not crucial to localisation accuracy based on the results obtained. This proposition is supported by the fact that the MS using FSFs was more effective than the MS created using only notch frequencies. Langendijk and Bronkhorst (2002) also found no evidence for the role of spectral cues below 4 kHz and proposed that spectral notches in this lower frequency region, defined as narrow frequency features, were not crucial to localisation accuracy based on the results obtained. These are findings that have been mirrored in the current body of work in terms of a validation using the subjects' perceptual judgements.

The results from the PCA also allowed for further insight into what the most perceptually relevant components of the HRTF are, with some dimensions, and their corresponding PCs, being clearly more effective than others. This is a novel approach to PCA of DTFs in the sense that it has always been assumed that all the PCs, or at least the PCs that are describing the majority of the variation in the dataset, are describing relevant differences in the spectra between subjects or positions due to their linear dependence. The corresponding PCs of the most effective dimensions in the current study can be thought of as covert components of the HRTFs that best describe features relevant to what subjects judged as being a high fidelity rendering in VAS in the listening test. Dimensions not used in the most effective MSs can be thought of as having PCs that described variations in DTF spectra that were not necessarily relevant to this criteria. It is important to note that these selected PCs are specific to this database and cannot be generalised to other PCAs of DTFs. Furthermore, in this analysis each PC described individual positions in space in a concatenated representation due to the fact that the input data to the PCA was organized as rows of concatenated subject DTFs for the left and right ear. This makes it hard to draw conclusions relating to the form of the PCs; the PCs can only be visualised for discrete positions across all subjects.

For studies that performed a PCA across all positions and subjects (Kistler and Wightman, 1992; Middlebrooks and Green, 1992), as opposed to using a PCA to establish inter-subject differences as was the case in the current study, there have been attempts to describe the function of PCs. These studies found consistent results across databases of HRTFs in terms of the PCs calculated but have been unable to successfully elaborate on any underlying function attributed to individual PCs. In the study by Kistler and Wightman (1992) it was inferred from the first PC that the main source of variance among the DTFs was the amount of energy in the high frequencies and that this was important for perceiving the laterality of sound sources in terms of ITDs and ILDs. The remaining basis functions (PCs 2 to 5) were explained to be less interpretable, probably mediating the distinction between up-down and front-back location of sound sources. Despite the difficulty in interpreting the individual PCs, the findings in this study demonstrate that a PCA can draw out a selection of components of the HRTF that are clearly preferred, and clearly not preferred, across all subjects in describing perceptual judgements. This finding is of interest because there is nothing implicit in the way PCs are calculated via a PCA that defines them as being perceptually distinct, whilst the results in this study suggest that they are.

Taken together, the results from the current study support the notion that different regions and components of the spectra are significant for perceiving sound sources in space and that neither covert peaks, described in studies using narrow-band stimuli (Blauert, 1969; Middlebrooks, 1992; Humanski and Butler, 1988), or frequency notches (Hebrank and Wright, 1974; Bloom, 1977) alone are sufficient as auditory cues to sound source location (Greff and Katz, 2007; Iida et al., 2007, ; see also Carlile et al., 2005 for a review). Most of the audible spectrum of the HRTF appears to be utilised for effectively interpreting sounds in our environment, rather than narrow-band cues; these results are supported by findings by Alves-Pinto and Lopez-Poveda (2005) showing that high frequency notches themselves are detected in the overall spectral shape rather than over certain frequency regions. It is most likely that the results from the study in this chapter point to a very important feature of the human auditory system; its ability to code for a vast array of spectral features. This redundancy in auditory processing may have evolved as a survival advantage due the fact that cues can be severely deteriorated, such as in a reverberant space, depending on the environment, and depend on the spectral content of the source in question, which can be limited and can vary between sources.

8.5 CONCLUSION

The contribution of different aspects of the HRTF to binaural synthesis has been examined from a psychoacoustic viewpoint. A model for describing HRTF inter-subject differences, using an optimal frequency range of 3.25 to 14.25 kHz and a PCA of HRTFs on a linear frequency scale, was shown to best reflect perceptual judgements from a listening test for a large number of subjects. The order in which the different models ranked against each other provides a valuable insight into what spectral features of the HRTF are most perceptually relevant. The analysis also showed that an alignment of peaks and notches between subject HRTFs, via an optimal frequency scale factor, better represented the perceptual judgements than methods used to describe inter-subject differences using notch frequencies or the location in space of covert peaks alone.

To the extent that the PCs from the PCA described the perceptual judgements, the results indicated that some PCs were more effective than others, highlighting perceptually relevant variations in the spectrum of the HRTFs across subjects.

The confirmation that methods such as those using frequency scale factors are good descriptors of the inter-subject differences between HRTFs suggests that information about complex spectral differences can be effectively distilled in the form of an optimal frequency scale factor for example. This may prove useful for applications that seek to correlate listener morphology to differences in HRTFs in order to produce sophisticated customisation techniques similar to that explored in the next chapter.

9 SIGNIFICANT MORPHOLOGICAL PARAMETERS FOR BINAURAL SYNTHESIS

The purpose of this final study was to examine a variety of techniques used to highlight the role of a listener's morphology with respect to the filtering effects described by the Head-Related Transfer Function (HRTF). The methods used aimed to distinguish the most perceptually significant morphological parameters with respect to their ability to describe perceptual judgements of a binaural synthesis. More specifically, this chapter draws on the results of the previous two chapters and the knowledge that there are key physical dimensions of the listener, namely the size of the head and dimensions of the outer ear, that influence the HRTF most. The results from the current chapter looked to find the link between measured morphological parameters and HRTFs judged to be effective in Virtual Auditory Space (VAS) in an effort to better understand how the body interacts with incident waves from a sound source in space. It was hoped that out of this analysis a method for selecting an appropriate set of HRTFs from a database for a listener could be validated. If effective, the selection procedure would allow one to bypass the laborious procedure of recording individualised HRTFs, making it amenable to consumer markets.

9.1 BACKGROUND

As detailed in in the previous chapter, the asymmetrical form of the human outer ear, and to some degree the dimensions of other parts of the body, greatly affect how a particular listener will perceive sounds in space. The human auditory system is finely tuned to the diffraction caused by a listener's head, the reflections caused by the upper body, and the reflections/diffraction or resonances caused by the pinna, in his or her acoustic environment. The pinna in particular plays a crucial role in helping listeners resolve the so-called cone of confusion, which is a function of interaural time and level differences, and as such is the focus of much work on the topic. Previous studies have looked at the specific modes of resonance that may contribute to components of the HRTF such as peaks and notches (Shaw and Teranishi, 1968; Musicant et al., 1990), while others have been more restrained in drawing direct links between spectral features and specific parts of pinna suggesting that a one-to-one correspondence is almost impossible due to the fact that reflections from any one part of pinna influence others (Lopez-Poveda and Meddis, 1996). Despite this body of work, there exists few studies, which use a perceptual rather than a signal viewpoint to investigate the most significant morphological parameters, which will be one of the main purposes of the current chapter.

9.2 PREDICTION OF SUBJECT LOCATION IN MULTIDIMENSIONAL SPACES

The previous chapter described a study in which subjects from the LISTEN database (see section 6.2.3) were positioned in Multidimensional Spaces (MSs) according to inter-subject differences. These inter-subject differences were based on analyses of the subject HRTFs, with a variety of different methods tested, each describing different aspects of the HRTF spectrum. The study allowed for a description of the most salient features of subject HRTFs with respect to the results from a listening test (described in section 7.3). Whilst these results were the focus of the study, the analyses also allowed for the selection of the most effective MS that could be used to study the significance of different morphological parameters.

9.2.1 *Method*

The most effective MS from the study described in the previous chapter was that using the non-filtered Directional Transfer Functions (DTFs) via a principal component analysis (see section 8.2.3.1). As previously explained, this MS was shown to most effectively describe the perceptual judgements made by the subjects in the LISTEN database using a listening test. The effectiveness of the MS was computed by comparing the distance between each pair of subjects (with a smaller distance signifying a more similar pair of HRTFs) to the perceptual judgements (registered as either *excellent*, *fair*, or *bad*) made by one subject in the pair concerning the other subject's HRTFs and vice versa.

With the most effective MS computed, the next phase in the study was to establish whether the subjects' position could be predicted without using any HRTF data. This would allow for a subject from outside of the HRTF database to have a selection of HRTFs chosen for them by predicting their position and choosing the nearest HRTFs in the MS. This is desirable for commercial applications, as described previously, since measuring HRTFs is a laborious and expensive task. Subject morphological parameters are an obvious candidate for use in this predictive technique as they have a direct bearing on the variation in the spectrum of HRTFs and are relatively easy to measure (e.g. via a photo of the listener). The author developed a procedure implemented in the numerical coding environment Matlab to predict HRTFs for the listener based on morphology.

This analysis used morphological data for subjects in the database, and by way of deriving the most effective predictive model, generated a subset of the most significant morphological parameters. The analysis thus used the predictive model as a vehicle for establishing what might be some of the most significant morphological parameters of the ear and body with respect to subjects' perceptual judgements. The method used was akin to that used to find the optimal frequency range and selection of dimensions for the MSs using Principal Component Analysis (PCA) in section 8.3.3. For this analysis however, the frequency range and number of dimensions was held constant whilst the selection of morphological parameters varied as part of a forward sequential feature selection. One morphological parameter was added at a time, testing how well they were able to predict the subjects' positions in the MS via a regression. The overall effectiveness of the subjects' predicted positions was assessed by how well they described the results from the listening test. A statistical validation was performed across all subjects for each combination of sequentially added morphological parameters. The validation involved first generating subject-specific MSs by removing each subject's filtered DTF data from the input matrix, and then performing a PCA and using the corresponding weights. The weights from the subject-specific MSs represented the coordinates of the remaining subjects judged by the removed subject. The optimal frequency range from section 8.3.3 was used in the creation of the input matrix as it was shown to highlight the most perceptually relevant window of the spectrum relative to HRTF selection from a database.

The removed subject's position in the MS was then predicted by firstly performing a regression of the other subjects' (i.e. all subjects except the removed subject) coordinates on the corresponding selected morphological parameters, one dimension at a time. The coefficients from the regression and the removed subject's morphological parameters were then used to determine the removed subject's coordinates. The regression aimed at creating a link between Principal Component (PC) weights for each subject and their corresponding morphological measurements. Thus for each subject a set of predicted coordinates was calculated, and these were added to the existing subject-specific MSs so that all the subjects had coordinates in the space. A validation was then performed across all subject-specific MSs for each combination of selected morphological parameters.

The validation of the subject-specific MSs for each subset of morphological parameters was the same as described in section 8.3.2. Distances from the removed subject (the predicted position) to all the other subjects in the MS were evaluated and paired with the removed subject's perceptual judgements of the other subjects' HRTFs. These distance values and corresponding judgements were calculated for each subject (and each corresponding subject-specific MS) and grouped by judgement type: *excellent*, *fair*, or *bad*. The process was repeated for all but eight subjects; these subjects were removed from the analysis due to the fact that they had missing values for the morphological parameters used in the regression (see section 8.2.2). The same optimal dimensions, calculated for the MS in section 8.3.3 using filtered DTFs and all subjects, were used in the subject-specific MSs (i.e. with one subject removed at a time). The use of these dimensions was validated, via an inspection of the PCs for input matrices that were subject-specific and the input matrix that used all subject DTFs, showing them to be almost identical. The mean difference between corresponding PCs was seven orders of magnitude smaller than the PC values themselves.

9.2.2 *Results*

The F-ratio of the analysis of variance of distance values (see section 8.2.4), across all subjects for the three judgement groups, was then used as the loss function for the forward sequential feature selection so that only morphological parameters that helped minimise the loss function were added; parameters were added until the loss function could be minimised no further. Using this methodology, the optimal morphological parameters, from most to least significant, were: d_6, x_6, x_{12}, θ_1, x_1, x_2, x_{16}, and d_7, corresponding to pinna width, neck width, shoulder width, pinna rotation angle, head width, head height, head circumference, and intertragal incisure width (see figure 36). A similar study by Hugeng et al. (2011) also found eight morphological parameters as being most significant, using multiple linear regressions and a signal validation; the same set of CIPIC dimensions were used along with a total of 37 subjects. Four of the eight parameters (d_6, x_6, x_{12}, and x_1) selected in the study by Hugeng et al. (2011) were also identified in the current study.

The distribution of the normalised distance values for the three judgement groups (*excellent*, *fair*, and *bad*) is shown in figure 49(b) along with corresponding F-ratio, p-value and number of data points, similar to the results shown in figure 48 from the previous chapter. Figure 49(b) is not directly comparable with the distributions in figure 48 due to the fact that each MS used to generate the former contained one less set of subject DTFs. In addition, MSs created via the regression were validated with eight less subjects, represented by the lower n value with respect to plots in figure 48. In order to have a comparison of the effectiveness of this regression methodology, a second analysis was performed using the MS with all subjects, and validated only using the same reduced number of subjects as for the regression. This analysis still involved the use of a MS created using one more subject than the subject-specific MSs, as one subject was removed for each, yet gives a better comparison of the effectiveness of the regression.

Figure 49(a) shows the distribution for this adapted validation procedure without regression. It can be seen that the effectiveness of the technique using the regression approached to that of the MS using only subject DTFs as input and no regression; F-ratio values were 82 and 34 respectively (both statistically significant results tested at the 95% confidence interval). This suggests that the selected morphological parameters were able describe a significant amount of the variance in subject DTFs in terms of the results from the listening test. More importantly, the described procedure was shown to be able to predict subjects' best HRTFs from a database at statistically significant level. The absolute numbers of correctly predicted HRTFs are less important than the relative statistical significance given that the subject judgements themselves are known to be somewhat unreliable for such a large number of judged HRTFs (see chapter 7).

Due to the fact that the regression methodology treated the removed subjects as DTFs coming from outside the database by completely removing them from the calculation of the MS, it follows that this prediction technique might be a viable tool for HRTF selection for listeners on a commercial scale.

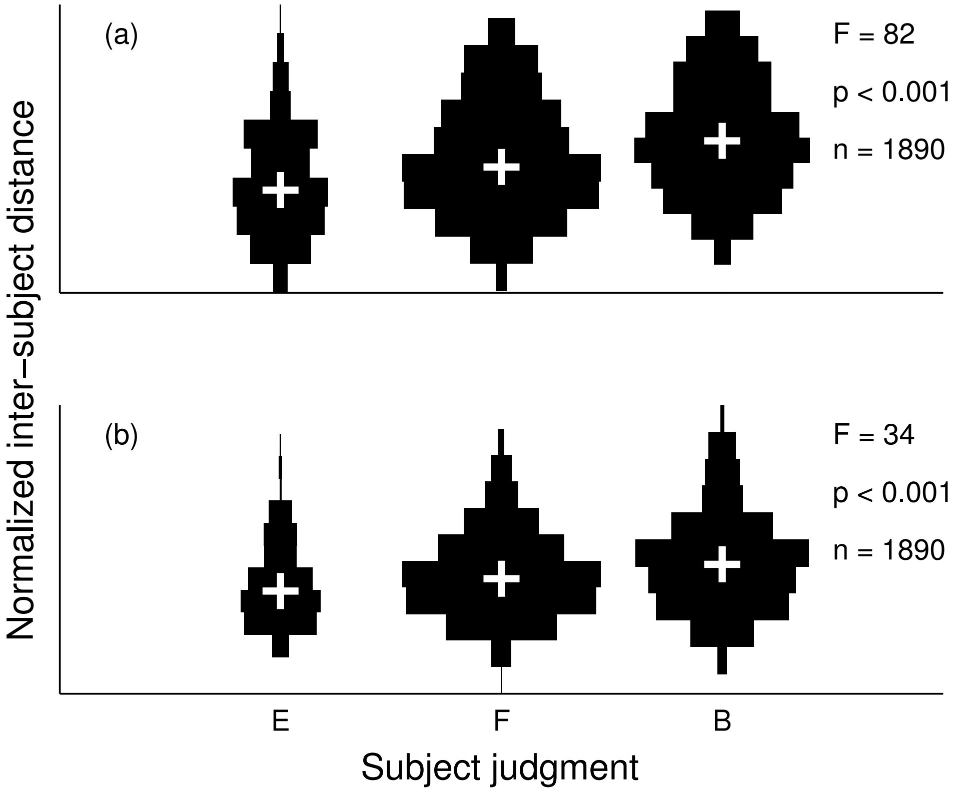


Figure 49: Histograms of distributions, across all subjects, of distances between subjects and HRTFs judged as either *excellent*, *fair*, or *bad* (labeled *E*, *F*, and *B* respectively) for the different MSs using: (a) non-filtered DTFs and subjects' calculated positions and (b) non-filtered DTFs with a regression using morphological parameters.

9.3 MACHINE LEARNING

The previous section demonstrated the potential for a HRTF selection technique and the results were used to highlight what might be some of the most significant morphological parameters based on those measured in the database. These parameters were selected with the assumption that inter-subject differences, and the MSs they help generate, are best represented using weights from a PCA of subject DTFs, along with an optimal frequency range and selection of dimensions. In reality, there exist other methods for selecting and analysing morphological parameters based on the perceptual judgements from the listening test. As demonstrated in the previous chapter (section 8.2.3.6 in which a PCA was performed on the measured morphological data), the link between the judgements and subject morphology can be analysed in the absence of subject HRTF data. This section describes two such techniques from the field of machine learning that were used to further validate the optimal selection from the previous section. An open source data mining software called Weka (version 3-6-2) was used for all calculations.

Machine learning, and the broad field of data mining itself, aims to distinguish patterns within large quantities of data. In machine learning, examples in the data that illustrate relations between observed variables can be used to capture characteristics of interest. In the case of this study, the examples, or instances, that might contain relations of interest were represented by the subjects' morphology and perceptual judgements. Algorithms along with the data were used to teach the computer in a sense how to recognise complex patterns and make intelligent decisions. The outcome is a set of rules that can be applied to the whole database in general and therefore can be used as a predictive model, in the case of this study, to select HRTFs from a database for a listener based on their morphology. As with the previous section, given the known variance in perceptual judgements of HRTFs demonstrated in chapter 7, which introduces noise into the analysis, the predictive power of these models, with respect to how many HRTFs could be effectively selected for each subject, is less significant when compared to what morphological parameters are actually used in the process. The overall statistical significance of the results for a particular model, related to how well it described the listening judgements, is merely a vehicle for drawing out the optimal morphological parameters as described in section 8.4 of the previous chapter. The morphological parameters used offer an insight into their significance in the filtering of sound sources in space by the body, and the role of inter-subject differences based on these dimensions. These insights are based on the effectiveness of a binaural synthesis using non-individualised HRTFs since they use results from the listening test described in chapter 7.

9.3.1 *Decision trees*

The previous section demonstrated the potential for an HRTF selection technique and the results were used to highlight some of the most significant morphological parameters. In the current section, a machine learning algorithm known as decision trees, which generates a predictive model, mapping the described instances and generating rules about the different values of the features used was studied. Decision trees are a set of rules, based on a dataset, organised as a hierarchy, or more specifically, organised as a tree as the name suggests. These tree structures are intended to classify instances, which are records, in any dataset that is to be modelled. The rules are organised as leaves and a set of branches that represent conjunctions of particular features or fields of the dataset. Once a decision tree has been built using a dataset, any record can be classified by following the branches and rules until there are no more branches, at which stage there is a classification based on a final rule. The purpose of this analysis was to use the optimal selection of eight morphological parameters from the previous section to try and develop a predictive model that did not require HRTF measurements, while also gaining insights into the role of the different selected parameters.

The particular decision tree algorithm used in this study was that created by Quinlan (1993) called the C4.5 algorithm. A decision tree is built based on a large number of instances using the concept of information entropy. The algorithm operates recursively by finding the attribute, or in this case the morphological parameter, that most effectively splits the instances according to the two or more assigned classes. In this sense the algorithm is a divide-and-conquer procedure in which the problem of classifying the instances is recursively broken down until all instances are satisfactorily classified. The effectiveness of the split at each node is measured using what is known as information gain. The information content is first calculated as the sum of the average logarithm of the ratios of classified instances split down each branch at a node for the set of variables or values being classified.

The information gain is calculated as the difference between the overall information content for a particular attribute, or morphological parameter in this study, and the average information content based on the classified variables or values as described above. This process is more intuitive for nominal data in which there are variables, but works equally well for continuous data. The information gain is calculated for each attribute being tested and the largest gain determines the choice of attribute at a particular node and the process continues recursively. An advantage of decision trees is that missing values can also be incorporated into the instance data using a variation of the information gain calculation (see Kohavi and Quinlan, 1999, for details).

The C4.5 algorithm can incorporate what is known as pruning in order to simplify the number of nodes created and provide results that are easier to interpret. This is necessary as the described algorithm will continue recursively until it has a perfect classification, causing an excessive number of rules. The pruning also makes any rule structure created by a decision tree more generalised and applicable to the training data by avoiding a strict separation of the classes, which is known as overfitting. Two pruning methods were employed in this study; the first is known as subtree replacement in which a node is turned into a leaf, effectively reducing the number of rules down a path, and the second is called subtree raising in which a node is moved closer to the root node (the first node to be created) by replacing other nodes above it. The two methods use either error rates and a portion of withheld training data or a calculation of the natural variance in the data to determine how much pruning should take place. Confidence thresholds can be set for the latter measure. Other pruning techniques involve specifying the minimum number of instances allowed per leaf and forcing a binary split (based on an inequality) of numeric data at the nodes effectively transforming the data into nominal values.

9.3.2 *Method*

As previously described, the data mining software called Weka (version 3-6-2) was used for all calculations. The subjects' morphological data and perceptual judgements of all HRTFs in the database were formatted for the decision tree generation. Each instance used for the analysis corresponded to one subject's judgement of another subject's HRTFs from the listening test results. This allowed for a total of 1,035 instances, which equated to the number of pairs of subjects that can be selected from the 46 subjects in the database. In order to represent the morphological data, and the corresponding pairwise judgements, the morphology of the first subject had to be calculated relative to the other subject. This was achieved by subtracting the morphology values of the subject's HRTFs that were judged from the morphology values of the subject making the judgement. One instance therefore represented a subject's set of relative morphology values, corresponding to the different parameters used, along with the subject's perceptual judgement from the listening test.

Given that decision trees can more readily incorporate multiple classes to classify, a variety of judgement grouping methods were used. This included groupings of HRTFs based on whether they were judged as *bad* or what we might call *not-bad*, which included HRTFs judged as either *fair* or *excellent* combined as one group. This allowed for an almost even number of instances in each of the two groups; there was a total of 825 and 1,200 instances in the *bad* and *not-bad* groups respectively. Similarly, groups of HRTFs judged as *excellent* and *not-excellent* were created, in which responses *fair* and *bad* were grouped together and named *not-excellent*. This grouping of *excellent* and *not-excellent* however lead to a disproportionately large number of instances in the *not-excellent* group given that only a small number of HRTFs were ever judged as *excellent*. The original judgement classes *excellent*, *fair*, and *bad* were also tested.

Given that it is likely that there would be a large degree of variance, or noise, in the subject responses (see section 8.4 of the previous chapter) from the listening test, the decision trees should tolerate incorrectly classified instances and allow for a generalised set of rules (i.e. a larger than normal degree of pruning). This was achieved by decreasing the confidence threshold for subtree replacement pruning to values lower than the standard 25%. The minimum number of instances allowed at a leaf was also set to approximately 10% of the total number of instances to be classified.

Despite the intuitive nature of the algorithm, a key drawback to using decision trees is their inability to distinguish correlated or irrelevant attributes (see Perner, 2001). Since the algorithm is recursive, an attribute that was not significant in terms of information gain when nodes were being created high up in the tree using a large proportion of the instances may become significant for creating rules lower down in the tree when using only a small portion of the instances to classify. This leads to attributes being used in the decision tree that have little to do with the effective classification of the data in a more general sense. To avoid this, the attributes, or morphological parameters, that are fed into the algorithm need to be relevant and have some sort of feature selection applied to them. The obvious choice for a subset of morphological parameters is that calculated in section 9.2 using the regression and subject specific MSs along with the results from the listening test. This feature selection method is preferred over others such as support vector machines (see section 9.3.4) as it does not require an arbitrary cutoff of the number of parameters to be used; Support Vector Machines (SVMs) only provided a ranking of the most significant parameters to be used rather than a subset.

Using this subset of morphological parameters, a decision tree model is built based on the instances in what is termed the training data. However this model will be using all instances as training and there is no way to test how this predictive model would perform on a new set of instances. In order to effectively assess the decision tree, a technique known as cross-validation was used. Instances are broken up randomly and evenly into a certain number of groups; 10 groups were used in this study, known as a 10-fold cross-validation, which is the standard. To validate the method of using a decision tree to classify instances, one of the 10 groups is held back and not used in the training of the model. Once a decision tree is built using the parameters described, the remaining instances that were held back are then used in order to assess how successfully they can be classified. This is performed 10 times in which each group is held back and used to assess the predictive power of the created decision tree.

It is important to note that a cross-validation is not used to find the optimal model as for each fold a different set of instances is used and this can lead to a different set of rules being defined. The cross-validation is used merely to test the quality of the model. The final model uses all the instances to create a decision tree once the effectiveness of classifying instances has been assessed using a cross-validation.

9.3.3 *Results*

The final decision tree that was generated used a minimum number of instances allowed at a leaf of 100 (approximately 5% of the 2,025 instances in the data set) and confidence threshold values below 25%. The minimum number of instances at a leaf ensures that rules are not generated based on a very small sample of data, and that there are not an excessive number of nodes. The confidence threshold turned out to not have a bearing on the effectiveness of the decision tree. The different grouping methods were tested and it was shown that only when judgements were split into the *bad* and *not-bad* categories (with a somewhat even split in the number of instances) were the results valid. Any decision tree that used the judgement category *excellent* inevitably incorrectly classified all these instances into the *not-excellent* group with little cost to the overall effectiveness of the decision tree given that there was a small number of instances with the *excellent* judgement.

The final decision tree, created using the *excellent* and *not-excellent* classification, was a simple split on the parameter d_6, pinna width, at a value of 7 mm, and had an overall percentage of correctly classified instances of approximately 61%. This classification of subject DTFs, based on the difference in the measured morphological parameters between each pair of subjects, is obviously an oversimplification and a poor classification given that it is only slightly better than chance. The cross-validation, which is used to test the predictive power of the decision tree, showed approximately the same percentage of correctly classified instances, suggesting that the model was not overfitting the data (which is clearly not the case given that only one node exists in the decision tree). It is interesting to note that the only parameter used in this global decision tree of the subset of eight selected from the previous section, using the F-ratio loss function, was pinna width (d_6), which was ranked as the most significant morphological parameter in the current study.

Given that the global decision tree used to classify all instances performed poorly, it was of interest to test the selected morphological parameters to create individual decision trees (from section 9.2) on a subject by subject basis. To achieve this, the analysis using decision trees was run separately for each subject in the database, and only instances for which a particular subject judged the other subjects (from the listening test in chapter 7) were used in the classification. In this manner, a unique decision tree was created for each subject that took part in the listening test. Figure 50 shows an example of one of the decision trees created from this analysis. The values that are at the nodes show how instances are split along the different levels of the tree. The number of classified instances are shown in the rectangular boxes with the number of correctly and incorrectly classified instances on the left and right respectively. Fractional values are shown due to how the C4.5 algorithm accounts for missing values for a small number of morphological parameters. The confidence threshold was kept at 25% and the minimum number of instances at each node was reduced to 2, which equated to approximately 5% of all instances; this was the same percentage used when classifying instances of all subjects together as previously described.

The analysis showed that for approximately 70% of subjects a decision tree was generated from the eight optimal parameters. For the subjects that did not have a decision tree generated, the algorithm was unable to create any rule set based on the morphological parameters given as input that could effectively predict the subject's responses from the listening test. It is possible that these subjects were what could be described as *naïve assessors* (see table 5 from chapter 6) and that their responses from the listening were not

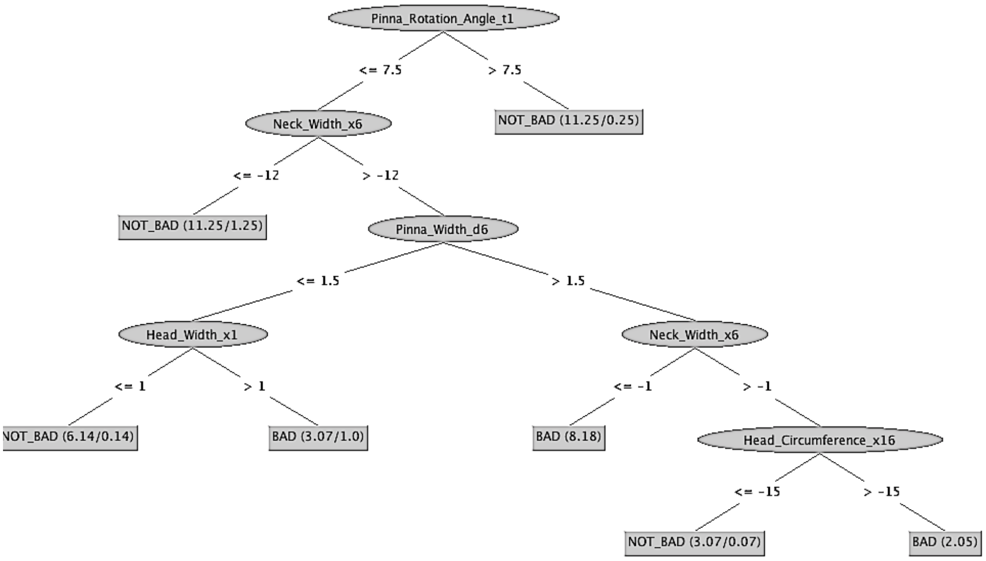


Figure 50: Example decision tree created for subject 1032 in the database. Values on the branches show how instances are split based on a particular parameter at the node. The number of classified instances are shown in the rectangular boxes with the number of correctly/incorrectly classified instances indicated.

reliable. The study from chapter 6 showed that subjects with little expertise had a large degree in variability in their responses, albeit for a slightly different design of listening test used in the current analysis, as explained in section 8.4. Yet, given that the level of expertise of the subjects was not available, it is difficult to know whether this result was due to subject expertise or other factors such as limitations in the algorithm itself for this particular data set.

Of the subjects that had a decision tree created for their responses, the results showed varying degrees of effectiveness. Figure 51 shows the performance of the individual decision trees for each subject in the database (for 13 subjects a decision tree was not generated, and these subjects were not shown in the figure), represented as the percentage of correctly classified HRTFs. The number of decision trees that performed better than chance (marked in green) was 24 out of 32, or 75%. The average percentage of correctly classified HRTFs across all subjects (only including subjects for which a decision tree was created) was approximately 60%, with a minimum and maximum of approximately 38% and 82% respectively. These subject specific decision trees can help to highlight significant morphological parameters but are not predictive in the sense that there is no global decision tree that can be used for a subject that did not participate in the listening test. At best, some form of mapping might be used based on the measured morphological data, as shown in the previous chapter using principal component weights from a PCA to map subjects into a MS, in order group subjects coming from outside the database into a subset of decision trees. This type of mapping however was shown in section 8.2.3.6 to not be effective in describing the perceptual judgements from the listening test. As previously discussed, the power of

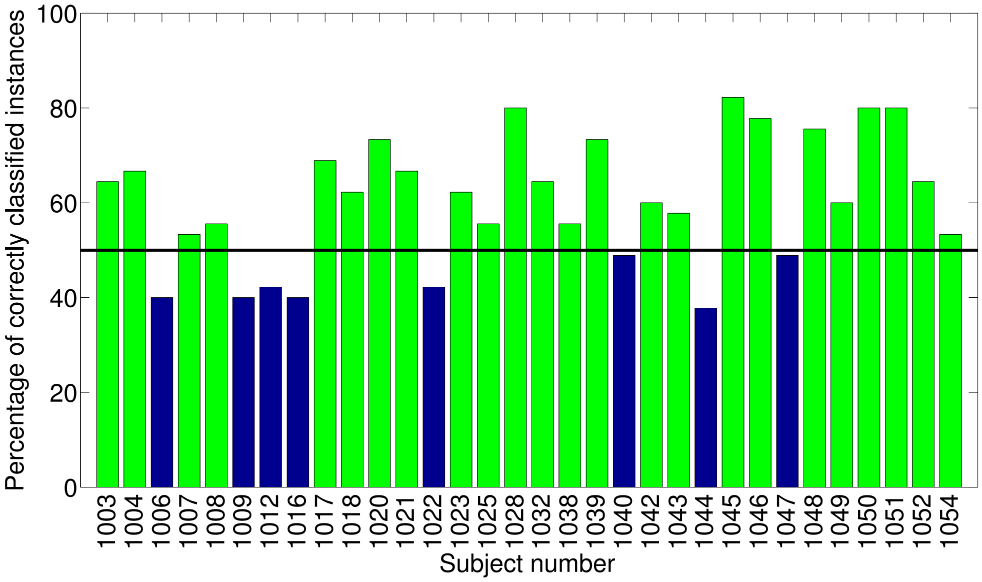


Figure 51: Percentage of correctly classified instances for each of the subject-specific decision trees generated. Subjects for which no decision tree was generated are not displayed. Decision trees that performed above chance (black line) are coloured green.

the analyses presented in this chapter was to draw out significant morphological parameters rather than present a model that had an high predictive ability given the observed variance in subject responses in the listening test (see section 8.4).

Figure 52 offers a more detailed breakdown of the generated decision trees showing the number of times the eight optimal parameters, taken from the forward sequential feature selection described in section 9.2.1, were used across all the subject-specific decision trees. The figure shows the different levels of the decision trees, with the first node corresponding to a value of one. The size and colour of the circles reflect a count for each parameter at a specific level in the tree. The results show that if the count is summed across all levels in the decision tree, the optimal morphological parameters can be ordered from most to least effective: d_6, x_2, x_6, x_1, d_7, θ_1, x_{16}, and x_{12}. Noteworthy is the fact that the parameter d_6 (pinna width) was shown to be the most significant in this decision tree analysis, chosen as the only parameter for the global decision tree analysis, and ranked as the most significant for the analysis using the MSs in section 9.2.2.

9.3.4 *Support vector machines*

Further to the selection of the eight optimal parameters in section 9.2 using a recursive forward feature selection, the following section utilised a different machine learning algorithm called SVMs to rank all morphological parameters. The main purpose of this analysis was to use the ranking as a comparison to the recursive feature selection in an effort to validate the chosen parameters. It was also of interest to determine if this classification

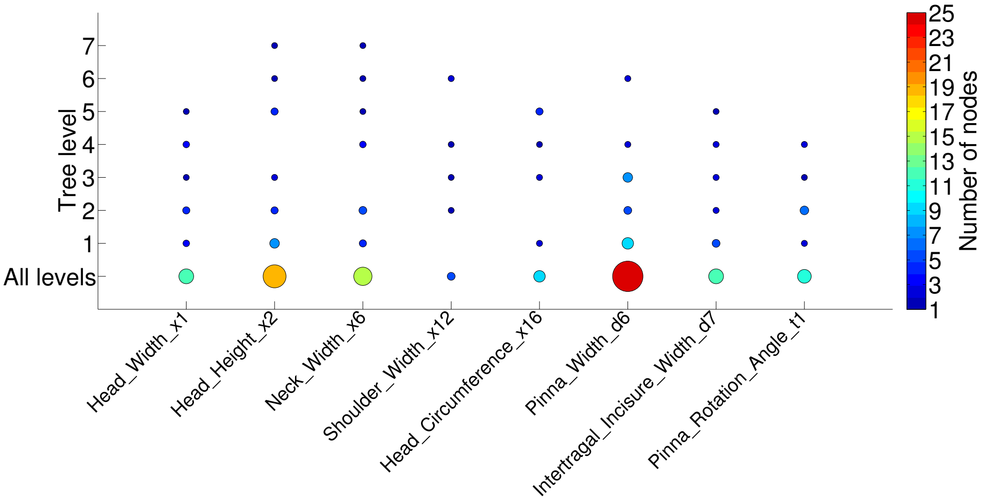


Figure 52: The number of times each of the eight optimal morphological parameters appeared in the subject-specific decision trees across all subjects at each node level, along with the total count. The tally for each parameter is represented by the size of circles and their corresponding colour.

technique could be used to predict subject responses from the listening test and act as a global classification tool similar to the global decision tree that was generated in the previous section.

SVMs are a state-of-art algorithm used for classification (Boser et al., 1992; Vapnik, 1998), and as explained previously, they use instances in the data to train a predictive model, which can then categorise any instance based on input for a set of variables such as morphological parameters. SVMs have been shown to be an effective tool for discovering patterns in data (Guyon et al., 1996) by representing instances, such as judgements by subjects of HRTFs in the database, as points in a MS. The instances are mapped so that the corresponding categories assigned to each instance (i.e. judgements of either *excellent*, *fair*, or *bad*) are divided by a clear gap that is as wide as possible. More specifically, SVMs classify instances by mapping them to a high- or infinite-dimensional space in which hyperplanes are used to divide the space according to the categories used (in most cases only two categories). The higher dimensionality of the space makes the instances more linearly separable. The feature of SVMs that makes them particularly attractive to this study is that it sets the chosen hyperplane or hyperplanes that allow for the largest margins between the mapped points. In addition, if a set of points cannot be perfectly split according to their categories, a *soft margin* method is used to create a clean split whilst allowing for a certain degree of mislabeled or incorrectly categorised points, while still maximizing the distance to the nearest cleanly split instances. This is particularly relevant to the results in the current study given the assumed noise in the listening test results (see section 8.4).

9.3.4.1 *Method*

Again, the data mining software called Weka (version 3-6-2) was used for all calculations. The subjects' morphological data and perceptual judgements of HRTFs in the database were formatted for an analysis using SVMs in the same manner as for the analysis using decision trees. One instance represented a subject's set of relative morphology values, corresponding to the different parameters used, along with the subject's perceptual judgement from the listening test. SVMs usually work by separating a set of instances into two classes, therefore the judgement responses were combined to form a group of responses *bad* comprising of HRTFs judged as *bad*, and a group of *not-bad* as for the previous section. This grouping provided an almost even split into the two groups across all instances.

The full set of instances were used as a vehicle, as in the previous section, to determine the significance of the morphological parameters with respect to the listening test results via SVMs. Different combinations of morphological parameters were systematically tried and the resulting effectiveness of splitting the instances based on subject judgement was used as a criterion for the selection of the different subsets of parameters. More specifically, morphological parameters were recursively eliminated, removing the parameter with the smallest positive effect on the classification of the instances as done in section 9.2 using the F-ratio loss function rather than classification accuracy as in the current analysis.

This recursive feature elimination, which is a backward elimination method, was then used with linear SVMs to train the instances using all the parameters and iteratively eliminate features based on the classification. This technique has been used and benchmarked against a variety of techniques, for example in analyses to find cancer genes (Guyon et al., 2002). With this method, the parameters were ranked from most to least significant.

9.3.5 *Results*

A variable in the classification algorithm when using SVMs is the complexity value C, which controls the tolerance of classifications errors in the calculation, and introduces a penalty function (Gunn, 1997). A C value of zero indicates that the penalty function is not taken into account, and a large value of C (C towards infinity) indicates that the penalty function is dominant. According to the different values of C, the ranking of the morphological parameters and the effectiveness of the classification varied. A value of C = 100 was used for classification with SVMs, selected to have a somewhat high tolerance for classification errors (i.e. reduce overfitting) and gave the best results using a 10-fold cross validation classification. Non-linear SVMs, using for example the radial basis function kernel, allow for a mapping into a higher dimensional space and are usually recommended for the type of data used in this study (i.e. data with a small number of features relative to the number of instances). However, a C value of 100 using linear SVMs produced classification results equivalent to using non-linear SVMs. The most significant morphological parameters, from most significant to least significant, as ranked using SVMs, were: x_2, x_3, x_{10}, x_{11}, x_{12}, x_6, x_9, x_1, x_5, d_1, x_7, x_{17}, θ_1, d_4, d_6, x_4, d_7, x_{16}, x_8, d_3, d_2, and d_5. For an effective comparison of the ranking of parameters using SVMs and the selection of eight optimal parameters using the F-ratio loss function detailed in section 9.2, only the first eight parameters were selected. Of these top eight parameters, four were found to be in common between the two methods; these were: x_2, x_{12}, x_6, x_1, corresponding to head height, shoulder width, neck width, and head width.

The percentage of correctly classified instances based on the judgement groups *bad* and *not-bad* using SVMs was shown to be approximately 62%, which is only slightly better than the accuracy of the global decision tree from the previous section. The results suggest that SVMs could not be validated as an effective predictive tool in this case, which may or may not be due to noise in the database; the significant result is the ranking of morphological parameters and how this compares to previous methods.

9.4 MODIFICATION OF DUMMY HEAD PINNAE

The previous sections have sought to highlight the most significant morphological parameters using the perceptual judgements from a listening test. The following section describes an analysis that aimed to draw out insights relating to the different morphological parameters by actually physically modifying them, on a mannequin or dummy head, and observing the impact of these modifications on HRTF measurements. The concept behind the dummy head measurements is to allow for controlled variations of the different morphological parameters or pinna structures and to observe the effect on the HRTF. More specifically, an analysis of the frequency magnitude changes in the HRTFs for each specific modification can help distinguish the role of a particular morphological parameter in terms of how listeners perceive their acoustic environment. The analysis also aimed at assessing how the different morphology modifications were distinct from each other, in terms of their influence on the HRTF, by comparing the changes in HRTF magnitude between a variety of different modifications.

9.4.1 *Method*

The recordings of the dummy head HRTFs were not performed by the author and were part of an previously mentioned LISTEN project at the IRCAM research centre. A full pinna mould was made from subject 1034 in the LISTEN database for use in the analysis. The pinna replica cast was made in hard rubber, mounted in a cylindrical form, diameter 7 cm, centred on the ear canal of the dummy head. This design ensured that the microphone position was the same for all moulds, and that rotation of the pinna did not alter the position of the microphone.

The dummy head was constructed from a Styrofoam fashion mannequin and corresponded to average male head size. The position of the ear canals

ID	LEFT EAR MODIFICATION	RIGHT EAR MODIFICATION
1500	No modification	No modification
1501	Reduced concha height	Reduced concha width
1502	Reduced notch to 1.5mm	Concha filled completely
1503	Filled behind ear	Filled details except concha
1504	Notch closed	Reduced concha depth 5-7mm
1505	Enlarged pinna	Reduced concha height (same as 1501 left)
1523	Concha only (pinna cut away)	N/A

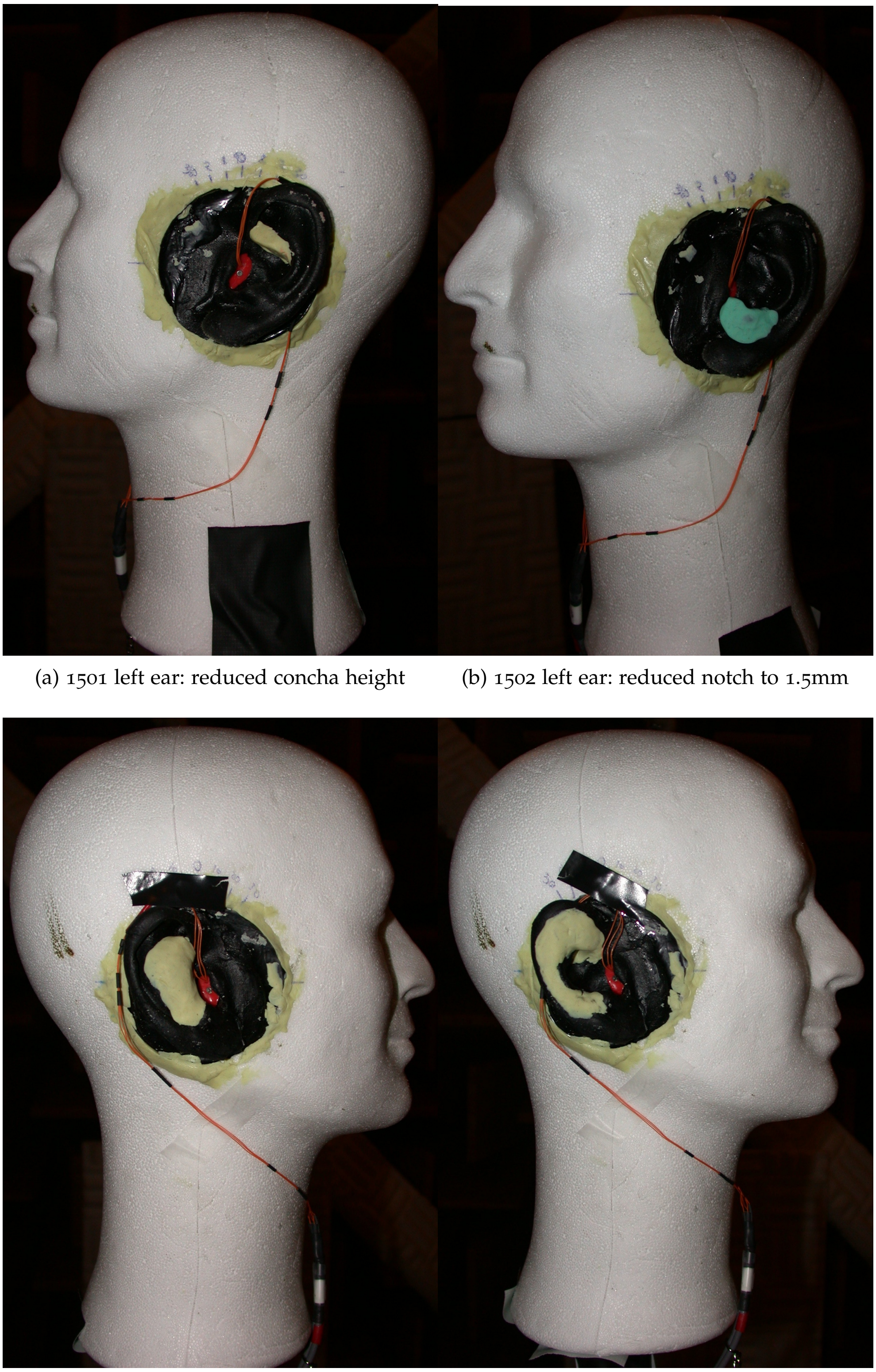
Table 9: List of the different modifications for each recording. The left and right ears were independently modified for a total of eight HRTF recording sessions, giving a total of 13 different modifications.

was determined based on average values for the various parameters in the CIPIC database for male and female subjects. The head was altered to allow for placement of the pinna replicas. Modeling clay was used to ensure a smooth seam between the pinna and head. Figure 53 shows a few examples of the different modifications made to the pinna of the dummy head. Each modification is labelled in the figure by giving a number to each of the recording sessions (e.g. 1501) and specifying whether it was the left or right pinna being modified.

Using this modification approach, it was possible to change certain parameters independently by either adding to or removing material from the replica. Material was added through the use of modeling clay. Removal of material was accomplished by simply cutting the material. As several replicas could be made from the same pinna mould, various configurations could be tested.

In contrast to live subject measurements, which were measured every 15° in the LISTEN database, the dummy head measurements (not performed by the author of this body of work) were made every 5°. This higher degree of resolution resulted in a measurement time of approximately four hours for a complete spatial set. The exact same measurement setup was used as for the recording of the subject HRTFs in the LISTEN database (see section 6.2.3). Table 9 displays the different modifications made to the dummy head for this analysis. A total of seven recording sessions were performed, using a different modification for the left and right ear. From these seven recordings, there was one in which only the left ear was used giving a total of 13 modifications. Not all the modifications were unique however; one of the modifications (1505 right ear) was replicating a previous one (1501 left ear) and was used in the study to validate the analysis.

The HRTFs for each modification were processed much in the same way that the LISTEN database HRTFs were in the previous chapter. The recorded impulse responses, or Head-Related Impulse Responses (HRIRs), were normalised in root-mean-square first across the left and right ears and then across all recording sessions for both the left and right ear combined. The HRIRs for the right ear modifications were then reordered so that record-



(a) 1501 left ear: reduced concha height

(b) 1502 left ear: reduced notch to 1.5mm

(c) 1502 right ear: concha filled

(d) 1503 right ear: filled details except concha

Figure 53: Examples of the different pinna modifications on the dummy head. Each modification is assigned a number referring to the specific HRTF recording and whether it was the left or right ear.

ing locations were mirrored across the midsagittal plane; for example all HRIRs that were recorded at locations with azimuth 30° were switched with the recordings at locations with azimuth 330° (angle of elevation was unchanged). The purpose of this reordering was to be able to compare left and right ear HRIRs for any given location in space by effectively changing all right ear HRIRs so that it was as if they were recorded at the left ear. The next step was to convert all HRIRs to the frequency domain and subtract out the mean magnitude across all positions in space from each HRTF. This is analogous to the DTFs described in previous chapters (see section 8.2.3.1), however a spatially weighted mean magnitude was not used.

The DTFs for the recording session 1500 (without any modifications) were then subtracted from all DTFs that had had the pinna modified using the corresponding location in space and corresponding ear to create 11 different modifications (i.e. the two conditions with no modifications were no longer included). This procedure effectively took the difference in the DTF magnitude between the unmodified pinna and corresponding modified pinna, allowing for a representation of how a particular modification changed the filtering effects of the pinna relative to the baseline condition. The 512 point DTFs were then all reduced to the optimal frequency range calculated in the previous chapter (see section 8.3.3), and then concatenated across all 2,016 locations in space for each modification, generating a single vector of magnitude values. The DTF vectors for each modification were then combined to occupy a single row in a matrix of all concatenated DTFs. The columns in this matrix thus represented the frequencies for magnitude values in the DTFs. Finally a PCA was performed on this matrix in order to produce a set of principal components, or eigenvectors, and their corresponding weights (see section 8.2.3.1 for a more detailed explanation). The weights from the PCA were used to map the 11 different modifications into MSs, in the same manner as described in the previous chapter, in order to visualise how each one compared in terms of how they influenced the HRTF.

9.4.2 *Results*

Figure 54(a) and 54(b) show the different dummy head modifications plotted using the first and second, and second and third PCs as coordinates in a two-dimensional space. The representation of the modifications in a two-dimensional space is valid given that almost 70% of the variance in the data is explained in the first three PCs. Figure 54(a) shows a clear separation of the modifications that used the left or right ear; all modifications that performed on the left ear are on the left of the plot and all modifications performed on the right ear are on the right. The first PC is describing the global differences between HRTFs recorded at either ear (i.e. all differences that are common across all modifications for a particular ear), and in this sense can be used as a form of normalisation, with all successive PCs representing differences among the HRTFs relating to the modifications at both the left and right ear. Inspection of the control left and right ear HRTFs showed a gain difference for positions of high elevation for the left ear. This is despite the fact that the left and right ear HRIRs were normalised in root-mean-square. A global difference of this nature could be due to an overall difference in the location of the left and right ear socket relative to the speakers.

Figure 54(b) shows the modifications without representing the first PC and only using the second and third. The validity of this representation, which included the removal of the first PC in order to gain insights into how the modifications compared to each other, is confirmed by comparing the position of the *reduced concha height* modification in the two-dimensional space, which was performed on both the left and right ear independently. These two modifications are relatively close to each other despite the fact that the first PC has been removed from the representation, which is the expected result as these modifications should have resulted in similar variations in the HRTF spectra.

In addition to this finding, there are other pairs of modifications that lie close to each other within this representation that are similar in the manner in which they occluded or exaggerated certain features of the pinna. For example, the modifications *reduced notch to 1.5mm (left)* and *notch closed*, both resulting from an occlusion of the pinna notch (see figure 53), are in proximity of one another using these two PCs. A similar finding can be seen for the modifications *filled details except concha (right)* and *concha only - pinna cut away (left)*.

Of particular importance is how this representation of the different modifications in figure 54(b) using the second and third PCs shows how certain aspects of the pinna are unique in terms of the measured HRTFs. In summary, the results suggest that modifications that reduced the concha width were distant from those that reduced its height, and reducing the concha depth is somewhere between these two modifications in terms of its location in this two-dimensional space.

9.5 DISCUSSION

The different analyses performed as part of the current study have highlighted some insights into the significance of different morphological parameters. The techniques used were able to explore a selection of parameters and test a model in terms of its predictive power. The first analysis, presented a methodology for selecting an optimal set of HRTFs for a listener based on a subset of key morphological parameters. The process described a prediction technique using morphological parameters to estimate subjects' positions in a MS. The results showed that using a predictive approach was almost as effective, in terms of the statistical significance of the model, as using the subjects' calculated positions that made use of their own HRTF data. This means that the method used to predict a subject's position based on their morphology (using a regression) was almost as accurate as using the subject's HRTFs to calculate their true location in the MS. In short, the predictive method, via its use of a PCA of subject DTFs to calculate inter-subject differences, was shown to go some way towards describing the perceptual judgements made by the subjects in terms of the Euclidean distance between them

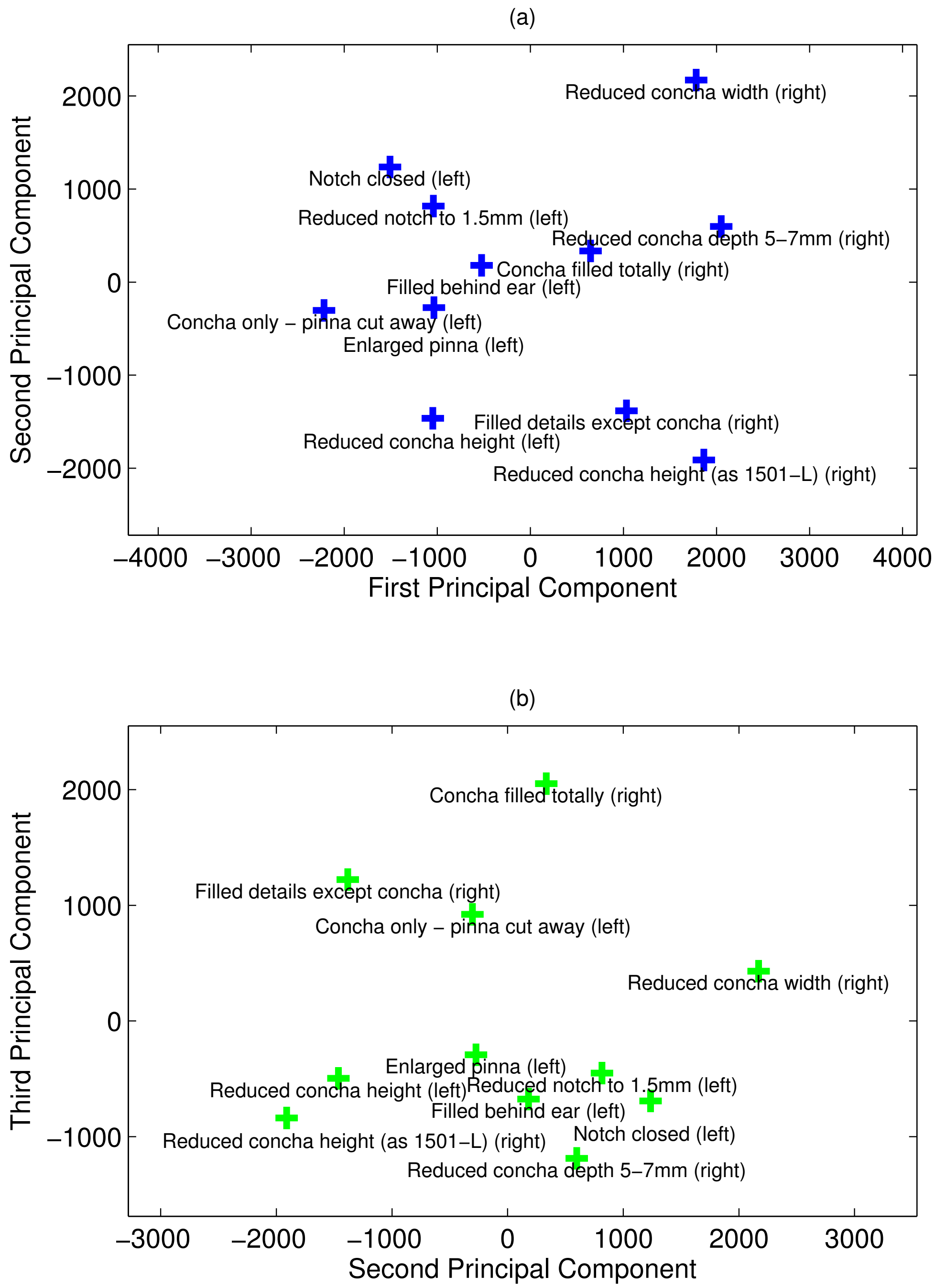


Figure 54: Each modification of the dummy head pinna displayed as a function of (a) the first and second, (b) the second and third PC weights from a PCA across all modifications.

in the generated MS. Given the statistical significance of the result, measured by the F-ratio, the calculated optimal selection of morphological parameters is considered to reflect the influence that they might have on the HRTF in terms of the quality of the binaural synthesis tested. This selection is obviously limited to the set of measured parameters in the database, which is not exhaustive. The power of the proposed method was that it was able to incorporate a large set of subject HRTFs from the chosen database in order to build a model for the selection of morphological parameters. In addition, the procedure used a large number of subjects and perceptual judgements so that variability or noise within the data set, in the form of inaccurate responses from subjects in the listening test (see chapter 7) and imperfect morphological or HRTF measurements, would wash out in the overall statistical result.

In contrast, the second and third analyses in the current study did not use any subject HRTF information. This can be advantageous for a predictive model since it avoids the laborious task of performing HRTF recordings for a large number of subjects. However, both machine learning algorithms, using decision trees and SVMs, were shown to be only slightly better than chance in terms of their ability to predict the perceptual judgements from the listening test. In the case of the decision tree analysis, only one morphological parameter was selected for classification of instances across all subjects (i.e. a global decision tree). The results suggest that the use of only subject morphology to classify responses (i.e. not using HRTF data) is not sufficient for generating a global predictive model with any meaning. This is not a surprising result given that attempts in the previous chapter (see section 8.2.3.6) to represent inter-subject differences using only morphological parameters via a PCA was shown to be ineffective in terms of describing the listening test results. It follows from this that the conclusions drawn from the two analyses using machine learning algorithms in the current study, with respect to the significance of the different morphological parameters, is difficult to interpret using the current database. Future work, using a large number of responses, which have been tested in terms of their reproducibility, may be able to use the mentioned machine learning techniques to distinguish more robust patterns in the data.

The analysis using pinna modifications on a dummy head does not directly relate to the previous sections in this chapter as they involve three-dimensional variations of the pinna using clay, and previous work used two-dimensional morphological parameters. That is to say, a modification on the dummy head may relate to a combination of CIPIC parameters making it difficult to compare results. Nevertheless, the representation of the different modifications as a function of their corresponding PC weights was able to show how they might vary from one another in terms of the measured HRTFs. This insight into the uniqueness of different controlled modifications of a dummy head pinna can be used to show the relative importance of different morphological parameters if each modification can be correlated with different regions or features in the spectra of the HRTFs that have been shown to be perceptually significant, such as those detailed in the previous chapter. Future work might seek to isolate these regions or features and analyse how they vary as a function of small changes to the parts of the pinna that were shown to be unique in the current analysis. A particularly elegant and efficient way to achieve this would be to generate a polygon mesh of the human head and pinna using three-dimensional computer graphics and have it vary along a set of key morphological parameters that have been shown to be significant and unique in the current study. This would allow for much more precise, somewhat orthogonal, variations in the shape of the head and pinna compared to the modifications outlined in section 9.4 using clay. For each of the small orthogonal head or pinna variations, a numerical acoustic simulation could be performed such as those computed in studies using the boundary element method (Katz, 2001; Kahana and Nelson, 2007). An analysis of the computed HRTFs for a series of pinna modifications would be a powerful tool for gaining insights into the link between morphology and the perceptually relevant components of the HRTF.

9.6 CONCLUSION

The current study aimed to address the question of which morphological parameters are significant in terms of their influence on the HRTF for binaural synthesis. The main finding was the selection of a subset of measured parameters via an analysis using the compressed format of the HRTF information from a large database via a PCA. The results suggested that the morphological parameters selected could be significant in terms of their perceptual effect on the HRTF, such as the pinna width, or that they are good predictors of, or proxies to, other morphological parameters that were not measured and that have a perceptual effect on the HRTF, such as neck width and shoulder width. In either case, the process used to generate the subset of significant morphological parameters could be applied on a commercial scale for the selection of optimal HRTFs for a listener using a small selection of morphological measurements. These morphological measurements could possibly be derived from a photo or scan. Further analysis using listening test results with a high degree of repeatability in subject responses is recommended in order to truly test the effectiveness of such a method. Other machine learning techniques that only used morphology data and not HRTF data were less successful in classifying perceptual responses. Some initial insights into the uniqueness of the different parts of the pinna were established using modifications on a dummy head and recorded HRTFs.

With respect to the predictive model, the focus of the analysis was not on how many HRTFs could be accurately selected for each subject in the database, given the observed variability in perceptual judgements of HRTFs shown by Schönstein and Katz (2011), but rather on the selected morphological parameters and the degree of statistical significance obtained using the described procedure. To this end, the proposed selection method was shown to be effective, with a high level of statistical significance ($p < 0.001$), and an F-ratio value of 34 compared to 84 for the model that used subject HRTFs instead of morphological parameters. If morphological parameters can be used instead of subject HRTFs then there is no need for the laborious and expensive task of recording HRTFs in order to provide a personalised rendering in virtual auditory space. This has implications for the adoption of the technology in consumer markets as the individualisation of HRTFs remains one of the main barriers to date. Future work should include another listening test, with perceptual judgements shown to have a high degree of reproducibility as in chapter 7, in order to further validate this HRTF prediction method.

10 GENERAL CONCLUSION

The studies presented in this body of work have aimed to highlight some of the most significant issues relating to the individualisation or personalisation of Head-Related Transfer Functions (HRTFs). The findings have been presented in the context of the many applications of binaural synthesis in the consumer market, from spatial hearing aids to immersive virtual environments. The work centred on understanding the role of spectral cues and how to use this knowledge in order to develop criteria for the selection of an optimal HRTF for the listener. Due to the complexity of how the auditory system interprets sounds in our environment, the task required systematic and varying approaches. Two methods were used to quantify the effectiveness of a binaural synthesis: localisation tasks and listening tests. Localisation tasks were used in the context of a study of different headphone types and headphone equalisation. Listening tests were used in studies that aimed to determine the variance in perceptual judgements of HRTFs, the most salient spectral cues, and the most salient morphological features of the listener. The listening tests were a vehicle for generating a process that was able to select an optimal HRTF from a database for a listener based on a few key morphological parameters. The impetus behind such a process was to avoid the need for recording HRTFs on the listener, which is an expensive and laborious task that requires specialised equipment and expertise.

10.1 FINDINGS FROM THE RESEARCH

The study of the different headphone types in chapter 6 did indeed show that the hardware used in a binaural synthesis is of importance and that there are significant variations in localisation accuracy for the different types, depending on whether the headphone was research grade, open or closed circumaural (i.e. covering the ear without contact with the ear), intra-aural, or bone conduction. It is recommended, based on the localisation accuracy results, that open circumaural or research grade headphones be used for the most effective rendering of sound sources in Virtual Auditory Space (VAS). The type of headphone equalisation used in this study did not result in a global improvement across all headphones.

The following studies in chapters 7, 8, and 9 then moved away from localisation accuracy as a criterion for an effective binaural synthesis to using listening tests. This was done in order to concentrate on the aspects of a rendering in VAS that might seem more relevant to the consumer market; small localisation errors were deemed less significant when compared to whether a virtual sound source was perceived as being coherent, correctly perceived in front of the listener and not inside the head, or poorly externalised, for example.

In the process of determining the best listening test format for the study in chapter 7, the question of whether subjects could in fact make reliable judgements in this context was assessed. The repeatability of perceptual judgements of HRTFs is crucial for their use in any type of evaluation of an HRTF selection procedure, particularly as the assessment of the quality of different HRTFs using listening tests is a challenging task. The results from the many different versions of the presented listening test showed that there was indeed a significant degree of variation across replicates for some subjects, particularly when the number of HRTFs being judged was higher than approximately six. Furthermore, the expertise of the subject was shown to play a role in response repeatability, which suggests that naïve listeners might not be able to make reliable perceptual judgements of HRTFs at all.

The iterative approach to developing a listening test lead to a reduction in the variability in subject responses. Further validation of the final version of the listening test, using a larger number of subjects in each of the assessor categories, could be performed in order to bolster the correlation between expertise and variance in subject responses, and in turn show that response variability can be reduced. Regardless of the observed effect of subject expertise, the study as a whole was significant given that response repeatability had been seldom analysed in the literature for judgements of different HRTFs. Given that the listening test task was shown to be difficult even for subjects with extensive experience judging audio quality, the research was seen as an important piece of work in the field. In addition, many studies in the past have based their conclusions regarding the effectiveness of a particular HRTF customisation technique on listening test results without measuring repeatability. In this context, the study has contributed to an understanding of the degree of variance to expect in subject responses, and the listening test design offers a method for evaluating whether responses from a particular subject are reliable or not.

Chapter 8 presented a study of inter-subject differences in HRTFs using knowledge of response variability from chapter 7. Based on the listening test results it can be assumed that there would be a number of responses that would change if the subjects were to repeat the experiment. It is for this reason that the judgements could not be used in isolation, and that any analysis would need to deal with a significant amount noise in the results. To this end, a robust statistical procedure, with a large number of subjects, was used in order to assess the most salient spectral cues in terms of the perceptual judgements from the listening test. By performing an analysis on all pairwise judgements between the 46 subjects, and tying the results to the HRTF data, which contains very little noise, the analysis was able to harness subtle patterns.

A principal component analysis, using concatenated HRTF data within a range of 3.25 to 14.25 kHz, was shown to best model the perceptual differences between subjects. Other techniques for describing inter-subject differences, such as the method using the frequency scale factor, were relatively effective and highlighted the role of covert features in the HRTF. It was found that covert features, which are features that require HRTFs from all locations in space to be considered in the analysis, were generally more important than overt features, such as notch frequencies.

The final study in chapter 9 used a feature selection to generate a subset of the most significant morphological parameters for all subjects in the database. The results showed that a small subset of parameters could be used to produce the most effective model, and that dimensions such as pinna width were particularly significant. The subset of morphological parameters were then used to predict optimal HRTFs from the database for the subjects. The procedure was effective within the limits of the known variance in the perceptual judgements from chapter 7. Most notably, the predictive model, using only subject morphology, was able to perform almost as well as the model using the subjects' own HRTF data, in terms of describing the listening test results. Future work using a large number of responses from subjects showing a high degree of repeatability could be used to further assess the effectiveness of this prediction method. If the procedures detailed in chapter 7 were to be used and the repeatability of subject responses tested, then the accuracy of the HRTF prediction technique could be more accurately assessed. This would be a significant result and one of the main objectives of this body of work; that is to be able to select an HRTF from a database without the need to perform any recordings.

Future work should also further develop the link between morphology and the HRTF. The CIPIC morphological parameters, whilst descriptive of the listener's ear and body sizes, were not chosen based on results and evidence from an experiment. An analysis of the many ways in which the human ear can vary, and the influence that these variations have on the HRTF, possibly via the use of numerical acoustic simulations using the boundary element method, should be used to build up knowledge on the importance of key dimensions. A better knowledge of what aspects of the human ear and body are most significant would feed into the HRTF selection method detailed in chapter 9, and improve its effectiveness.

Taken together, the studies have presented compelling methods for quantifying the quality of a binaural synthesis and progressing the field. The main goal of producing a predictive HRTF selection tool, in order to bypass the process of recording individualised HRTFs, was achieved to a level of statistical significance. Possibly the most challenging aim of this body of work was to understand further the complexity of how the auditory system uses the HRTF in order to extract cues to sound source location, and which parts of the listener's morphology play a significant role. To this end, some key achievements were made, in particular the ability to compare a number of the most studied features of the HRTF using the same validation metric. This allowed for an understanding of the relative significance of these spectral features, and added to studies highlighting the importance over covert features of overt features, such as spectral notches. The subset of morphological parameters chosen for the selection procedure could potentially highlight their influence on the HRTF. Despite the fact that some commonalities in the selected parameters existed between the different analyses from the last chapter, and that the similarity of the different morphological parameters was explored using the dummy head recordings, the significance of the chosen parameters was not as important as the performance of the HRTF selection procedure on the whole.

10.2 POTENTIAL APPLICATIONS OF RESEARCH

Findings pertaining to the role of headphones in binaural synthesis have implications for the type of headphones that should and should not be used when applying the technology to the consumer market.

The localisation accuracy results suggested that the tube insert headphones that were not research grade (i.e. those used by security guards and the like) lead to the highest degree of errors, both in terms of global accuracy and front-back confusions. This was most definitely due to the fact that the headphones did not produce a signal in the high frequency range. The other intra-aural headphone (i.e. not the tube intra-aural headphones), was shown to be effective when compared to the accuracy of the reference research headphones. The implication of this finding is significant given the large number of people using such headphones in the current market. Intra-aural headphones and other similar headphones appear to be suitable for applications of binaural synthesis in, for example, the fields of gaming, or immersive film and video on handheld devices. The bone conduction headphones were shown to result in a significant degree of localisation errors. Whilst bone conduction headphones have been shown to be effective for binaural synthesis in other studies, those tested in the current study might not be appropriate. This has implications for their application as a headphone that can be used to render virtual sound sources and allow for the audition of environmental sounds; for example as a headphone that could provide additional auditory cues to the blind. It should be noted however that conclusive results would require the addition of more subjects, as only one subject was tested. In terms of the circumaural headphones, their effectiveness was confirmed for use in binaural synthesis, yet this was expected given their price range and the fact that they were developed for the audiophile market. Based on the results, open circumaural headphones would be recommended over closed circumaural headphones.

The fact that headphone equalisation, performed as part of the study of different headphone types, did not prove to be effective when taken on the aggregate across all headphones tested, suggests that the method used would not be a way in which headphones with a poor frequency response could be adapted to perform better. This would mean that for the headphone types tested, it would be a better option for consumers to purchase headphones that are known to be effective for binaural synthesis (such as the intra-aural headphones and open circumaural headphones) rather than try and manipulate the frequency response of a headphone that has not been proven to be effective. This is of course only valid for the method of headphone equalisation used in the study, and does not preclude that another equalisation method might be more effective. In addition, the equalisation did have a positive effect for the open circumaural headphone, which means that for this type of headphone an equalisation could further improve the quality of a rendering in VAS.

For the study involving the listening test design, there are potential applications of the results in terms of future validations of HRTF personalisation techniques. In fact the listening test design, having shown to help reduce the variability in subject responses, is not a method that is restricted to only judgements of different HRTFs; it could possibly be applied to any type of audio quality evaluation. The results also showed that expertise has an effect on the repeatability of perceptual judgements, and that expertise alone could not ensure a low degree of variance. The implication of the observed variance in responses across replicates, for even the experienced listeners, is that response variance must be taken into account when evaluating any procedure that aims to improve audio quality. If replicates of a particular listening test are not possible, then selecting subjects with a very high level of expertise can help to improve the reliability of the results.

The study also highlighted that the evaluation of renderings of sound sources in VAS can vary depending on which attributes of the quality of a binaural synthesis is being assessed. The attributes *sense of direction* and *front image quality* were shown to describe different components of the renderings; the attribute *sense of distance* was shown to be correlated with *sense of direction*. These findings are significant to the development of applications of binaural synthesis in the consumer market, as customisation techniques are often validated using perceptual judgements via listening tests, and depending on the application of the technology, some attributes may be more relevant than others.

For the studies relating to the analysis of HRTF inter-subject differences and the HRTF selection procedure, significant advances were made towards the goal of producing a viable customisation tool for the application of binaural synthesis to consumer market. It is of great interest to companies such as Arkamys, which funded this industry-linked research and specialises in software solutions for improving sound quality, to establish whether an affordable personalisation solution can be brought to the consumer. To this end, the effectiveness of a predictive HRTF selection procedure was shown to be statistically significant, and a patent has been published. The next phase of the development of the predictive tool should be to test, using the knowledge of subject response variance from this body of work, the effectiveness of a selected optimal HRTF against a generic HRTF using soundscapes that would be similar to those used by the consumer, such as those used for video gaming.

The applications of binaural synthesis are far reaching because they allow us to simulate any soundscape via headphones. This technology could allow us to drastically improve hearing aids, provide spatial cues to the blind, help aeroplane pilots navigate in cockpits, or create completely immersive environments for cinema and video gaming, just to name a few possible uses of the technology. Binaural synthesis has been used extensively in academic research in the past, but has yet to be taken up on mass by the consumer market, despite other virtual technologies such as three-dimensional visual effects being widely adopted. One of the main reasons for this adoption lag is that renderings of virtual sound sources using generic HRTFs are not very effective in terms of their realism; personalised HRTFs need to be used if we are to expect the consumer market to embrace the technology. The findings from the studies presented in the research chapters, along with the suggested HRTF selection procedure, have brought us closer to being able to provide a personalised binaural synthesis using simple tools such as a photo of the listener's ear. The simplicity of the solution and ease of use for the consumer being paramount to bringing this technology to the masses.

Part III

APPENDIX

APPENDIX A

A.1 LISTENING TEST 3.2 PROTOCOL

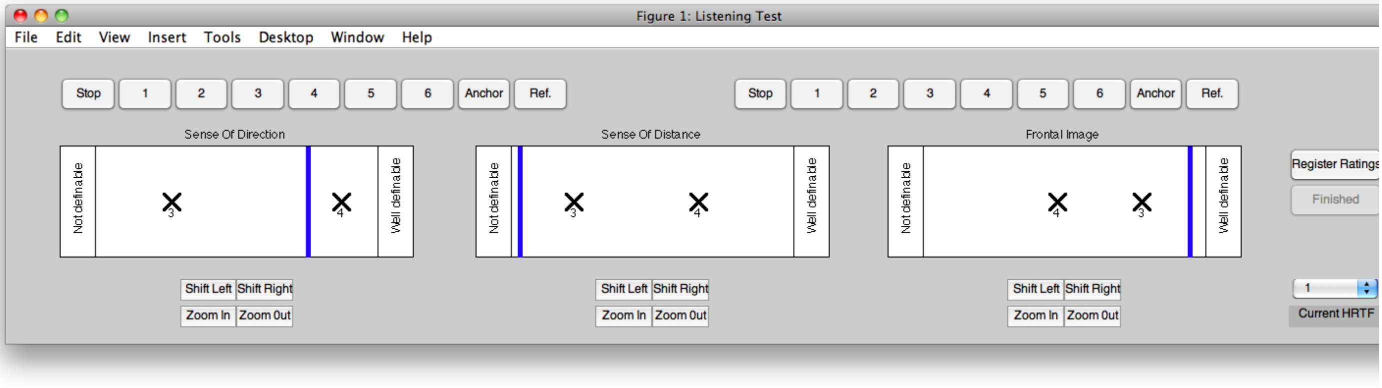
Below is the document provided to the subjects detailing the listening test protocol. The document was written in French as all the subjects were native French speakers.

Protocole du test d'écoute

Sommaire

Dans ce test d'écoute, vous serez invités à classifier 6 rendus audios tri-dimensionnels[1] (profils) en utilisant l'interface spécifiée. Cela se fera en 5 fois avec un entrainement au début. Vous ne serez pas informés des profils utilisés dans chacun des tests.

Exemple de l'interface (continuer la lecture pour une explication)

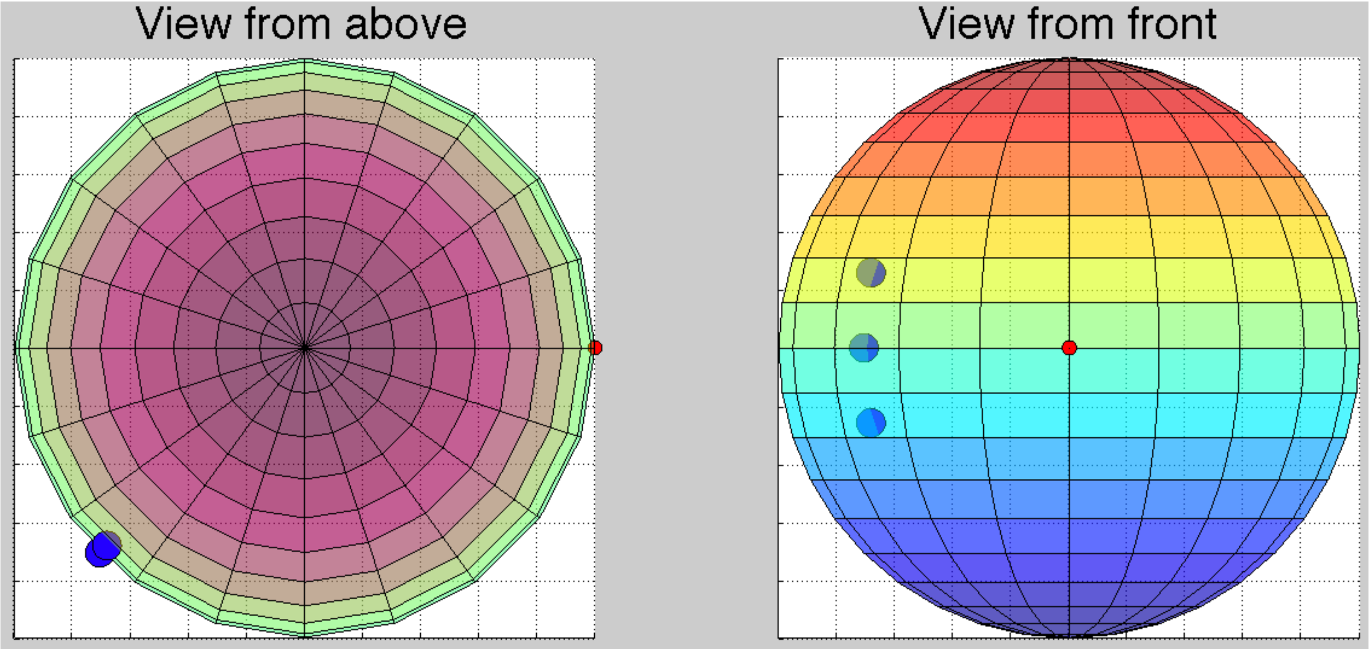


Méthodes

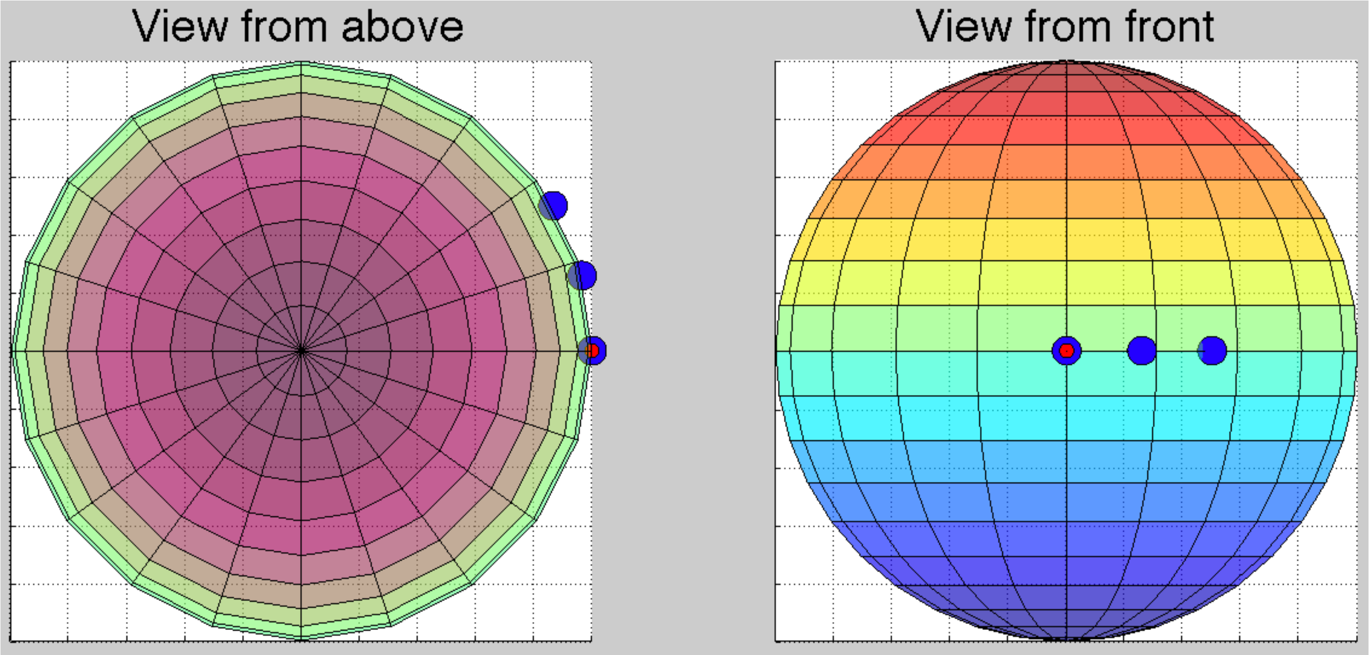
Il y a 6 profils différents afin de comparer et classer dans chaque test. Vous allez écouter un stimulus bruit blanc[2] qui suivra 2 trajectoires différentes dans l'espace pour chaque profil. Ces trajectoires sont affichées ci-dessous:

Imaginez que le point rouge est directement devant l'auditeur à une distance d'un mètre. Les sons partira de bas vers le hait pour trajectoire 1 et du centre vers la gauche pour trajectoire 2.

Trajectoire 1



Trajectoire 2



En plus des 6 différents profils étant considérés il y a une référence 'Ref.', qui constitue une adaptation du bruit blanc en utilisant vos propres HRTF[3] que nous supposons se fera entendre tout à fait naturel, comme si il y avait un haut-parleur à jouer à des positions dans les trajectoires. Il y a aussi un point d'ancrage 'Anchor' qui est une version dégradée de votre référence, que nous supposons ne sonnera pas naturel du tout. La référence et d'ancrage doit être utilisé comme un guide pour vous aider à évaluer les 6 profils différents.

Vous pouvez écouter les différents profils, y compris la référence et l'ancre, à tout moment pour autant et aussi longtemps que vous le souhaitez.

Les évaluations doivent être faites en utilisant l'interface curseur pour chaque attribut. Les évaluations doivent se situer entre les deux lignes verticales, la ligne verticale de gauche représente une évaluation **Pas définissable** (**Not definable**) et la ligne verticale de droite représente une évaluation **bien définissable** (**Well definable).**

Pour chaque profil, vous serez invité à faire une évaluation basée sur les 3 attributs suivants:

1. **Sens de distance.** Cet attribut décrit la façon dont la distance entre la source sonore et l'auditeur peut être défini. Cet attribut ne se rapporte pas à *quelle distance* vous percevez la source, mais plutôt de savoir si vous *pouvez* le *percevoir* comme venant d'une certaine distance.
2. **Sens de direction.** Cet attribut décrit la façon dont la direction de la source sonore peut être définie. Cet attribut ne se rapporte pas à la corrélation de la *position* de la source sonore par rapport à la référence, mais plutôt de savoir si la direction des sons dans la trajectoire sont *bien définis* et *distincts.* En gros, cet attribut se rapporte si la position de la source est diffuse (pas définissable) ou semble d'avoir une position spécifique et precise dans l'espace (bien définissable).
3. **Image frontale.** Cet attribut décrit la façon dont la perception du son provenant de l'avant de l'auditeur peut être définie. Cet attribut ne se rapporte pas à la corrélation de la *position* de la source sonore par rapport à la référence, mais plutôt de savoir si la source sonore est *perçu* comme venant d'en face de vous.

La première trajectoire doit être utilisé pour évaluer les attributs **sens de direction** et **sens de distance.** La seconde trajectoire doit être utilisé pour évaluer **l'image frontale.**

L'interface vous permet de zoomer et dézoomer pour chaque attribut et également au déplacement de l'échelle à gauche ou à droite si vous avez besoin d'un meilleur indice tandis résolution.

Afin d'évaluer un profil, vous devez le sélectionner dans le menu déroulant. Le profil étant considérée est affiché dans le menu déroulant. Une fois tous les trois attributs ont été évalués pour un profil, vous devez cliquer sur le bouton 'Register Ratings'. Après avoir cliqué sur le bouton 'Register Ratings' vous pouvez sélectionner un nouveau profil.

Pour modifier votre évaluation pour un profil déjà classé, vous devez finir l'évaluation du profil actuel que vous écoutez et cliquer sur 'Register Ratings', puis sélectionnez le profil que vous souhaitez modifier. Une fois que vous avez changé vos évaluations pour ce profil, vous devez cliquer à nouveau sur 'Register Ratings'.

Lorsque vous avez terminé de classer tous les profils et vous êtes confiants de vos évaluations, vous devez cliquer sur le bouton 'Finished'.

Merci!

Footnotes

[1] Un rendu audio trois-dimensionnel est un son joué via un casque qui a été manipulé pour qu'il donne l'impression que le son vient de l'extérieur de la tête de l'auditeur, par opposition à l'intérieur de la tête, comme c'est normalement le cas.

[2] S'il vous plaît demandez que ce son soit joué à vous maintenant

[3] Les HRTF sont les mesures qui ont été effectuées par Van à l'IRCAM avec les petits micros dans les oreilles

B

APPENDIX B

B.1 RELEVANT COMPUTING SCRIPTS

Below are the two relevant scripts, written in the numerical coding environment Matlab, for the analyses performed in chapter 8 and 9.

```
function [Chisq] = predict_best_HRTF(morph_data,model,use_regression,only_dist_error)
% close all hidden
load new_results
% 1034 created using df_equalize_ds.m because there was an error in original
plot_3D_model = 0;
colour_coded = 1;
percent_explained = 0;
use_input = 0;
all_subjects = 0;
load_saved_results = 0;
freq_range = 1;
use_optimal_dims = 1;
use_optimal_params = 1;
save_zscores = 0;
non_linear_pca = 0;
plot_position = 0;
regress_morph = 3;
show_waitbar = 0;
scrsz = get(0,'ScreenSize');
if use_input == 1
    use_regression = input(' 1) Use regression \n 2) Don''t use regression \n Enter 1 or 2: ');
    if use_regression == 1
        use_regression = 1;
    elseif use_regression == 2
        use_regression = 0;
    end
end
if use_input == 1 && use_regression == 1
    model = input([' 1) Morphology \n 2) FFT magnitude \n 3)'...
        ' Frequency Scaling \n 4) Octave \n 5) FFT magnitude '...
        'contralateral removed \n 6) ERB magnitude \n Please type the number of the '...
        'corresponding model: ']);
end
save_path = '/Users/davidschonstein/Documents/code/phd/DatabaseMatching/PCA/';
for aa = 1:length(Names)
    listen_subjects_C{aa} = num2str(Names(aa));
end
if model == 2
    if use_optimal_dims == 1
        % optimal dimensions using sequential_feature_selection_pca.m
        optimal_dims = [1,2,6,20,30,19,24,39,25,33,37,38,16,28,43,22,12,41];
    end
    % taken from pca_find_freq_range.m
    start_freq = 3250;
    stop_freq = 14250;
    % keep only subjects in listening test
    if all_subjects == 0
        for aa = 1:length(Names)
            HRTF_ind(aa) = find(strcmp(num2str(Names(aa)),listen_subjects_C));
        end
        listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    else
        HRTF_ind = 1:1:length(listen_subjects_C);
    end
    if freq_range == 1
        if exist([save_path 'smoothed_mag_est_dtf_all_freq_range.mat'],'file') ~= 2
            % created using create_pca_data.m
            load pca_fft_data_norm_comp.mat
            % create pca_data
            load IRC_1002_NORM_C_HRIR
            Fs = l_eq_norm_hrir_S.sampling_hz;
            fft_length = size(l_smoothed_mag_est_dtf_all,2);
            freqs = linspace(0,Fs/2,fft_length);
            bin_ind = find(freqs >= start_freq & freqs <= stop_freq);
            % create pca data
            l_data = l_smoothed_mag_est_dtf_all(:,bin_ind);
            r_data = r_smoothed_mag_est_dtf_all(:,bin_ind);
            offset = 1;
            no_subjects = size(l_smoothed_mag_est_dtf_all,1)/length(l_eq_norm_hrir_S.azim_v);
            pca_data_mag_subj = [];
            for subject_no = 1:no_subjects
                l_data_subj =  l_data(offset:offset+size(l_smoothed_mag_est_dtf_all,1)/no_subjects-↙
1,:);
                r_data_subj =  r_data(offset:offset+size(r_smoothed_mag_est_dtf_all,1)/no_subjects-↙
1,:);
                offset = offset + size(l_smoothed_mag_est_dtf_all,1)/no_subjects;
                l_data_cat = [];
                r_data_cat = [];
                for position_ind = 1:size(l_smoothed_mag_est_dtf_all,1)/no_subjects
                    l_data_cat = cat(2,l_data_cat,l_data_subj(position_ind,:));
                    r_data_cat = cat(2,r_data_cat,r_data_subj(position_ind,:));
                end
                pca_data_mag_subj(subject_no,:) = cat(2,l_data_cat,r_data_cat);
            end
            save([save_path 'smoothed_mag_est_dtf_all_freq_range'],'pca_data_mag_subj')
        else
            load smoothed_mag_est_dtf_all_freq_range
        end
    else
        if exist([save_path 'smoothed_mag_est_dtf_all.mat'],'file') ~= 2
            % created using create_pca_data.m
            load pca_fft_data_norm_comp.mat
            % create pca_data
            load IRC_1002_NORM_C_HRIR
            % create pca data
            l_data = l_smoothed_mag_est_dtf_all;
            r_data = r_smoothed_mag_est_dtf_all;
            offset = 1;
            no_subjects = size(l_smoothed_mag_est_dtf_all,1)/length(l_eq_norm_hrir_S.azim_v);
            pca_data_mag_subj = [];
            for subject_no = 1:no_subjects
                l_data_subj =  l_data(offset:offset+size(l_smoothed_mag_est_dtf_all,1)/no_subjects-↙
1,:);
                r_data_subj =  r_data(offset:offset+size(r_smoothed_mag_est_dtf_all,1)/no_subjects-↙
1,:);
                offset = offset + size(l_smoothed_mag_est_dtf_all,1)/no_subjects;
                l_data_cat = [];
                r_data_cat = [];
                for position_ind = 1:size(l_smoothed_mag_est_dtf_all,1)/no_subjects
                    l_data_cat = cat(2,l_data_cat,l_data_subj(position_ind,:));
                    r_data_cat = cat(2,r_data_cat,r_data_subj(position_ind,:));
                end
                pca_data_mag_subj(subject_no,:) = cat(2,l_data_cat,r_data_cat);
            end
            save([save_path 'smoothed_mag_est_dtf_all'],'pca_data_mag_subj')
        else
            load smoothed_mag_est_dtf_all
        end
    end
    listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    if non_linear_pca == 1
        [eigenvectors,net,network] = nlpca(pca_data_mag_subj(HRTF_ind,1:193)',3,'pre_pca','yes');
        figure
        nlpca_plot(net)
        axis equal
    else
        [eigenvectors,zscores_plot,eigenvalues] = princomp(pca_data_mag_subj(HRTF_ind,:),'econ');
        if use_optimal_dims == 1 && use_regression == 0
            zscores_plot = zscores_plot(:,optimal_dims);
        end
    end
elseif model == 1
    % keep only subjects in listening test
    if all_subjects == 0
        for aa = 1:length(Names)
            HRTF_ind(aa) = find(strcmp(num2str(Names(aa)),listen_subjects_C));
        end
        listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    else
        HRTF_ind = 1:1:length(listen_subjects_C);
    end
    if freq_range == 1
        cpa_file_name = 'cpa_correlation_freq_range.mat';
    else
        cpa_file_name = 'cpa_correlation.mat';
    end
    if exist([save_path cpa_file_name],'file') ~= 2
        oct_fft = 2;
        switch oct_fft
            case 1
                % taken from pca_find_freq_range.m
                start_freq = 3250;
                stop_freq = 14250;
                % created using create_pca_data.m
                load pca_fft_data_norm_comp
            case 2
                % taken from pca_find_freq_range.m
                if freq_range == 1
                    start_freq = 4750;
                    stop_freq = 12250;
                else
                    start_freq = 0;
                    stop_freq = 22050;
                end
                % created using pca_oct_freq_filt.m
                load oct_filt_hrtfs_bpo_35_comp
        end
        h_subject = waitbar(0,'Processing each pair of subjects...');
        for subject_ind = 1:length(listen_subjects_C)
            waitbar(subject_ind/length(listen_subjects_C),h_subject)
            subject_s = listen_subjects_C{subject_ind};
            switch oct_fft
                case 1
                    norm = 1;
                    type = 'compensated';
                    [l_hrir_S,r_hrir_S,l_hrtf_log_mag,r_hrtf_log_mag,nfft,NumUniquePts] ...
                        = pca_load_HRIR(subject_s,type,norm);
                    l_mag_est_dtf = l_hrtf_log_mag;
                    r_mag_est_dtf = r_hrtf_log_mag;
                    % take optimal frequency range
                    Fs = 44100;
                    f = (0:NumUniquePts-1)*Fs/nfft;
                    freq_ind = f >= start_freq & f <= stop_freq;
                    % critical band smoothing
                    l_mag_S = rmfield(l_hrir_S,'content_m');
                    r_mag_S = rmfield(r_hrir_S,'content_m');
                    % make positive for smoothing
                    l_mag_S.content_m = l_mag_est_dtf - min_mag;
                    r_mag_S.content_m = r_mag_est_dtf - min_mag;
                    l_smoothed_mag_S = critical_band_smoothing(l_mag_S);
                    r_smoothed_mag_S = critical_band_smoothing(r_mag_S);
                    % data
                    l_data = l_smoothed_mag_S.content_m;
                    r_data = r_smoothed_mag_S.content_m;
                case 2
                    % data
                    l_data = l_oct_filt_hrtfs{subject_ind};
                    r_data = r_oct_filt_hrtfs{subject_ind};
                    % take optimal frequency range
                    freq_ind = F0_used >= start_freq & F0_used <= stop_freq;
                    norm = 1;
                    type = 'compensated';
                    [l_hrir_S,r_hrir_S] ...
                        = pca_load_HRIR(subject_s,type,norm);
                    l_smoothed_mag_S = l_hrir_S;
            end
            % frequency range
            l_data = l_data(:,freq_ind);
            r_data = r_data(:,freq_ind);
            % split into lower sample regions for high elevations
            ind_under_60 = l_smoothed_mag_S.elev_v < 60;
            azim_n = length(unique(l_smoothed_mag_S.azim_v(ind_under_60)));
            elev_n = length(unique(l_smoothed_mag_S.elev_v(ind_under_60)));
            l_data_vol = reshape(l_data(ind_under_60,:),elev_n,azim_n,size(l_data(ind_under_60,:),2));
            r_data_vol = reshape(r_data(ind_under_60,:),elev_n,azim_n,size(r_data(ind_under_60,:),2));
            % add separately elevations 60, 75 and 90
            left_over_el = unique(l_smoothed_mag_S.elev_v(ind_under_60 == false));
            for gg = 1:length(left_over_el)
                ind = l_smoothed_mag_S.elev_v == left_over_el(gg);
                ind = ismember(unique(l_smoothed_mag_S.azim_v),unique(l_smoothed_mag_S.azim_v(ind)));
                l_data_vol(size(l_data_vol,1)+1,ind,:) = l_data(ind,:);
                % fill rest of positions with nan
                l_data_vol(size(l_data_vol,1),ind == false,:) = nan(length(find(ind == false)),size↙
(l_data,2));
            end
            % find azimuth and elevation for each frequency bin
            azimuth_values = unique(l_smoothed_mag_S.azim_v);
            elevation_values = unique(l_smoothed_mag_S.elev_v);
            for cpa_ind = 1:length(l_data_vol)
                % left ear
                [C,I] = nanmax(l_data_vol(:,:,cpa_ind));
                [C,cpa_azim_ind] = nanmax(C);
                cpa_elev_ind = I(cpa_azim_ind);
                l_cpa_position(cpa_ind,:) = [azimuth_values(cpa_azim_ind) ...
                    elevation_values(cpa_elev_ind)];
                % right ear
                [C,I] = nanmax(r_data_vol(:,:,cpa_ind));
                [C,cpa_azim_ind] = nanmax(C);
                cpa_elev_ind = I(cpa_azim_ind);
                r_cpa_position(cpa_ind,:) = [azimuth_values(cpa_azim_ind) ...
                    elevation_values(cpa_elev_ind)];
            end
            l_cpa_position_all{subject_ind} = l_cpa_position;
            r_cpa_position_all{subject_ind} = r_cpa_position;
            plot_cpa = 0;
            if plot_cpa == 1
                figure(1)
                ah_1 = axes;
                figure(2)
                ah_2 = axes;
                for hh = 1:length(l_cpa_position)
                    cmap = colormap(hsv(length(l_cpa_position)));
                    plot(ah_1,l_cpa_position(:,1),l_cpa_position(:,2),'*','Color',cmap(hh,:))
                    plot(ah_2,r_cpa_position(:,1),r_cpa_position(:,2),'*','Color',cmap(hh,:))
                end
                keyboard
            end
        end
        close(h_subject)
        % calculate corr2 between each pair of subjects
        choose_values = nchoosek(HRTF_ind,2);
        h_all = waitbar(0,'Processing each pair of subjects...');
        for choose_ind = 1:length(choose_values)
            waitbar(choose_ind/length(choose_values),h_all)
            l_r(choose_ind) = corr2(l_cpa_position_all{choose_values(choose_ind,1)},...
                l_cpa_position_all{choose_values(choose_ind,2)});
            r_r(choose_ind) = corr2(r_cpa_position_all{choose_values(choose_ind,1)},...
                r_cpa_position_all{choose_values(choose_ind,2)});
        end
        close(h_all)
        save([save_path cpa_file_name],'l_r','r_r')
    else
        if freq_range == 1
```

```
            load cpa_correlation_freq_range
        else
            load cpa_correlation
        end
    end
    % create MS
    l_D = ones(size(l_r))-abs(l_r);
    [l_Y,l_e] = cmdscale(l_D);
    r_D = ones(size(r_r))-abs(r_r);
    [r_Y,r_e] = cmdscale(r_D);
    % right ear best
    left_right = input([' 1) Left ear \n 2) Right ear \n'...
        ' Please type the number of the corresponding setting: ']);
    if left_right == 1
        zscores_plot = l_Y;
        if use_optimal_dims == 1
            calculate_dims = 1;
            if calculate_dims == 1
                morph_data = [];
                pca_data_oct = [];
                subjects_C = {}; % removed numbers
                % calculate the optimal dimensions for each subject
                [optimal_dims] = sequential_feature_selection_pca(zscores_plot,...
                    subjects_C,model,use_regression,pca_data_oct,morph_data);
            else
                % optimal dimensions using sequential_feature_selection_pca.m
                optimal_dims = [1 9 12 13 16 19 23];
            end
            zscores_plot = zscores_plot(:,optimal_dims);
        end
    else
        zscores_plot = r_Y;
        if use_optimal_dims == 1
            calculate_dims = 1;
            if calculate_dims == 1
                morph_data = [];
                pca_data_oct = [];
                subjects_C = {}; % removed numbers
                % calculate the optimal dimensions for each subject
                [optimal_dims] = sequential_feature_selection_pca(zscores_plot,...
                    subjects_C,model,use_regression,pca_data_oct,morph_data);
            else
                % optimal dimensions using sequential_feature_selection_pca.m
                optimal_dims = [1 4 7 8 9 11 14 16 22 23];
            end
            zscores_plot = zscores_plot(:,optimal_dims);
        end
    end
    % analyse the results of cmdscale
    l_e_norm = [l_e l_e/max(abs(l_e))];
    r_e_norm = [r_e r_e/max(abs(r_e))];
    l_relerr = abs(l_D - pdist(l_Y))/max(l_D);
    l_percent_error = sum(l_relerr <= 0.2)/length(l_relerr);
    r_relerr = abs(r_D - pdist(r_Y))/max(r_D);
    r_percent_error = sum(r_relerr <= 0.2)/length(r_relerr);
    l_maxrelerr = max(abs(l_D - pdist(l_Y)))/max(l_D);
    r_maxrelerr = max(abs(r_D - pdist(r_Y)))/max(r_D);
elseif model == 3
    if all_subjects == 0
        % keep only subjects in listening test
        for aa = 1:length(Names)
            HRTF_ind(aa) = find(strcmp(num2str(Names(aa)),listen_subjects_C));
        end
    else
        HRTF_ind = 1:1:length(listen_subjects_C);
    end
    [l_scale_factor_min,r_scale_factor_min,listen_subjects_C] = pca_freq_scaling_mdscaling;
    listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    % right best
    ear = input([' 1) Left ear \n 2) Right ear \n'...
        ' Please type the number of the corresponding setting: ']);
    if ear == 1
        l_D = l_scale_factor_min - ones(1,length(l_scale_factor_min));
        [l_Y,l_e] = cmdscale(l_D);
        zscores_plot = l_Y(HRTF_ind,:);
        eigenvalues = l_e;
    elseif ear == 2
        r_D = r_scale_factor_min - ones(1,length(r_scale_factor_min));
        [r_Y,r_e] = cmdscale(r_D);
        zscores_plot = r_Y(HRTF_ind,:);
        eigenvalues = r_e;
    end
    if use_optimal_dims == 1
        each_subject = 0;
        if each_subject == 1
            % calculate the optimal dimensions for each subject
            for i = 1:length(listen_subjects_C)
                if strcmp(listen_subjects_C{i},'1013') == 0
                    subjects_C = {listen_subjects_C{i}};
                    [optimal_dims] = sequential_feature_selection_pca(zscores_plot,subjects_C,model);
                    optimal_dims_all{i} = optimal_dims;
                end
            end
            %                         hist([optimal_dims_all{:}],1:45)
        else
            morph_data = [];
            pca_data_oct = [];
            subjects_C = {}; % removed numbers
            % calculate the optimal dimensions for each subject
            [optimal_dims] = sequential_feature_selection_pca(zscores_plot,...
                subjects_C,model,use_regression,pca_data_oct,morph_data);
        end
        % optimal dimensions using
        % sequential_feature_selection_pca.m
        zscores_plot = zscores_plot(:,optimal_dims);
    end
elseif model == 6
    if isempty(morph_data) == 1
        % left = 1, right = 2
        left_right = 1;
        morph_data = [head_m pinna_m(:,:,left_right)];
        if regress_morph == 1
            [morph_data] = missing_data_regression(morph_data);
        elseif regress_morph == 2
            % set any NaNs to mean of that parameter
            mean_values = nanmean(morph_data);
            [I,J] = ind2sub(size(morph_data),isnan(morph_data));
            for ii = 1:length(I)
                morph_data(ii,logical(I(ii,:))) = mean_values(logical(I(ii,:)));
            end
        elseif regress_morph == 3
            % leave NaNs in data
            treat_nans = 1;
            if treat_nans == 1
                % find NaNs in X
                nan_sum = sum(isnan(morph_data));
                col_rmv = nan_sum > size(morph_data,1)/2;
                % remove columns of data that have over half NaN
                morph_data(:,col_rmv) = [];
            end
        end
    end
    if use_optimal_params == 1
        % from sequential_feature_selection_morph.m
        load forward_fs_regress_morph
        if treat_nans == 1
            morph_data = morph_data(:,optimal_params);
        else
            morph_data = morph_data(:,ind_params{1});
        end
    end
    % remove subjects for which there is no morphological data
    ind_nan = sum(isnan(morph_data),2) > 0;
    morph_data(ind_nan,:) = [];
    [eigenvectors,zscores_plot,eigenvalues] = princomp(morph_data);
    % remove subject from listen_subjects_C
    listen_subjects_C = {listen_subjects_C{ind_nan == 0}};
    calculate_opt_dims = 1;
    if use_optimal_dims == 1
        if calculate_opt_dims == 1
            [optimal_dims] = sequential_feature_selection_pca(zscores_plot,...
                listen_subjects_C,model,use_regression,[],morph_data);
        elseif use_optimal_dims == 0
            optimal_dims = [5 7 8];
        end
        zscores_plot = zscores_plot(:,optimal_dims);
    end
elseif model == 7
    load pca_data_fb_subj_raw
    listen_subjects_C = {}; % removed numbers
    [eigenvectors,zscores_plot,eigenvalues] = princomp(pca_data_fb_subj,'econ');
elseif model == 8
    load pca_variables_raw_contralateral_removed
    listen_subjects_C = {}; % removed numbers
    ear = input([' 1) Left ear \n 2) Right ear \n 3) Left and right ear \n'...
        ' Please type the number of the corresponding setting: ']);
    if ear == 1
        pca_data_mag_subj = pca_data_mag_subj_contra_left;
    elseif ear == 2
        pca_data_mag_subj = pca_data_mag_subj_contra_right;
    elseif ear == 3
        pca_data_mag_subj = pca_data_mag_subj_contra_all;
    end
    [eigenvectors,zscores_plot,eigenvalues] = princomp(pca_data_mag_subj,'econ');
    for aa = 1:length(Names)
        listen_subjects_C{aa} = num2str(Names(aa));
    end
elseif model == 4
    if use_optimal_dims == 1
        % optimal dimensions using sequential_feature_selection_pca.m
        optimal_dims = [1,2,6,21,17,11,18,23,39,31,28,32,33,25,27,37];
    end
    load oct_filt_hrtfs_bpo_35_comp
    % taken from pca_find_freq_range.m
    start_freq = 4750;
    stop_freq = 12250;
    if all_subjects == 0
        for aa = 1:length(Names)
            HRTF_ind(aa) = find(strcmp(num2str(Names(aa)),listen_subjects_C));
        end
        listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    else
        HRTF_ind = 1:1:length(listen_subjects_C);
    end
    if freq_range == 1
        freq_ind = F0_used >= start_freq & F0_used <= stop_freq;
    else
        freq_ind = 1:size(l_oct_filt_hrtfs{1},2);
    end
    for ii = 1:length(l_oct_filt_hrtfs)
        l_hrtfs_cat = [];
        r_hrtfs_cat = [];
        for jj = 1:size(l_oct_filt_hrtfs{ii},1)
            l_hrtfs_cat = [l_hrtfs_cat l_oct_filt_hrtfs{ii}(jj,freq_ind)];
            r_hrtfs_cat = [r_hrtfs_cat r_oct_filt_hrtfs{ii}(jj,freq_ind)];
        end
        pca_data_oct(ii,:) = [l_hrtfs_cat r_hrtfs_cat];
    end
    listen_subjects_C = {}; % removed numbers
    listen_subjects_C = {listen_subjects_C{HRTF_ind}};
    [eigenvectors,zscores_plot,eigenvalues] = princomp(pca_data_oct(HRTF_ind,:),'econ');
    if use_optimal_dims == 1 && use_regression == 0
        zscores_plot = zscores_plot(:,optimal_dims);
    end
end
if save_zscores == 1
    save([save_path 'zscores_plot_model_' num2str(model)],'zscores_plot','Names')
    display(['zscores saved as zscores_plot_model_' num2str(model) '.mat'])
    return
end
if percent_explained == 1
    percent_explained = 100*eigenvalues/sum(eigenvalues);
    fh_v = figure('Position',scrsz);
    ah_v = axes;
    pareto(ah_v,percent_explained)
    xlabel('Principal Component')
    ylabel('Variance Explained (%)')
    pause
end
calculate = 1;
zscores = [];
subj_coord = [];
% make sure that the subjects are only from LISTEN
listen_subjects = str2double(listen_subjects_C);
if all_subjects == 0 && model == 2
    pca_data_mag_subj = pca_data_mag_subj(HRTF_ind,:);
elseif all_subjects == 0 && model == 4
    pca_data_oct = pca_data_oct(HRTF_ind,:);
end
if calculate == 1 && use_regression == 1
    if isempty(morph_data) == 1
        pinna_mean_m = (pinna_m(:,:,1)+pinna_m(:,:,2))/2;
        morph_data = cat(2,head_m,pinna_mean_m);
        if regress_morph == 1
            [morph_data] = missing_data_regression(morph_data);
        elseif regress_morph == 2
            % set any NaNs to mean of that parameter
            mean_values = nanmean(morph_data);
            [I,J] = ind2sub(size(morph_data),isnan(morph_data));
            for ii = 1:length(I)
                morph_data(ii,logical(I(ii,:))) = mean_values(logical(I(ii,:)));
            end
        elseif regress_morph == 3
            % leave NaNs in data
            treat_nans = 1;
            if treat_nans == 1
                % find NaNs in X
                nan_sum = sum(isnan(morph_data));
                col_rmv = nan_sum > size(morph_data,1)/2;
                % remove columns of data that have over half NaN
                morph_data(:,col_rmv) = [];
            end
        end
        % calculated using sequential_feature_selection_morph.m using
        % optimal dimensions i.e. calculate_opt_dims = 1
        load forward_fs_regress_morph
        if treat_nans == 1
            morph_data = morph_data(:,optimal_params);
        else
            morph_data = morph_data(:,ind_params);
        end
    end
    if use_optimal_dims == 1
```

```
        % calculating optimal dimensions is done once a global analysis (use_optimal_dims = 0) is
        % performed using the optimal dimensions from octave filtered DTF
        % MS using PCA
        calculate_opt_dims = 3;
        if calculate_opt_dims == 1
            [optimal_dims] = sequential_feature_selection_pca(zscores_plot,...
                listen_subjects_C,model,use_regression,pca_data_oct,morph_data);
        elseif calculate_opt_dims == 2
            % optimal dimensions using sequential_feature_selection_pca.m
            % above
            optimal_dims = [1 13 27 10 2 30 35 38 21 14 34 19 36 12];
            zscores_plot = zscores_plot(:,optimal_dims);
        elseif calculate_opt_dims == 3
            % use the same PCs as for model 2
            optimal_dims = [1,2,6,20,30,19,24,39,25,33,37,38,16,28,43,22,12,41];
        end
        zscores_plot = zscores_plot(:,optimal_dims);
    end
    if show_waitbar == 1
        h_all = waitbar(0,'Processing each set of HRTFs...');
    end
    % find subjects to regress for
    for subject_ind = 1:length(Names)
        if show_waitbar == 1
            waitbar(subject_ind/length(Names),h_all)
        end
        % remove current subject from data
        subj_morph = morph_data(subject_ind,:);
        morph_data_use = morph_data;
        morph_data_use(subject_ind,:) = [];
        Names_without_subj = Names;
        Names_without_subj(subject_ind,:) = [];
        listen_subjects_use = listen_subjects;
        listen_subjects_use(subject_ind) = [];
        HRTF_ind = [];
        for aa = 1:length(Names_without_subj)
            HRTF_ind(aa) = find(listen_subjects_use == Names_without_subj(aa));
        end
        % get position of subject
        pos_subject = zscores_plot(subject_ind,:);
        % create MS without the subject's data
        [eigenvectors,zscores,eigenvalues] = princomp(pca_data_mag_subj(ismember(Names,↙
Names_without_subj),:),'econ');
        if use_optimal_dims == 1
            % take first however many is needed as they are ranked
            optimal_dims = optimal_dims(optimal_dims <= size(zscores,2));
            zscores = zscores(:,optimal_dims);
        end
        % save this subject-specific MS
        zscores_all_subjects{subject_ind} = zscores;
        % generate morphology data without subject
        X = cat(2,ones(length(morph_data_use),1),morph_data_use);
        % regress for subject's position using optimal number of PCs
        b = [];
        for zscore_ind = 1:size(zscores,2)
            b(:,zscore_ind) = regress(zscores(:,zscore_ind),X);
            % regression can only be based on subjects in Names for which
            % we have morphological parameters
            if sum(isnan(subj_morph)) > 0
                subj_coord(subject_ind,zscore_ind) = NaN;
            else
                subj_coord(subject_ind,zscore_ind) = b(1,zscore_ind) + nansum(b(2:end,zscore_ind).↙
*subj_morph');
            end
        end
        % calculate distance between regression position and real position
        dist_error(subject_ind) = pdist([subj_coord(subject_ind,:); pos_subject(1:size(subj_coord,↙
2))]);
        if plot_position == 1
            % plot position
            figure(1)
            cla
            plot(zscores(:,1),zscores(:,2),'*')
            hold on
            plot(subj_coord(subject_ind,1),subj_coord(subject_ind,2),'g*')
            plot(pos_subject(1),pos_subject(2),'r*')
            dist_error(subject_ind)
            pause
        end
    end
    if show_waitbar == 1
        close(h_all)
    end
else
    subj_coord = [];
    dist_error = [];
end
if only_dist_error == 0
    if load_saved_results == 1
        if model == 2 && calculate == 1 && use_regression == 1
            save(strcat(save_path,save_name),'zscores','subj_coord')
        end
        if model == 4 && calculate == 1
            save(strcat(save_path,save_name),'zscores','subj_coord')
        end
        if model == 5 && calculate == 1
            if ear == 1
                save(strcat(save_path,save_name),'zscores','subj_coord')
            elseif ear == 2
                save(strcat(save_path,save_name),'zscores','subj_coord')
            elseif ear == 3
                save(strcat(save_path,save_name),'zscores','subj_coord')
            end
        end
    end
    if plot_3D_model == 1
        for ss = 1:length(Names)
            % plot PC1 weights against PC2 weights against PC3 weights
            fh_w = figure('Position',scrsz);
            ah_w = axes;
            for ii = 1:length(listen_subjects_C)
                listen_subjects_C_3D{ii} = ['   ' listen_subjects_C{ii}];
            end
            if colour_coded == 1
                ratings = ResSq_m(:,ss);
                ind_excellent = ratings == 4;
                plot3(ah_w,zscores_plot(ss,1),zscores_plot(ss,2),zscores_plot(ss,3),'m+','MarkerSize',↙
18,'LineWidth',5)
                hold(ah_w,'on')
                plot3(ah_w,zscores_plot(ind_excellent,1),zscores_plot(ind_excellent,2),zscores_plot↙
(ind_excellent,3),'g+','MarkerSize',15,'LineWidth',5)
                ind_good = ratings == 3;
                plot3(ah_w,zscores_plot(ind_good,1),zscores_plot(ind_good,2),zscores_plot(ind_good,↙
3),'y+','MarkerSize',15,'LineWidth',5)
                ind_bad = ratings == 2;
                plot3(ah_w,zscores_plot(ind_bad,1),zscores_plot(ind_bad,2),zscores_plot(ind_bad,↙
3),'r+','MarkerSize',15,'LineWidth',5)
                if use_regression == 1
                    plot3(ah_w,subj_coord(ss,1),subj_coord(ss,2),subj_coord(ss,3),'b+','MarkerSize',↙
15,'LineWidth',5)
                end
            else
                plot3(ah_w,zscores_plot(:,1),zscores_plot(:,2),zscores_plot(:,3),'b+','MarkerSize',↙
15,'LineWidth',5)
            end
            axis equal
            hold(ah_w,'on')
            text(zscores_plot(:,1),zscores_plot(:,2),zscores_plot(:,3),listen_subjects_C_3D,'FontSize',↙
15)
            grid(ah_w,'on')
            xlabel(ah_w,'First principal component','FontSize',30);
            ylabel(ah_w,'Second principal component','FontSize',30);
            zlabel(ah_w,'Third principal component','FontSize',30);
            view(0,90)
            axis(ah_w,[-35 40 -40 35])
            if colour_coded == 1
                legend↙
(ah_w,'Excellent','Okay','Bad','Subject','Prediction','Location','NorthEastOutside')
            end
            set(ah_w,'FontSize',20,'XTick',-30:10:30,'YTick',-30:10:30)
            if Names(ss,:) == 1047
                keyboard
            end
            pause
            close(fh_w)
        end
    end
    save_distributionPlot = 1;
    if exist('optimal_dims','var') == 0
        optimal_dims = [];
    end
    if exist('zscores_all_subjects','var') == 0
        zscores_all_subjects = [];
    end
    [p,Chisq,rank_corr] ...
        = pca_rank_and_graph_regress_all_subjects(zscores_plot,subj_coord,...
        listen_subjects_C,use_regression,model,save_distributionPlot,zscores_all_subjects);
    % mean(p)
    [max_Chisq,no_dimensions] = max(Chisq)
    p_max_Chisq = p(no_dimensions);
    %     keyboard
else
    p = [];
    p_excellent_top_HRTFs = [];
    ind_subject_remove = [];
end
```

```
function [p,Chisq,rank_corr] ...
    = pca_rank_and_graph_regress_all_subjects(zscores_plot,subj_coord,...
    subjects_C,use_regression,model,save_distributionPlot,zscores_all_subjects)
% rank HRTFs based on regression coordinates of subject and model
% close all hidden
load new_results
plot_rank_all = 0;
golden_heads = 0;
coloured_model_plot = 0;
rank_my_HRTF = 0;
plot_anova = 0;
excellent_good = 0;
OK = 1;
compare_regression = 0;
top_HRTFs = 0;
plot_top_HRTFs = 0;
plot_ranked_HRTFs_best_model = 0;
save_for_plot = 0;
test_each_dimension = 0;
% use distance = 3 or rank = 1 or corr = 2
use_rank = 3;
show_waitbar = 0;
save_percentages = 0;
plot_position = 0;
no_HRTFs_top = 10;
scrsz = get(0,'ScreenSize');
save_path = '/Users/dschonstein/MyFiles/PhD/MatlabMapped/DatabaseMatching/PCA/';
model_names = {'Covert peak','FFT magnitude','Frequency scaling',...
    'Octave magnitude','Notch frequencies','Morphology'};
model_name = model_names{model};
listen_subjects_C = {}; % removed numbers
if model == 6
    ind = ismember(str2double(listen_subjects_C),str2double(subjects_C));
    ResSq_m = ResSq_m(ind,ind);
    Names = Names(ind);
end
if use_regression == 1
    % remove subjects for which a NaN exists for their position or did not
    % make judgements
    ind_subject = sum(isnan(ResSq_m)) + sum(isnan(subj_coord),2)' == 0;
    subj_coord = subj_coord(ind_subject,:);
else
    ind_subject = sum(isnan(ResSq_m)) == 0;
end
% all subjects are judged
ind_ms = 1:length(zscores_plot);
subjects_C = {subjects_C{ind_subject}};
% names of subjects being judged
zscores_plot = zscores_plot(ind_ms,:);
ResSq_m = ResSq_m(ind_ms,:);
listen_subjects = str2double(subjects_C);
% HRTF_ind = ismember(listen_subjects,Names);
% number of subjects judging the HRTFs
no_subjects = length(listen_subjects);
% number of subjects being judged
no_subjects_judged = length(find(ind_ms));
% generate vectors of pairwise judgment indices
if plot_top_HRTFs
    fh_top = figure('Position',[1 900 1440 900]);
    ah_top = axes;
end
if show_waitbar == 1
    h_all = waitbar(0,'Processing each set of HRTFs...');
end
for subject_ind = 1:no_subjects
    if show_waitbar == 1
        waitbar(subject_ind/length(subjects_C),h_all)
    end
    ind_anova_cat = [];
    ind_pcs = [];
    offset = 0;
    Names_list = Names;
    % index of subject for jugments matrix
    current_subj_ind = find(ismember(Names,listen_subjects(subject_ind)));
    % consider all judgments except the judgments of oneself
    % list of subjects judged without current subject
    Names_list(ismember(Names,listen_subjects(subject_ind))) = [];
    % find ranking from listening test
    results_per_subj = ResSq_m(:,current_subj_ind);
    results_per_subj(ismember(Names,listen_subjects(subject_ind)),:) = [];
    [results_sorted_listen,sort_ind_listen] = sort(results_per_subj,'descend');
    sorted_HRTFs_listen = Names_list(sort_ind_listen);
    if use_regression == 1
        % create MS of subjects judged without the subject's data
        zscores = zscores_all_subjects{current_subj_ind};
        % add the subject's regressed position
        % test = 1 for using PCA with all subjects
        test = 0;
        if test == 1
            zscores_regression = zscores_plot;
        else
```

```
            zscores_regression = [zscores(1:current_subj_ind-1,:); subj_coord(subject_ind,1:size↙
(zscores,2)); ...
                zscores(current_subj_ind:end,:)];
        end
        % calculate distance between regression position and real position
        dist_error(subject_ind) = pdist([subj_coord(subject_ind,:); zscores_plot(current_subj_ind,1:↙
size(subj_coord,2))]);
    else
        zscores_regression = zscores_plot;
    end
    if plot_position == 1 && use_regression == 1
        % plot position
        figure(1)
        cla
        plot(zscores_regression(:,1),zscores_regression(:,2),'*')
        hold on
        plot(zscores_regression(current_subj_ind,1),zscores_regression(current_subj_ind,2),'g*')
        plot(zscores_plot(current_subj_ind,1),zscores_plot(current_subj_ind,2),'r*')
        dist_error(subject_ind);
        ind_subj_dist = find(subject_ind ~= 1:no_subjects);
        test_dist_subj = [];
        for m = 1:no_subjects-1
            test_dist_subj(m) = mean(pdist([zscores_regression(subject_ind,:); zscores_regression↙
(ind_subj_dist(m),:)]));
        end
        test_dist(subject_ind) = max(test_dist_subj)
        %          subject_ind
        pause
    end
    % select a random set of top HRTFs for the subject
    %     rand_ind = randperm(length(Names)-1);
    for zscore_ind = 1:size(zscores_regression,2)
        % add coordinates of subject from regression
        if use_regression == 1
            if compare_regression == 1 && zscore_ind == 1
                fh_c = figure('Position',[1 900 1440 900]);
                ah_c = axes;
                plot(ah_c,zscores_plot(subject_ind,:))
                hold on
                plot(ah_c,subj_coord(subject_ind,:),'r')
                pause
                close(fh_c)
            end
            if compare_regression == 1
                distance_test(zscore_ind) = pdist(cat(1,zscores_plot(subject_ind,1:zscore_ind),↙
subj_coord(subject_ind,1:zscore_ind)));
            end
        end
        % starts at 1D
        if test_each_dimension == 1
            points = zscores_regression(:,zscore_ind);
        else
            points = zscores_regression(:,1:zscore_ind);
        end
        % create distances from model with however many subjects
        distances = pdist(points)';
        distances_ind = nchoosek(1:no_subjects_judged,2);
        % find index of pairs for a particular subject
        subject_pairs = [current_subj_ind*ones(size(1:no_subjects_judged)); 1:no_subjects_judged]';
        subject_pairs(current_subj_ind,:) = [];
        % find corresponding distances for pairs
        tf = ismember(sort(distances_ind,2),sort(subject_pairs,2),'rows');
        % select distances for Names subjects without current subject
        distances_subj_listen = distances(tf);
        % get subjects ranked from distance index
        subjects_ranked = sum(distances_ind(tf,:),2) - current_subj_ind;
        [results_sorted_pca,ranking] = sort(distances_subj_listen,'ascend');
        % subjects ranked
        subjects_ranked_number = Names(subjects_ranked);
        sorted_HRTFs_pca = subjects_ranked_number(ranking);
        % compare rank with MS and with regression
        compare{subject_ind} = [sorted_HRTFs_listen sorted_HRTFs_pca];
        [tf,loc] = ismember(sorted_HRTFs_listen,sorted_HRTFs_pca);
        rank_corr(subject_ind,zscore_ind) = corr(loc,(1:length(loc))');
        best_HRTFs = sorted_HRTFs_pca(1:no_HRTFs_top);
        %          best_HRTF = sorted_HRTFs_pca(1);
        %          rand_HRTFs = sorted_HRTFs_pca(rand_ind(1:no_HRTFs_top));
        % number in excellent category
        ind_excellent_listen{subject_ind} = find(results_sorted_listen == 4);
        excellent_HRTFs_listen{subject_ind} = sorted_HRTFs_listen(ind_excellent_listen{subject_ind});
        %          excellent_HRTFs_pca{subject_ind} = sorted_HRTFs_pca(ind_excellent_listen↙
{subject_ind});
        %          excellent_count(subject_ind,zscore_ind) = 0;
        %          TF(subject_ind,zscore_ind) = ismember(best_HRTF,sorted_HRTFs_listen↙
(ind_excellent_listen{subject_ind}));
        % plot subjects in 2D to check
        if zscore_ind == 200
            figure(1)
            plot(points(:,1),points(:,2),'*')
            hold on
            ind = ismember(Names,excellent_HRTFs_listen{subject_ind});
            text(points(ind,1),points(ind,2),num2str(excellent_HRTFs_listen{subject_ind}),’Color’,’g’)
            ind = ismember(Names,good_HRTFs_listen{subject_ind});
            text(points(ind,1),points(ind,2),num2str(good_HRTFs_listen{subject_ind}),’Color’,’y’)
            ind = ismember(Names,bad_HRTFs_listen{subject_ind});
            text(points(ind,1),points(ind,2),num2str(bad_HRTFs_listen{subject_ind}),’Color’,’r’)
            plot(zscores_regression(current_subj_ind,1),zscores_regression(current_subj_ind,2),’m*’)
            plot(zscores_plot(current_subj_ind,1),zscores_plot(current_subj_ind,2),’k*’)
            sorted_HRTFs_pca
            results_sorted_pca
            pause
            cla
        end
        ind_anova = [];
        if use_rank == 1
            ind_anova = find(ismember(sorted_HRTFs_pca,excellent_HRTFs_listen{subject_ind}))’;
        else
            ind_anova = results_sorted_pca(ismember(sorted_HRTFs_pca,excellent_HRTFs_listen↙
{subject_ind}))’;
        end
        ind_anova_cat = cat(2,ind_anova_cat,ind_anova);
        ind_pcs = cat(2,ind_pcs,zscore_ind*ones(size(ind_anova)));
        for group_ind = 1:length(ind_anova)
            group{subject_ind,offset + group_ind} = ’Excellent’;
        end
        offset = length(ind_anova_cat);
        %         for dd = 1:length(excellent_HRTFs_listen{subject_ind})
        %             ind_excellent_pca{zscore_ind,subject_ind}(dd) = find(sorted_HRTFs_pca ==↙
excellent_HRTFs_listen{subject_ind}(dd));
        %             ind_excellent_count = excellent_HRTFs_pca{subject_ind} == excellent_HRTFs_listen↙
{subject_ind}(dd);
        %             if ~any(ind_excellent_count) == 0
        %                 excellent_count(subject_ind,zscore_ind) = excellent_count(subject_ind,↙
zscore_ind) + 1;
        %             end
        %         end
        % number in good category
        ind_good{subject_ind} = find(results_sorted_listen == 3);
        good_HRTFs_listen{subject_ind} = sorted_HRTFs_listen(ind_good{subject_ind});
        %         good_HRTFs_pca{subject_ind} = sorted_HRTFs_pca(ind_good{subject_ind});
        %         good_count(subject_ind,zscore_ind) = 0;
        if OK == 1
            ind_anova = [];
            if use_rank == 1
                ind_anova = find(ismember(sorted_HRTFs_pca,good_HRTFs_listen{subject_ind}))’;
            else
                ind_anova = results_sorted_pca(ismember(sorted_HRTFs_pca,good_HRTFs_listen↙
{subject_ind}))’;
            end
            ind_anova_cat = cat(2,ind_anova_cat,ind_anova);
            ind_pcs = cat(2,ind_pcs,zscore_ind*ones(size(ind_anova)));
            for group_ind = 1:length(ind_anova)
                group{subject_ind,offset + group_ind} = ’OK’;
            end
            offset = length(ind_anova_cat);
        end
        %         for dd = 1:length(good_HRTFs_listen{subject_ind})
        %             ind_good_pca{zscore_ind,subject_ind}(dd) = find(sorted_HRTFs_pca ==↙
good_HRTFs_listen{subject_ind}(dd));
        %             ind_good_count = good_HRTFs_pca{subject_ind} == good_HRTFs_listen{subject_ind}↙
(dd);
        %             if ~any(ind_good_count) == 0
        %                 good_count(subject_ind,zscore_ind) = good_count(subject_ind,zscore_ind) + 1;
        %             end
        %         end
        % number in bad category
        ind_bad{subject_ind} = find(results_sorted_listen == 2);
        bad_HRTFs_listen{subject_ind} = sorted_HRTFs_listen(ind_bad{subject_ind});
        %         bad_HRTFs_pca = sorted_HRTFs_pca(ind_bad{subject_ind});
        %         bad_count(subject_ind,zscore_ind) = 0;
        ind_anova = [];
        if use_rank == 1
            ind_anova = find(ismember(sorted_HRTFs_pca,bad_HRTFs_listen{subject_ind}))’;
        else
            ind_anova = results_sorted_pca(ismember(sorted_HRTFs_pca,bad_HRTFs_listen{subject_ind}))’;
        end
        ind_anova_cat = cat(2,ind_anova_cat,ind_anova);
        ind_pcs = cat(2,ind_pcs,zscore_ind*ones(size(ind_anova)));
        for group_ind = 1:length(ind_anova)
            group{subject_ind,offset + group_ind} = ’Bad’;
        end
        offset = length(ind_anova_cat);
        %         for dd = 1:length(bad_HRTFs_listen{subject_ind})
        %             ind_bad_pca{zscore_ind,subject_ind}(dd) = find(sorted_HRTFs_pca ==↙
bad_HRTFs_listen{subject_ind}(dd));
        %             ind_bad_count = bad_HRTFs_pca == bad_HRTFs_listen{subject_ind}(dd);
        %             if ~any(ind_bad_count) == 0
        %                   bad_count(subject_ind,zscore_ind) = bad_count(subject_ind,zscore_ind) + 1;
        %             end
        %         end
        % percentage of excellent, good and bad in top number of HRTFs
        no_excellent = 0;
        no_good = 0;
        no_bad = 0;
        for i = 1:length(best_HRTFs)
            if ~any(best_HRTFs(i) == excellent_HRTFs_listen{subject_ind}) == 0
                no_excellent = no_excellent + 1;
                if plot_top_HRTFs == 1
                    axes(ah_top)
                    text(1,i,num2str(best_HRTFs↙
(i)),'color','g','HorizontalAlignment','center','FontSize',20)
                end
            end
            if ~any(best_HRTFs(i) == good_HRTFs_listen{subject_ind}) == 0
                no_good = no_good + 1;
                if plot_top_HRTFs == 1
                    axes(ah_top)
                    text(1,i,num2str(best_HRTFs↙
(i)),'color','y','HorizontalAlignment','center','FontSize',20)
                end
            end
            if ~any(best_HRTFs(i) == bad_HRTFs_listen{subject_ind}) == 0
                no_bad = no_bad + 1;
                if plot_top_HRTFs == 1
                    axes(ah_top)
                    text(1,i,num2str(best_HRTFs↙
(i)),'color','r','HorizontalAlignment','center','FontSize',20)
                end
            end
        end
        percent_excellent(subject_ind,zscore_ind) = (no_excellent/length(best_HRTFs))*100;
        percent_good(subject_ind,zscore_ind) = (no_good/length(best_HRTFs))*100;
        percent_bad(subject_ind,zscore_ind) = (no_bad/length(best_HRTFs))*100;
        %         % random
        %         no_excellent_rand = 0;
        %         no_good_rand = 0;
        %         no_bad_rand = 0;
        %         for i = 1:length(rand_HRTFs)
        %             if ~any(rand_HRTFs(i) == excellent_HRTFs_listen{subject_ind}) == 0
        %                 no_excellent_rand = no_excellent_rand + 1;
        %                 if plot_top_HRTFs == 1
        %                     axes(ah_top)
        %                     text(2,i,num2str(rand_HRTFs↙
(i)),'color','g','HorizontalAlignment','center','FontSize',20)
        %                 end
        %             end
        %             if ~any(rand_HRTFs(i) == good_HRTFs_listen{subject_ind}) == 0
        %                 no_good_rand = no_good_rand + 1;
        %                 if plot_top_HRTFs == 1
        %                     axes(ah_top)
        %                     text(2,i,num2str(rand_HRTFs↙
(i)),'color','y','HorizontalAlignment','center','FontSize',20)
        %                 end
        %             end
        %             if ~any(rand_HRTFs(i) == bad_HRTFs_listen{subject_ind}) == 0
        %                 no_bad_rand = no_bad_rand + 1;
        %                 if plot_top_HRTFs == 1
        %                     axes(ah_top)
        %                     text(2,i,num2str(rand_HRTFs↙
(i)),'color','r','HorizontalAlignment','center','FontSize',20)
        %                 end
        %             end
        %         end
        %         if plot_top_HRTFs == 1
        %             set(ah_top,'XLim',[0 3],'YLim',[0 length(best_HRTFs)+1],'XTick',[1↙
2],'XTickLabel',{'Model' 'Random'})
        %             title(['subject ',subjects_C{subject_ind},' with ',num2str(zscore_ind),' out of↙
',num2str(size(subj_coord,2)),'basis functions used'])
        %             pause
        %             cla(ah_top)
        %         end
        %         percent_excellent_rand(subject_ind,zscore_ind) = (no_excellent_rand/length↙
(rand_HRTFs))*100;
        %         percent_good_rand(subject_ind,zscore_ind) = (no_good_rand/length(rand_HRTFs))*100;
        %         percent_bad_rand(subject_ind,zscore_ind) = (no_bad_rand/length(rand_HRTFs))*100;
    end
    % plot distances
    if compare_regression == 1
        fh_c = figure('Position',[1 900 1440 900]);
        ah_c = axes;
        bar(distance_test)
        pause
        close(fh_c)
    end
    percent_excellent_all(subject_ind) = (length(excellent_HRTFs_listen{subject_ind})/length↙
(sorted_HRTFs_listen))*100;
    percent_good_all(subject_ind) = (length(good_HRTFs_listen{subject_ind})/length↙
(sorted_HRTFs_listen))*100;
    percent_bad_all(subject_ind) = (length(bad_HRTFs_listen{subject_ind})/length(sorted_HRTFs_listen))↙
*100;
    if listen_subjects(subject_ind) == 1013
        ind_anova_cat = nan(1,size(anova_values,2));
        ind_pcs = nan(1,size(ind_all,2));
    end
    anova_values(subject_ind,:) = ind_anova_cat;
    ind_all(subject_ind,:) = ind_pcs;
end
if plot_position == 1 && use_regression == 1
    max(test_dist)
end
if save_percentages == 1
    save('subject_percentages','percent_excellent','percent_good','percent_bad',...
        'percent_excellent_all','percent_good_all','percent_bad_all')
end
mean_rank_corr = mean(rank_corr);
max_zscore = max(max(ind_all));
if show_waitbar == 1
    close(h_all)
end
for zscore_ind = 1:max_zscore
    [r,c] = find(ind_all == zscore_ind);
    for ii = 1:length(c)
        anova_values_per_pc(ii) = anova_values(r(ii),c(ii));
        group_per_pc{ii} = group{r(ii),c(ii)};
    end
    rank_data(zscore_ind).data = {anova_values_per_pc(strcmp('Excellent',group_per_pc)); ...
        anova_values_per_pc(strcmp('OK',group_per_pc)); ...
        anova_values_per_pc(strcmp('Bad',group_per_pc))};
    % standard deviation
    TF_excellent = strcmp('Excellent',group_per_pc);
    all_excellent_rankings = anova_values_per_pc(TF_excellent);
    %     std_excellent(zscore_ind) = std(all_excellent_rankings);
    %     mean_rank_excellent(zscore_ind) = mean(all_excellent_rankings);
    TF_good = strcmp('OK',group_per_pc);
    all_good_rankings = anova_values_per_pc(TF_good);
    %     std_good(zscore_ind) = std(all_good_rankings);
    %     mean_rank_good(zscore_ind) = mean(all_good_rankings);
    TF_bad = strcmp('Bad',group_per_pc);
    all_bad_rankings = anova_values_per_pc(TF_bad);
    %     std_bad(zscore_ind) = std(all_bad_rankings);
    %     mean_rank_bad(zscore_ind) = mean(all_bad_rankings);
    % anova
    if plot_anova == 1
        if zscore_ind == max_zscore
            if use_rank == 1
                [p(zscore_ind),stats] = kruskalwallis(anova_values_per_pc,group_per_pc);
                Chisq(zscore_ind) = stats{2,5};
            else
                [p(zscore_ind),table,stats] = anova1(anova_values_per_pc,group_per_pc);
                Chisq(zscore_ind) = table{2,5};
            end
            figure('Position',scrsz)
            distributionPlot({all_excellent_rankings all_good_rankings all_bad_rankings},...
                [],[],[],0,10)
            %               pause
            %               close all
        else
            if use_rank == 1
                [p(zscore_ind),stats] = kruskalwallis(anova_values_per_pc,group_per_pc,'off');
                Chisq(zscore_ind) = stats{2,5};
            else
                [p(zscore_ind),table,stats] = anova1(anova_values_per_pc,group_per_pc,'off');
                Chisq(zscore_ind) = table{2,5};
            end
        end
    else
        if use_rank == 1
            [p(zscore_ind),stats] = kruskalwallis(anova_values_per_pc,group_per_pc,'off');
            Chisq(zscore_ind) = stats{2,5};
        elseif use_rank == 2
            [r,p_value] = corrcoef(anova_values_per_pc,...
                (strcmp(group_per_pc,'Excellent')+2*strcmp(group_per_pc,'OK')+3*strcmp↙
(group_per_pc,'Bad')));
            Chisq(zscore_ind) = r(2,1);
            p(zscore_ind) = p_value(2,1);
        else
            [p(zscore_ind),table,stats] = anova1(anova_values_per_pc,group_per_pc,'off');
            Chisq(zscore_ind) = table{2,5};
        end
    end
    if excellent_good == 1
        % anova for excellent and good only
        data_excellent_good = cat(2,anova_values_per_pc(TF_excellent),...
            anova_values_per_pc(TF_good));
        group_excellent_good = cat(2,group_per_pc(TF_excellent),group_per_pc(TF_good));
        if plot_anova == 1
            if zscore_ind == max_zscore && subject_ind == length(Names)
                p_excellent_good(zscore_ind) = kruskalwallis(data_excellent_good,group_excellent_good);
                %               pause
                %               close all
            else
                p_excellent_good(zscore_ind) = kruskalwallis(data_excellent_good,↙
group_excellent_good,'off');
            end
        else
            p_excellent_good(zscore_ind) = kruskalwallis(data_excellent_good,↙
group_excellent_good,'off');
        end
    end
end
if test_each_dimension == 1
    [B,ranked_dimensions] = sort(Chisq,'descend');
end
if save_distributionPlot == 1
    max_Chisq = Chisq(end);
    max_Chisq_num_dim = length(Chisq);
    model_rank_data = rank_data(max_Chisq_num_dim).data;
    model_p = p(max_Chisq_num_dim);
    if use_regression == 1
        if test == 1
            regress_str = '_regression_test';
        else
            regress_str = '_regression';
        end
    else
        regress_str = [];
    end
    if use_rank == 1
        save([save_path 'rank_distr_model_' num2str(model) regress_str],'model_rank_data',...
            'model_p','max_Chisq','zscores_plot','max_Chisq_num_dim','model_name')
    else
        save([save_path 'dist_distr_model_' num2str(model) regress_str],'model_rank_data',...
            'model_p','max_Chisq','zscores_plot','max_Chisq_num_dim','model_name','ind_ms')
    end
    %     return
end
% for ICA Sydney 2010
if save_for_plot == 1
    % save distributions for plots
    save([save_path 'distribution_plot_pca_dtf'],'anova_values_per_pc','group_per_pc','p')
end
% top HRTFs
if top_HRTFs == 1 || golden_heads == 1 || coloured_model_plot == 1
    anova_ranked = 1;
    anova_random = 0;
    bar_graph_top_HRTFs = 0;
    bar_graph_ranked = 0;
    plot_anova_top_HRTFs = 0;
    % anova
    if anova_ranked == 1
        for aa = 1:size(percent_excellent,2)
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_excellent,2)
                    [h,p_excellent_top_HRTFs(aa)] = ttest2(percent_excellent(:,aa),↙
percent_excellent_all);
                    title(gca,'Excellent')
                else
                    [h,p_excellent_top_HRTFs(aa)] = ttest2(percent_excellent(:,aa),↙
percent_excellent_all);
                end
            else
                [h,p_excellent_top_HRTFs(aa)] = ttest2(percent_excellent(:,aa),percent_excellent_all);
            end
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_excellent,2)
                    [h,p_good_top_HRTFs(aa)] = ttest2(percent_good(:,aa),percent_good_all);
                    title(gca,'OK')
                else
                    [h,p_good_top_HRTFs(aa)] = ttest2(percent_good(:,aa),percent_good_all);
                end
            else
                [h,p_good_top_HRTFs(aa)] = ttest2(percent_good(:,aa),percent_good_all);
            end
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_bad,2)
                    [h,p_bad_top_HRTFs(aa)] = ttest2(percent_bad(:,aa),percent_bad_all);
                    title(gca,'Bad')
                else
                    [h,p_bad_top_HRTFs(aa)] = ttest2(percent_bad(:,aa),percent_bad_all);
                end
            else
                [h,p_bad_top_HRTFs(aa)] = ttest2(percent_bad(:,aa),percent_bad_all);
            end
        end
        % choose number of PCs for best results on p_excellent_top_HRTFs
        [value,number_pcs] = min(p_excellent_top_HRTFs);
        percent_best_HRTF_correct = 100*mean(TF(:,number_pcs));
    end
    % randomly chosen best HRTFs
    if anova_random == 1
        excellent_anova_data_all = [];
        excellent_group_percent_all = [];
        good_anova_data_all = [];
        good_group_percent_all = [];
        bad_anova_data_all = [];
        bad_group_percent_all = [];
        for aa = 1:size(percent_excellent,2)
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_excellent,2)
                    [h,p_excellent_rand(aa)] = ttest2(percent_excellent(:,aa),percent_excellent_rand(:,↙
aa));
                    title(gca,'Excellent Random')
                else
                    [h,p_excellent_rand(aa)] = ttest2(percent_excellent(:,aa),percent_excellent_rand(:,↙
aa));
                end
            else
                [h,p_excellent_rand(aa)] = ttest2(percent_excellent(:,aa),percent_excellent_rand(:,↙
aa));
            end
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_excellent,2)
                    [h,p_good_rand(aa)] = ttest2(percent_good(:,aa),percent_good_rand(:,aa));
                    title(gca,'OK Random')
                else
                    [h,p_good_rand(aa)] = ttest2(percent_good(:,aa),percent_good_rand(:,aa));
                end
            else
                [h,p_good_rand(aa)] = ttest2(percent_good(:,aa),percent_good_rand(:,aa));
            end
            if plot_anova_top_HRTFs == 1
                if aa == size(percent_excellent,2)
                    [h,p_bad_rand(aa)] = ttest2(percent_bad(:,aa),percent_bad_rand(:,aa));
                    title(gca,'Bad Random')
                else
                    [h,p_bad_rand(aa)] = ttest2(percent_bad(:,aa),percent_bad_rand(:,aa));
                end
            else
                [h,p_bad_rand(aa)] = ttest2(percent_bad(:,aa),percent_bad_rand(:,aa));
            end
        end
    end
    % percentage of excellent, good and bad in top HRTFs from model
    if bar_graph_top_HRTFs == 1
        data_top_HRTFs_model = cat(2,percent_excellent(:,number_pcs),percent_good(:,number_pcs),...
            percent_bad(:,number_pcs));
        %         figure('Position',[1 900 1440 900]);
        %         bar(data,'group')
        %         legend('Excellent','OK','Bad')
    end
    for qq = 1:size(percent_excellent,2)
        mean_percent_excellent(qq) = mean(percent_excellent(:,qq));
        mean_percent_good(qq) = mean(percent_good(:,qq));
        mean_percent_bad(qq) = mean(percent_bad(:,qq));
    end
    for qq = 1:size(percent_excellent_rand,2)
        mean_percent_excellent_rand(qq) = mean(percent_excellent_rand(:,qq));
        mean_percent_good_rand(qq) = mean(percent_good_rand(:,qq));
        mean_percent_bad_rand(qq) = mean(percent_bad_rand(:,qq));
    end
    percent_excellent_all_by_subject = percent_excellent_all;
    mean_percent_excellent_by_subject = mean(percent_excellent_all);
    percent_good_all_by_subject = percent_good_all;
    mean_percent_good_by_subject = mean(percent_good_all);
    percent_bad_all_by_subject = percent_bad_all;
    mean_percent_bad_by_subject = mean(percent_bad_all);
    if bar_graph_ranked == 1
        data_top_HRTFs_listening_test = cat(2,percent_excellent_all_by_subject',↙
percent_good_all_by_subject',...
            percent_bad_all_by_subject');
        offset = 0;
        for rr = 1:length(Names)
            data_for_bar_graph(3*(rr-1)+1,:) = data_top_HRTFs_listening_test(rr,:);
            data_for_bar_graph(3*(rr-1)+2,:) = data_top_HRTFs_model(rr,:);
            data_for_bar_graph(3*(rr-1)+3,:) = [0 0 0];
        end
        fh_bar = figure('Position',[1 900 1440 900]);
        ah_bar = axes;
        bar(data_for_bar_graph,'stack')
        xlim([0 151])
        set(ah_bar,'XTick',1.5:3:3*length(Names),'XTickLabel',1:1:length(Names),'FontSize',12)
        xlabel(ah_bar,'Subject')
        title(ah_bar,['Percentage of Excellent, OK and Bad ratings in top ' num2str(no_HRTFs_top) '↙
HRTFs'])
        legend(ah_bar,'Excellent','OK','Bad')
        % create box plot of percentage of excellent, good  and bad HRTFs
        % in top HRTFs with and without the model
        fh_box = figure('Position',[1 900 1440 900]);
        ah_box = axes;
        boxplot(ah_box,[data_top_HRTFs_listening_test(:,1) data_top_HRTFs_model(:,1)↙
data_top_HRTFs_listening_test(:,2) data_top_HRTFs_model(:,2) data_top_HRTFs_listening_test(:,3)↙
data_top_HRTFs_model(:,3)],...
            'notch','on','labels',{'Excellent','Excellent with model','OK','OK with model','Bad','Bad↙
with model'})
        title(ah_box,['Percentage of Excellent, OK and Bad ratings in top ' num2str(no_HRTFs_top) '↙
HRTFs'])
```

```
    end
end
if plot_rank_all == 1
    figure('Position',[1 900 1440 900]);
    for subject_no = 1:length(Names)
        text(zeros(1,length(excellent_HRTFs_listen{subject_no})),ind_excellent_listen{subject_no},↙
num2str(excellent_HRTFs_listen{subject_no}),'color','g','HorizontalAlignment','center','FontSize',8)
        hold on
        text(zeros(1,length(good_HRTFs_listen{subject_no})),ind_good{subject_no},num2str↙
(good_HRTFs_listen{subject_no}),'color','y','HorizontalAlignment','center','FontSize',8)
        text(zeros(1,length(bad_HRTFs_listen{subject_no})),ind_bad{subject_no},num2str(bad_HRTFs_listen↙
{subject_no}),'color','r','HorizontalAlignment','center','FontSize',8)
        for aa = 1:size(ind_excellent_pca,1)
            text(aa*ones(1,length(ind_excellent_pca{aa,subject_no})),ind_excellent_pca{aa,subject_no},↙
num2str(excellent_HRTFs_listen{subject_no}),'color','g','HorizontalAlignment','center','FontSize',8)
            text(aa*ones(1,length(ind_good_pca{aa,subject_no})),ind_good_pca{aa,subject_no},num2str↙
(good_HRTFs_listen{subject_no}),'color','y','HorizontalAlignment','center','FontSize',8)
            text(aa*ones(1,length(ind_bad_pca{aa,subject_no})),ind_bad_pca{aa,subject_no},num2str↙
(bad_HRTFs_listen{subject_no}),'color','r','HorizontalAlignment','center','FontSize',8)
        end
        xlim([-1 aa+6])
        ylim([0 length(Names)])
        title(['Subject ',num2str(Names(subject_no))])
        xlabel('Number of basis functions')
        ylabel('HRTF subject ranking')
        plot(1,50,'*g')
        plot(1,50,'*y')
        plot(1,50,'*r')
        legend(gca,'Excellent','OK','Bad')
        pause
        cla
    end
    mean(excellent_count)
    mean(good_count)
    mean(bad_count)
end
% generate plot, using the number of basis functions that gives the most
% statistically significant result, which shows the model ranking of each
% HRTF with the 'Golden Heads' included in bold
if golden_heads == 1
    figure('Position',[1 900 1440 900]);
    hold on
    golden_head_HRTFs = []; % removed numbers
    colours = colormap(hsv(length(golden_head_HRTFs)));
    for subject_no = 1:length(Names)
        ranked_HRTFs = [];
        ranked_HRTFs(ind_excellent_pca{number_pcs,subject_no}) = excellent_HRTFs_listen{subject_no};
        ranked_HRTFs(ind_good_pca{number_pcs,subject_no}) = good_HRTFs_listen{subject_no};
        ranked_HRTFs(ind_bad_pca{number_pcs,subject_no}) = bad_HRTFs_listen{subject_no};
        golden_head_ind = [];
        for qq = 1:length(golden_head_HRTFs)
            if ~any(ranked_HRTFs == golden_head_HRTFs(qq)) == 1
                golden_head_ind(qq) = 100;
            else
                golden_head_ind(qq) = find(ranked_HRTFs == golden_head_HRTFs(qq));
            end
        end
        % remove the points that are golden heads
        ind_excellent = ind_excellent_pca{number_pcs,subject_no};
        ind_good = ind_good_pca{number_pcs,subject_no};
        ind_bad = ind_bad_pca{number_pcs,subject_no};
        for ff = 1:length(golden_head_HRTFs)
            ind_remove = find(ind_excellent == golden_head_ind(ff),1);
            if isempty(ind_remove) == 0
                ind_excellent(ind_remove) = [];
                golden_head_colours(ff) = 'g';
            end
            ind_remove = find(ind_good == golden_head_ind(ff),1);
            if isempty(ind_remove) == 0
                ind_good(ind_remove) = [];
                golden_head_colours(ff) = 'y';
            end
            ind_remove = find(ind_bad == golden_head_ind(ff),1);
            if isempty(ind_remove) == 0
                ind_bad(ind_remove) = [];
                golden_head_colours(ff) = 'r';
            end
        end
        plot(subject_no*ones(1,length(ind_excellent)),ind_excellent,'*g')
        plot(subject_no*ones(1,length(ind_good)),ind_good,'*y')
        plot(subject_no*ones(1,length(ind_bad)),ind_bad,'*r')
        for qq = 1:length(golden_head_HRTFs)
            text(subject_no,golden_head_ind(qq),num2str(golden_head_HRTFs(qq)),'Color',↙
golden_head_colours(qq),'HorizontalAlignment','center','FontSize',8)
        end
    end
    xlim([-1 length(Names)+6])
    ylim([0 length(Names)])
    set(gca,'XTick',1:1:length(Names),'XTickLabel',1:1:length(listen_subjects),'FontSize',10)
    title('Model predicting rank of rated HRTFs with Golden Heads','FontSize',14)
    xlabel('Subject','FontSize',12)
    ylabel('HRTF subject ranking','FontSize',12)
    % plot some points that are out of view and give legend for those
    plot(1,100,'*g')
    plot(1,100,'*y')
    plot(1,100,'*r')
    legend(gca,'Excellent','OK','Bad')
end
% make a plot of ranked HRTFs using best model
if plot_ranked_HRTFs_best_model == 1
    figure('Position',[1 900 1440 900]);
    hold on
    for subject_no = 1:length(Names)
        text(subject_no*ones(1,length(ind_excellent_pca{number_pcs,subject_no})),ind_excellent_pca↙
{number_pcs,subject_no},num2str(excellent_HRTFs_listen↙
{subject_no}),'color','g','HorizontalAlignment','center','FontSize',8)
        text(subject_no*ones(1,length(ind_good_pca{number_pcs,subject_no})),ind_good_pca{number_pcs,↙
subject_no},num2str(good_HRTFs_listen↙
{subject_no}),'color','y','HorizontalAlignment','center','FontSize',8)
        text(subject_no*ones(1,length(ind_bad_pca{number_pcs,subject_no})),ind_bad_pca{number_pcs,↙
subject_no},num2str(bad_HRTFs_listen↙
{subject_no}),'color','r','HorizontalAlignment','center','FontSize',8)
    end
    xlim([-1 length(Names)+6])
    ylim([0 length(Names)])
    set(gca,'XTick',1:1:length(Names),'XTickLabel',1:1:length(listen_subjects),'FontSize',10)
    title('Ranking for best model for each subject','FontSize',14)
    xlabel('Number of basis functions','FontSize',12)
    ylabel('HRTF subject ranking','FontSize',12)
    plot(1,100,'*g')
    plot(1,100,'*y')
    plot(1,100,'*r')
    legend(gca,'Excellent','OK','Bad')
end
% plot in 3D plot the ranked HRTFs and label then as rated as either 'excellent',
% 'good' or 'bad'
if coloured_model_plot == 1
    figure('Position',scrsz);
    for subject_no = 1:no_subjects
        stem3(zscores_plot(ind_excellent_pca{number_pcs,subject_no},1),zscores_plot(ind_excellent_pca↙
{number_pcs,subject_no},2),zscores_plot(ind_excellent_pca{number_pcs,subject_no},↙
3),'Color','g','BaseValue',min(zscores_plot(:,3)))
        hold on
        stem3(zscores_plot(ind_good_pca{number_pcs,subject_no},1),zscores_plot(ind_good_pca{number_pcs,↙
subject_no},2),zscores_plot(ind_good_pca{number_pcs,subject_no},3),'Color',[1 0.6 0],'BaseValue',min↙
(zscores_plot(:,3)))
        stem3(zscores_plot(ind_bad_pca{number_pcs,subject_no},1),zscores_plot(ind_bad_pca{number_pcs,↙
subject_no},2),zscores_plot(ind_bad_pca{number_pcs,subject_no},3),'Color','r','BaseValue',min↙
(zscores_plot(:,3)))
        % plot on the x-y plane
        plot3(zscores_plot(ind_excellent_pca{number_pcs,subject_no},1),zscores_plot(ind_excellent_pca↙
{number_pcs,subject_no},2),min(zscores_plot(:,3))*ones(size(zscores_plot(ind_excellent_pca{number_pcs,↙
subject_no},3))),'MarkerEdgeColor','k','MarkerSize',↙
10,'MarkerFaceColor','g','LineStyle','none','Marker','o')
        plot3(zscores_plot(ind_good_pca{number_pcs,subject_no},1),zscores_plot(ind_good_pca{number_pcs,↙
subject_no},2),min(zscores_plot(:,3))*ones(size(zscores_plot(ind_good_pca{number_pcs,subject_no},↙
3))),'MarkerEdgeColor','k','MarkerSize',10,'MarkerFaceColor',[1 0.6 0],'LineStyle','none','Marker','o')
        plot3(zscores_plot(ind_bad_pca{number_pcs,subject_no},1),zscores_plot(ind_bad_pca{number_pcs,↙
subject_no},2),min(zscores_plot(:,3))*ones(size(zscores_plot(ind_bad_pca{number_pcs,subject_no},↙
3))),'MarkerEdgeColor','k','MarkerSize',10,'MarkerFaceColor','r','LineStyle','none','Marker','o')
        axis equal
        grid on
        title('Ranking for best model for each subject','FontSize',14)
        xlabel('First principal component','FontSize',12)
        ylabel('Second principal component','FontSize',12)
        zlabel('Third principal component','FontSize',12)
        legend(gca,'Excellent','OK','Bad')
        pause
        cla
    end
end
% this option does not work if the new IRCAM HRTFs and namely my HRTF
% (1063) is not included in the analysis
if rank_my_HRTF == 1
    % do ranking for my HRTF (1063)
    listening_test_list = []; % removed numbers
    sorted_HRTFs_pca = [];
    sort_ind_pca = [];
    for aa = 1:length(subjects_C)-2
        points = zscores(:,1:aa+1); % starts at 2D
        distances = pdist(points)';
        % find ranking from model
        current_HRTF_all_ind = find(listen_subjects == 1063);
        Names_list = listen_subjects';
        Names_list(current_HRTF_all_ind) = [];
        subj_choose_ind = find(choose_values == current_HRTF_all_ind);
        ind_col_1 = subj_choose_ind <= length(choose_values);
        ind_col_2 = subj_choose_ind > length(choose_values);
        subj_choose_ind_col_1 = subj_choose_ind(ind_col_1);
        subj_choose_ind_col_2 = subj_choose_ind(ind_col_2) - length(choose_values)*ones(length(find↙
(ind_col_2)),1);
        choose_subjects_ind = choose_values(subj_choose_ind_col_1,2);
        choose_subjects_ind = [choose_subjects_ind; choose_values(subj_choose_ind_col_2,1)];
        choose_subjects = listen_subjects(choose_subjects_ind);
        distances_subj = distances(subj_choose_ind_col_1);
        distances_subj = [distances_subj; distances(subj_choose_ind_col_2)];
        [results_sorted_pca,sort_ind_pca(:,aa)] = sort(distances_subj,'ascend');
        sorted_HRTFs_pca(:,aa) = choose_subjects(sort_ind_pca(:,aa));
        for qq = 2:length(listening_test_list)
            ind_listening_test_HRTFs(qq,aa) = find(sorted_HRTFs_pca(:,aa) == listening_test_list(qq));
        end
    end
    figure('Position',[1 900 1440 900]);
    for aa = 1:size(sorted_HRTFs_pca,2)
        text(aa+1*ones(1,length(sorted_HRTFs_pca(:,aa))),1:1:length(sorted_HRTFs_pca(:,aa)),num2str↙
(sorted_HRTFs_pca(:,aa)),'color','k','HorizontalAlignment','center','FontSize',8)
        hold on
        %     text(aa,ind_listening_test_HRTFs(1,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs↙
(1,aa),aa)),'color','c','HorizontalAlignment','center','FontSize',8)
        text(aa+1,ind_listening_test_HRTFs(2,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs(2,↙
aa),aa)),'color','g','HorizontalAlignment','center','FontSize',8)
        text(aa+1,ind_listening_test_HRTFs(3,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs(3,↙
aa),aa)),'color','y','HorizontalAlignment','center','FontSize',8)
        text(aa+1,ind_listening_test_HRTFs(4,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs(4,↙
aa),aa)),'color','m','HorizontalAlignment','center','FontSize',8)
        text(aa+1,ind_listening_test_HRTFs(5,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs(5,↙
aa),aa)),'color','r','HorizontalAlignment','center','FontSize',8)
        text(aa+1,ind_listening_test_HRTFs(6,aa),num2str(sorted_HRTFs_pca(ind_listening_test_HRTFs(6,↙
aa),aa)),'color','b','HorizontalAlignment','center','FontSize',8)
    end
    xlim([1 aa+6])
    ylim([0 length(listen_subjects)])
    title('PCA model with listening test HRTF ranking')
    xlabel('Number of basis functions')
    ylabel('HRTF subject ranking')
    plot(1,100,'*g')
    plot(1,100,'*y')
    plot(1,100,'*m')
    plot(1,100,'*r')
    plot(1,100,'*b')
    legend(gca,[]) % removed numbers
end
```

BIBLIOGRAPHY

D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004. (Cited on pages 7 and 26.)

V. R. Algazi. Structural composition and decomposition of HRTFs. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 103–106, New Platz, NY , USA, October 2001. (Cited on page 32.)

V. R. Algazi, C. Avendano, and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *Journal of the Acoustical Society of America*, 109(3):1110–1122, 2001a. (Cited on pages 27 and 31.)

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, New Platz, NY ,USA, October 2001b. (Cited on pages xxii, 127, and 128.)

V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. H. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *Journal of the Acoustical Society of America*, 112(5):2053–2064, 2002. (Cited on page 32.)

A. Alves-Pinto and E. A. Lopez-Poveda. Detection of high-frequency spectral notches as a function of level. *Journal of the Acoustical Society of America*, 118 (4):2458–2469, 2005. (Cited on page 158.)

American National Standards Inst. ANSI S1.11-2004 (R2009) octave-band and fractional octave-band analog and digital filters. (American National Standards Inst., New York), 2004. (Cited on page 131.)

S. R. Arnott, M. A. Binns, C. L. Grady, and C. Alain. Assessing the auditory dual-pathway model in humans. *Neuroimage*, 22(1):401–408, May 2004. (Cited on page 7.)

F. Asano, Y. Suzuki, and T. Sone. Role of spectral cues in median plane localization. *Journal of the Acoustical Society of America*, 88(1):159–168, 1990. (Cited on pages 24 and 29.)

M. R. Bai and K. Y. Ou. Head-related transfer function (HRTF) synthesis based on a three-dimensional array model and singular value decomposition. *Journal of Sound and Vibration*, 281(3-5):1093–1115, 2005. (Cited on page 129.)

S. Bech. Selection and training of subjects for listening tests on sound-reproducing equipment. *Journal of the Audio Engineering Society*, 40(7-8): 590–610, 1992. (Cited on pages 106, 113, 116, and 142.)

J. Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, 1977. (Cited on page 86.)

J. Berge. How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality? In *Poceedings of Spatial Audio And Sensory Evaluation Techniques*, Guildford, UK, April 6-7 2006. (Cited on page 98.)

V. Best, S. Carlile, C. Jin, and A. van Schaik. The role of high frequencies in speech localization. *The Journal of the Acoustical Society of America*, 118(1): 353–363, 2005. (Cited on page 106.)

J. Blauert. Sound localization in median plane. *Acustica*, 22(4):205–213, 1969. (Cited on pages 27, 30, 137, and 158.)

J. Blauert. *Spatial Hearing: the Psychophysics of Human Sound Localization*. Massachusetts Institute of Technology, London, England, 1997. (Cited on pages 8, 10, 15, 32, and 71.)

P. J. Bloom. Creating source elevation illusions by spectral manipulation. *Journal of the Audio Engineering Society*, 25(9):560–565, 1977. (Cited on pages 30 and 158.)

B. Boren and A. Roginska. The effects of headphones on listener hrtf preference. In *Proceedings of the 131st Convention of the Audio Engineering Society*, New York, NY, USA, October 2011. (Cited on pages 71 and 116.)

B. Boser, I. Guyon, and V. Vapnik. An training algorithm for optimal margin classifiers. In *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, United States, 1992. (Cited on page 173.)

A. W. Bronkhorst. Localization of real and virtual sound sources. *Journal of the Acoustical Society of America*, 98(5):2542–2553, 1995. (Cited on page 156.)

T. Brookes and C. Treble. The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings. In *Proceedings of the 118th Convention of the Audio Engineering Society*, 5 2005. (Cited on page 138.)

R. A. Butler. An analysis of the monaural displacement of sound in space. *Attention, Perception, & Psychophysics*, 41(1):1–7, 1987. (Cited on pages 31 and 137.)

R. A. Butler and K. Belendiuk. Spectral cues utilized in localization of sound in median sagittal plane. *Journal of the Acoustical Society of America*, 61(5): 1264–1269, 1977. (Cited on page 30.)

R. A. Butler and C.C. Helwig. The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. *American Journal of Otolaryngology*, 4(3):165–173, 1983. (Cited on page 137.)

R. A. Butler, R. A. Humanski, and A.D. Musicant. Binaural and monaural localization of sound in two-dimensional space. *Perception*, 19(2):241–256, 1990. (Cited on pages 28, 31, and 137.)

S. Carlile. *Virtual auditory space: Generation and applications*. RG Landes, 1996. (Cited on page 58.)

S. Carlile and D. Pralong. The location-dependent nature of perceptually salient features of the human head-related transfer-functions. *Journal of the Acoustical Society of America*, 95(6):3445–3459, 1994. (Cited on pages 12, 20, 24, 29, and 129.)

S. Carlile and D. Schönstein. Frequency bandwidth and multi-talker environments. In *Proceedings of the 120th Convention of the Audio Engineering Society*, Paris, France, May 2006a. (Cited on page 72.)

S. Carlile and D. Schönstein. Bandwidth and talker segregation in 3d virtual auditory displays. In *Proceedings of the 9th Western Pacific Acoustics Conference*, Seoul, Korea, 6 2006b. (Cited on page 72.)

S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hearing Research*, 114(1-2):179–196, 1997. (Cited on pages 24, 36, 58, and 59.)

S. Carlile, C. Jin, and V. Raad. Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors. In *International Symposium on Multimedia Information Processing*, Sydney, Australia, 2000. (Cited on pages 23 and 37.)

S. Carlile, R. Martin, and K. McAnally. Spectral information in sound localization. In Manuel S. Malmierca and Dexter R. F. Irvine, editors, *Auditory Spectral Processing*, volume 70 of *International Review of Neurobiology*, pages 399 – 434. Academic Press, 2005. (Cited on pages xvii, 11, 14, 29, 137, and 158.)

J. S. Chen, B. D. Vanveen, and K. E. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *Journal of the Acoustical Society of America*, 97(1):439–452, 1995. (Cited on pages 23 and 129.)

C. S. Choe, R. B. Welch, R. M. Gilford, and J. F. Juola. Ventriloquist effect - visual dominance or response bias. *Perception & Psychophysics*, 18(1):55–60, 1975. (Cited on pages 26 and 83.)

W. Chung, S. Carlile, and P. Leong. A performance adequate computational model for auditory localization. *The Journal of the Acoustical Society of America*, 107(1):432–445, 2000. (Cited on page 32.)

P. D. Coleman. Failure to localize the source distance of an unfamiliar sound. *The Journal of the Acoustical Society of America*, 34(3):345–346, 1962. (Cited on page 98.)

R. Conetta. Scaling and predicting spatial attributes of reproduced sound using an artificial listener. MPhil/PhD Upgrade Report, University of Surrey, Institute of Sound Recording, 2007. (Cited on page 85.)

W. P. J. de Bruijn and M. M. Boone. Subjective experiments on the effects of combining spatialized audio and 2D video projection in audio-visual systems. In *Proceedings of the 112th Audio Engineering Society Convention*, April 2002. (Cited on page 83.)

S. D. Erulkar. Comparative aspects of spatial localization of sound. *Physiological Reviews*, 52(1):238–360, 1972. (Cited on page 7.)

A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proceedings of the 108th Convention of the Audio Engineering Society*, 2 2000. (Cited on page 52.)

N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, October 1995. (Cited on page 64.)

N. I. Fisher and A. J. Lee. A correlation coefficient for circular data. *Biometrika*, 70(2):327–332, 1983. (Cited on page 64.)

N. I. Fisher, T. Lewis, and B. J. J. Embleton. *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge, 1987. (Cited on page 60.)

A. Gabrielsson. Statistical treatment of data from listening tests on sound reproducing systems. Rep. ta 92, Karolinska Institutet, Technical Audiology, Stockholm, Sweden, 1979. (Cited on pages 88 and 123.)

A. Gabrielsson, U. Rosenberg, and H. Sjogren. Judgments and dimension analyses of perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, 55(4):854–861, 1974. (Cited on pages 113, 116, and 123.)

M. B. Gardner and R. S. Gardner. Problem of localization in median plane - effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America*, 53(2):400–408, 1973. (Cited on pages 12, 33, and 36.)

B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, 1990. (Cited on page 29.)

R. Greff and B. F. G. Katz. Perceptual evaluation of hrtf notches versus peaks for vertical localisation. In *19th International Congress on Acoustics*, Madrid, Spain, 2-7 September 2007. (Cited on pages 30, 136, 156, and 158.)

T. D. Griffiths and J. D. Warren. What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887–892, 11 2004. (Cited on page 7.)

J. P. Guilford. *Psychometric methods.* New York, NY, US: McGraw-Hill, 1936. (Cited on pages 111 and 122.)

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, 2009. (Cited on pages xvii, 11, 33, 38, and 116.)

P. Guillon, R. Nicol, and L. Simon. Head-related transfer functions reconstruction from sparse measurements considering a priori knowledge from database analysis: A pattern recognition approach. In *Proceedings of the 125th Convention of the Audio Engineering Society*, October 2008. (Cited on page 38.)

R. S. Gunn. Support vector machines for classification and regression. Technical report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997. (Cited on page 174.)

I. Guyon. *Practical feature selection: from correlation to causality*. Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security. IOS Press, Amsterdam, The Netherlands, 2008. (Cited on page 148.)

I. Guyon, N. Matic, V. Vapnik, et al. Discovering informative patterns and data cleaning. *Advances in knowledge discovery and data mining*, 181:203, 1996. (Cited on page 173.)

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3): 389–422, 2002. (Cited on page 174.)

W. M. Hartmann and A. Wittenberg. On the externalization of sound images. *Journal of the Acoustical Society of America*, 99(6):3678–3688, 1996. ISSN 0001-4966. (Cited on page 26.)

K. Hartung, J. Braasch, and S.J. Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In *The Proceedings of the AES 16th International Conference: Spatial Sound Reproduction*, pages 319–329. Audio Engineering Society, 1999. (Cited on pages 23, 38, and 81.)

J. Hebrank and D. Wright. Spectral cues used in localization of sound sources on median plane. *Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974. (Cited on pages 27, 30, 33, 156, and 158.)

P. M. Hofman and A. J. Van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *Journal of the Acoustical Society of America*, 103(5):2634–2648, 1998. (Cited on pages 28, 32, 48, and 108.)

P. M. Hofman and A. J. Van Opstal. Binaural weighting of pinna cues in human sound localization. *Experimental Brain Research*, 148(4):458–470, 2003. (Cited on page 32.)

P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal. Relearning sound localization with new ears. *Nature Neuroscience*, 1(5):417–421, 1998. (Cited on pages 16 and 41.)

Hugeng, W. Wahab, and D. Gunawan. The effectiveness of chosen partial anthropometric measurements in individualizing head-related transfer functions on median plane. *J. of Inform. and Commun. Technol.*, 5(1):35–56, 2011. (Cited on pages 41 and 164.)

Hugeng, W. Wahab, and D. Gunawan. Enhanced Individualization of Head-Related Impulse Response Model in Horizontal Plane Based on Multiple Regression Analysis. In *Proceedings of the 2nd International Conference on Computer Engineering and Applications ICCEA*, pages 226–230, Bali Island, Indonesia, 19-21 March 2010. (Cited on page 41.)

R. A. Humanski and R. A. Butler. The contribution of the near and far ear toward localization of sound in the sagittal plane. *Journal of the Acoustical Society of America*, 83(6):2300–2310, 1988. (Cited on pages 30 and 158.)

J. Huopaniemi, N. Zacharov, and M. Karjalainen. Objective and subjective evaluation of head-related transfer function filter design. *Journal of the Audio Engineering Society*, 47(4):218–239, 1999. (Cited on page 75.)

S. Hwang and Y. Park. HRIR Customization in the Median Plane Via Principal Components Analysis. In *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*, 6 2007. (Cited on page 129.)

S. Hwang and Y. Park. Interpretations on principal components analysis of head-related impulse responses in the median plane. *Journal of the Acoustical Society of America*, 123(4):EL65–EL71, 2008. (Cited on page 129.)

K. Iida, M. Itoh, A. Itagaki, and M. Morimoto. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Applied Acoustics*, 68(8):835–850, 2007. (Cited on pages 30, 136, 156, and 158.)

K. Inanaga, Y. Yamada, and H. Koizumi. Headphone system with out-of-head localization applying dynamic hrtf (head-related transfer function). In *Proceedings of the 98th Convention of the Audio Engineering Society*, February 1995. (Cited on page 26.)

ISO/IEC IS 5497:1982:. Sensory analysis – methodology – guidelines for the preparation of samples for which direct sensory analysis is not feasible. Technical report, International Organization for Standardization, Geneva, Switzerland. (Cited on page 114.)

ISO/IEC IS 8253-1:1989. Acoustics - audiometric methods - part 1: Basic pure tone air and bone conduction threshold audiometry. Technical report, International Organization for Standardization, Geneva, Switzerland. (Cited on pages 80, 94, and 106.)

ISO/IEC IS 8586-2:1994. Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts. Technical report, International Organization for Standardization, Geneva, Switzerland. (Cited on pages xxv, 106, and 107.)

ITU-R BS.1116-1:1997. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Technical report, International Telecommunication Union, Geneva, Switzerland. (Cited on pages 25, 84, 91, and 97.)

ITU-R BS.1534-1:2003. Method for the subjective assessment of intermediate quality levels of coding systems. Technical report, International Telecommunication Union, Geneva, Switzerland. (Cited on pages 25, 91, and 111.)

Y. Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. *Acoustical Science and Technology*, 27(6):340–343, 2006. (Cited on pages 40 and 75.)

C. Jin. *Spectral Analysis and Resolving Spatial Ambiguities in Human Sound Localization*. PhD thesis, The University of Sydney, 2001. (Cited on pages 28 and 31.)

C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile. Enabling individualized virtual auditory space using morphological measurements. In *IEEE International Conference on Multimedia Information Processing*, volume Proceedings of the First IEEE Pacific-Rim Conference on Multimedia, 2000. (Cited on pages 41, 130, and 139.)

C. Jin, A. Corderoy, S. Carlile, and A. van Schaik. Contrasting monaural and interaural spectral cues for human sound localization. *Journal of the Acoustical Society of America*, 115(6):3124–3141, 2004. (Cited on pages 12, 59, 62, 106, and 116.)

Y. Kahana and P. A. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of Sound and Vibration*, 300(3-5):552–579, 2007. (Cited on pages 42 and 182.)

Y. Kahana and P.A. Nelson. Spatial acoustic mode shapes of the human pinna. In *Proceedings of the 109th Convention of the Audio Engineering Society*, Los Angeles, USA, 2000. Audio Engineering Society; 1999. (Cited on page 33.)

B. F. G. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *Journal of the Acoustical Society of America*, 110(5):2440–2448, 2001. (Cited on pages 42, 116, and 182.)

B. F. G. Katz and G. Parseihian. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, 131(2):EL99–EL105, 2012. (Cited on pages 24, 35, and 93.)

B. F. G. Katz, R. Nicol, and S. Busson. Subjective investigations of the interaural time difference in the horizontal plane. In *Proceedings of the 118th Convention of the Audio Engineering Society*, May 2005. (Cited on pages 11 and 19.)

B.F.G. Katz, E. Rio, and L. Picinali. LIMSI Spatialisation Engine, 2010. International Deposit Digital Number IDDN.FR.001.340014.000.S.P.2010.000.31235. (Cited on page 80.)

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, 1997. (Cited on pages 27, 31, and 156.)

R. E. Kirk. Learning, a major factor influencing preferences for high-fidelity systems. *Journal of the Audio Engineering Society*, 5(4):238–241, 1957. (Cited on page 85.)

D. J. Kistler and F. L. Wightman. A model of head-related transfer-functions based on principal components-analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91(3):1637–1647, 1992. (Cited on pages 19, 129, 130, 140, 151, and 157.)

R. Kohavi and R. Quinlan. Decision tree discovery. In *in Handbook of Data Mining and Knowledge Discovery*. Citeseer, 1999. (Cited on page 167.)

A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998. (Cited on pages 20, 29, and 116.)

A. Kulkarni and H. S. Colburn. Variability in the characterization of the headphone transfer-function. *Journal of the Acoustical Society of America*, 107(2):1071–1074, 2000. (Cited on pages 21 and 22.)

A. Kulkarni, S. K. Isabelle, and H. S. Colburn. Sensitivity of human subjects to head-related transfer-function phase spectra. *Journal of the Acoustical Society of America*, 105(5):2821–2840, 1999. (Cited on pages 10 and 19.)

E. H. A. Langendijk and A. W. Bronkhorst. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *Journal of the Acoustical Society of America*, 107(1):528–537, 2000. (Cited on page 23.)

E. H. A. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *Journal of the Acoustical Society of America*, 112 (4):1583–1596, 2002. (Cited on pages 27, 31, 71, 116, 156, and 157.)

H. T. Lawless and H. Heymann. *Sensory evaluation of food: principles and practices*. Aspen Publishers, Gaithersburg, MD, 1999. (Cited on page 122.)

P. Leong and S. Carlile. Methods for spherical data analysis and visualization. *Journal of Neuroscience Methods*, 80(2):191–200, 1998. (Cited on pages 58, 59, 60, and 108.)

J. Leung and S. Carlile. PCA Compression of HRTFs and localization performance. In *International Workshop on the Principles and Applications of Spatial Hearing*, Zao, Miyagi, Japan, 11-13 November 2009. (Cited on pages 130, 140, 141, and 155.)

H. Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B):467–477, 1971. (Cited on page 48.)

S. P. Lipshitz and J. Vanderkooy. The great debate: Subjective evaluation. In *Proceedings of the 65th Convention of the Audio Engineering Society*, 2 1980. (Cited on page 111.)

E. A. Lopez-Poveda and R. Meddis. Physical model of sound diffraction and reflections in the human concha. *Journal of the Acoustical Society of America*, 100(5):3248–3259, 1996. (Cited on pages 33 and 161.)

G. Lorho. Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction. In *Proceedings of the 119th Convention of the Audio Engineering Society*, 2005a. (Cited on page 96.)

G. Lorho. Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating. In *Proceedings of the 118th Convention of the Audio Engineering Society*, 2005b. (Cited on page 122.)

E. A. Macpherson and J. C. Middlebrooks. Localization of brief sounds: Effects of level and background noise. *Journal of the Acoustical Society of America*, 108(4):1834–1849, 2000. (Cited on pages 25, 28, 48, and 111.)

E. A. Macpherson and J. C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *Journal of the Acoustical Society of America*, 111(5, Part 1):2219–2236, 2002. (Cited on page 10.)

E. A. Macpherson and J. C. Middlebrooks. Vertical-plane sound localization probed with ripple-spectrum noise. *Journal of the Acoustical Society of America*, 114(1):430–445, 2003. (Cited on pages 20 and 29.)

E.A. Macpherson and W. Center. On the role of head-related transfer function spectral notches in the judgement of sound source elevation. In *Proceedings of the 2nd International Conference on Auditory Display*, pages 187–194, Sante Fe Institute, Sante Fe, 1994. (Cited on page 31.)

J. C. Makous and J. C. Middlebrooks. 2-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America*, 87(5):2188–2200, 1990. (Cited on pages 24, 59, and 66.)

D. Marston and A. Mason. Cascaded audio coding. Technical report, European Broadcasting Union Technical Report, 2005. (Cited on page 111.)

W. L. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proceedings of the International Computer Music Conference*, pages 274–281, 1987. (Cited on pages 116 and 129.)

W. L. Martens. Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis. *Acoustical Science and Technology*, 24(5):220–232, 2003. (Cited on pages 75 and 139.)

R. Martin and K. McAnally. Interpolation of head-related transfer functions. Technical report, Air Operations Division, Defence Science and Technology Organisation, 2007. (Cited on page 37.)

K. Matsui and A. Ando. Estimation of individualized head-related transfer function based on principal component analysis. *Acoustical Science and Technology*, 30(5):338–347, 2009. (Cited on page 75.)

V. V. Mattila and N. Zacharov. Generalized listener selection (GLS) procedure. In *Proceedings of the 110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, 2001. (Cited on page 105.)

K. I. McAnally and R. L. Martin. Variability in the headphone-to-ear-canal transfer function. *Journal of the Audio Engineering Society*, 50(4):263–266, 2002. (Cited on page 22.)

S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *The Journal of the Acoustical Society of America*, 61(6):1567–1576, 1977. (Cited on page 70.)

J. C. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *Journal of the Acoustical Society of America*, 92(5):2607–2624, 1992. (Cited on pages 28, 31, 156, and 158.)

J. C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *Journal of the Acoustical Society of America*, 106(3):1493–1510, 1999a. (Cited on pages 31, 36, 39, 132, 133, and 136.)

J. C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of the Acoustical Society of America*, 106(3):1480–1492, 1999b. (Cited on pages 39, 116, 131, 132, 133, 135, and 136.)

J. C. Middlebrooks and D. M. Green. Directional dependence of the interaural envelope delays. *Journal of the Acoustical Society of America*, 87(5): 2149–2162, 1990. (Cited on pages 10, 19, 71, and 110.)

J. C. Middlebrooks and D. M. Green. Observations on a principal components-analysis of head-related transfer-functions. *Journal of the Acoustical Society of America*, 92(1):597–599, 1992. (Cited on pages 129, 130, and 157.)

J. C. Middlebrooks, J. C. Makous, and D. M. Green. Directional sensitivity of sound-pressure levels in the human ear canal. *Journal of the Acoustical Society of America*, 86(1):89–108, 1989. (Cited on page 70.)

A. W. Mills. Lateralization of high-frequency tones. *Journal of the Acoustical Society of America*, 32:132–134, 1960. (Cited on page 10.)

P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller. Localization with binaural recordings from artificial and human heads. *Journal of the Audio Engineering Society*, 49(5):323–336, 2001. (Cited on page 36.)

H. Moller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4): 171–218, 1992. (Cited on page 18.)

H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sorensen. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, 43(4):203–217, 1995a. (Cited on pages 21, 22, 51, 54, and 72.)

H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321, 1995b. (Cited on pages 18 and 70.)

H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469, 1996. (Cited on pages 36 and 78.)

H. Møller, D. Hammershøi, C. B. Johnson, and M. F. Sørensen. Evaluation of artificial heads in listening tests. *Journal of the Audio Engineering Society*, 47 (3):83–100, 1999. (Cited on page 36.)

B. C. J. Moore, S. R. Oldfield, and G. J. Dooley. Detection and discrimination of spectral peaks and notches at 1 and 8kHz. *Journal of the Acoustical Society of America*, 85(2):820–836, 1989. (Cited on page 30.)

M. Morimoto, M. Yairi, K. Iida, and M. Itoh. The role of low frequency components in median plane localization. *Acoustical Science and Technology*, 24(2):76–82, 2003. (Cited on page 27.)

S. Müller and P. Massarani. Transfer-function measurement with sweeps. *Journal of the Audio Engineering Society*, 49(6):443–471, 2001. (Cited on page 52.)

A. D. Musicant, J. C. K. Chan, and J. E. Hind. Direction-dependent spectral properties of cat external ear - new data and cross-species comparisons. *Journal of the Acoustical Society of America*, 87(2):757–781, 1990. (Cited on page 161.)

T. Naes. Handling individual differences between assessors in sensory profiling. *Food Quality and Preference*, 2(3):187–199, 1990. (Cited on pages 78 and 86.)

T. Naes and R. Solheim. Detection and interpretation of variation within and between assessors in sensory profiling. *Journal of Sensory Studies*, 6(3): 159–177, 1991. (Cited on page 87.)

T. Neher, F. J. Rumsey, and T. Brookes. Training of listeners for the evaluation of spatial sound reproduction. In *Proceedings of the 112th Audio Engineering Society Convention*, April 2002. (Cited on pages 113, 116, and 123.)

R. Nicol. Binaural technology. AES Monograph, New York, NY, USA, 2010. (Cited on pages 19 and 26.)

S. E. Olive. Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study. *Journal of the Audio Engineering Society*, 51(9):806–825, 2003. (Cited on pages 113 and 116.)

A. R. Palmer and I. J. Russell. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing research*, 24(1):1–15, 1986. (Cited on page 10.)

C. Pantev, R. Oostenveld, A. Engelien, B. Ross, L. E. Roberts, and M. Hoke. Increased auditory cortical representation in musicians. *Nature*, 392(6678): 811–814, 1998. (Cited on page 106.)

C. Pantev, L. E. Roberts, M. Schulz, A. Engelien, and B. Ross. Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport*, 12(1):169–174, 2001. (Cited on page 106.)

J. Pernaux, M. Emerit, and R. Nicol. Perceptual evaluation of binaural sound synthesis: the problem of reporting localization judgments. In *Proceedings of the 114th Convention of the Audio Engineering Society*, March 2003. (Cited on page 84.)

P. Perner. Improving the accuracy of decision tree induction by feature preselection. *Applied Artificial Intelligence*, 15(8):747–760, 2001. (Cited on page 169.)

D. R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990. (Cited on page 80.)

E. C. Poulton. *Bias in quantifying judgments*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1989. (Cited on page 84.)

D. Pralong and S. Carlile. Measuring the human head-related transfer functions – a novel method for the construction and calibration of a miniature in-ear recording-system. *Journal of the Acoustical Society of America*, 95(6): 3435–3444, 1994. (Cited on pages 18 and 106.)

D. Pralong and S. Carlile. The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *Journal of the Acoustical Society of America*, 100(6):3785–3793, 1996. (Cited on pages 22, 47, 55, and 72.)

J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. (Cited on page 167.)

V. C. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *Journal of the Acoustical Society of America*, 118(1):364–374, 2005. (Cited on page 137.)

S. G. Rodriduez and M. A. Ramirez. Extracting and modeling approximated pinna-related transfer functions from hrtf data. In *Eleventh Meeting of the International Conference on Auditory Display*, Limerick, Ireland, 6-9 July 2005. (Cited on pages 30 and 136.)

S. K. Roffler and R. A. Butler. Factors that influence localization of sound in vertical plane. *Journal of the Acoustical Society of America*, 43(6):1255–&, 1968. (Cited on page 12.)

R. Romano, P. B. Brockhoff, M. Hersleth, O. Tomic, and T. Naes. Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. *Food Quality and Preference*, 19(2):197–209, 2008. (Cited on page 86.)

F. Rumsey. Controlled subjective assessments of two-to-five channel surround sound processing algorithms. *Journal of the Audio Engineering Society*, 47(7-8):563–582, 1999. (Cited on pages 85 and 97.)

F. Rumsey and J. Berg. Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *Audio Engineering Society Conference: 19th International Conference: Surround Sound - Techniques, Technology, and Perception*, 6 2001. (Cited on page 97.)

P. Runkle, A. Yendiki, and G. Wakefield. Active sensory tuning for immersive spatialized audio. In *Proceedings of the International Conference on Auditory Display*, 2000. (Cited on page 39.)

J. W. H. Schnupp and C. E. Carr. On hearing with more than one ear: lessons from evolution. *Nature Neuroscience*, 12(6):692–697, 2009. (Cited on page 7.)

D. Schönstein and B. F. G. Katz. Variability in perceptual evaluation of HRTFs. In *Proceedings of the 128th Convention of the Audio Engineering Society*, London, England, May 2010a. (Cited on pages 50 and 78.)

D. Schönstein and B. F. G. Katz. Sélection de HRTF dans une base de données en utilisant des paramètres morphologiques pour la synthèse binaurale. In *Proceedings of the 10th Congrès Français d'Acoustique*, number 431, Lyon, France, April 2010b. (Cited on page 77.)

D. Schönstein and B. F. G. Katz. (submitted). Variability in Perceptual Evaluation of HRTFs. *Journal of the Audio Engineering Society*, 2011. (Cited on pages 40, 155, and 182.)

D. Schönstein, B. F. G. Katz, and L. Ferré. Comparison of headphones and equalization for virtual auditory source localization. In *Proceedings of the ASA and EAA Joint Conference on Acoustics*, pages 4617–4622, Paris, France, June 2008. (Cited on page 106.)

C. L. Searle, L. D. Braida, D. R. Cuddy, and M. F. Davis. Binaural pinna disparity: another auditory localization cue. *The Journal of the Acoustical Society of America*, 57:448, 1975. (Cited on pages 26 and 138.)

G. A. F. Seber. *Multivariate observations*, volume 41. Wiley Online Library, 1984. (Cited on page 136.)

B. Seeber and H. Fastl. Subjective selection of non-individual head-related transfer functions. In *Proceedings of the 9th International Conference on*

Auditory Display, pages 259–262, Boston, United States, 2003. (Cited on pages 25, 40, and 75.)

M. A. Senova, K. I. McAnally, and R. L. Martin. Localization of virtual sound as a function of head-related impulse response duration. *Journal of the Audio Engineering Society*, 50(1-2):57–66, 2002. (Cited on page 29.)

E. A. G. Shaw. Earcanal pressure generated by a free sound field. *Journal of the Acoustical Society of America*, 39(3):465–&, 1966. (Cited on page 70.)

E. A. G. Shaw. *Handbook of sensory physiology*, chapter The external ear, pages 455–490. Springer, 1974. (Cited on page 10.)

E. A. G. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *The Journal of the Acoustical Society of America*, 44(1):240–249, 1968. (Cited on pages 33, 53, and 161.)

S. Shimada, N. Hayashi, and S. Hayashi. A clustering method for sound localization transfer-functions. *Journal of the Audio Engineering Society*, 42 (7-8):577–584, 1994. (Cited on page 40.)

K. H. Shin and Y. J. Park. Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, E91A(1):345–356, 2008. (Cited on page 129.)

B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107:1627, 2000. (Cited on page 15.)

A. Silzle. Selection and tuning of hrtfs. In *Proceedings of the 112th Convention of the Audio Engineering Society*, 4 2002. (Cited on page 38.)

W. H. Slattery and J. C. Middlebrooks. Monaural sound localization - acute versus chronic unilateral impairment. *Hearing Research*, 75(1-2):38–46, 1994. (Cited on page 28.)

J. Sodnik, R. Susnik, and S. Tomazic. Principal components of non-individualized head related transfer functions significant for azimuth perception. *Acta Acustica United with Acustica*, 92(2):312–319, 2006. (Cited on page 129.)

C. J. Tan and W. S. Gan. User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems. *Electronics Letters*, 34(25):2387–2389, 1998. (Cited on page 38.)

F. E. Toole. In-head localization of acoustic images. *Journal of the Acoustical Society of America*, 48(4):943–&, 1970. (Cited on page 26.)

J. Usher and W. Martens. Perceived naturalness of speech sounds presented using personalized versus non-personalized HRTFs. In *Proceedings of the*

13th International Conference on Auditory Display, Montréal, Canada, 2007. (Cited on pages 25, 36, 75, 78, 116, and 130.)

M. M. Van Wanrooij and A. J. Van Opstal. Contribution of head shadow and pinna cues to chronic monaural sound localization. *Journal of Neuroscience*, 24(17):4163–4171, 2004. (Cited on page 28.)

V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, New York, NY, USA, 1998. (Cited on page 173.)

D Vastfjall. Contextual influences on sound quality evaluation. *Acta Acustica United with Acustica*, 90(6):1029–1036, 2004. (Cited on page 85.)

J. Vliegen and A. J. Van Opstal. The influence of duration and level on human sound localization. *Journal of the Acoustical Society of America*, 115 (4):1705–1713, 2004. (Cited on pages 28, 29, 48, and 108.)

G. Von Békésy. *Experiments in hearing.* McGraw Hill, 1960. (Cited on page 15.)

M. Vorlander. Acoustic load on the ear caused by headphones. *The Journal of the Acoustical Society of America*, 107(4):2082–2088, 2000. (Cited on page 70.)

B. N Walker, R. M. Stanley, A. Przekwas, X.G. Tan, Z.J. Chen, H.W. Yang, P. Wilkerson, V. Harrand, C. Chancey, and A.J.M. Houtsma. High fidelity modeling and experimental evaluation of binaural bone conduction communication devices. In *Proceedings of the 19th International Congress on Acoustics*, 2007. (Cited on pages 21 and 51.)

L. Wang, F. L. Yin, and Z. Chen. Hrtf compression via principal components analysis and vector quantization. *IEICE Electronics Express*, 5(9):321–325, 2008. (Cited on page 129.)

C. S. Watson. Time course of auditory perceptual learning. *Annals of Otology Rhinology and Laryngology*, 89(5):96–102, 1980. ISSN 0003-4894. (Cited on pages 113 and 116.)

E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer-functions. *Journal of the Acoustical Society of America*, 94(1):111–123, 1993. (Cited on pages 35, 36, 59, 67, 70, 78, and 108.)

F. L. Wightman and D. Kistler. Multidimensional scaling analysis of head-related transfer functions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 98–101, 1993. (Cited on pages 35, 40, and 67.)

F. L. Wightman and D. Kistler. Measurement and validation of human HRTFs for use in hearing research. *Acta Acustica United with Acustica*, 91 (3):429–439, 2005. (Cited on pages 18, 22, 47, 55, and 72.)

F. L. Wightman and D. J. Kistler. Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis. *Journal of the Acoustical Society of America*, 85 (2):858–867, 1989a. (Cited on page 70.)

F. L. Wightman and D. J. Kistler. Headphone Simulation of Free-Field Listening. II: Psychophysical Validation. *Journal of the Acoustical Society of America*, 85(2):868–878, 1989b. (Cited on pages 24, 59, and 60.)

F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America*, 1992. (Cited on page 10.)

F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *Journal of the Acoustical Society of America*, 101(2):1050–1063, 1997. (Cited on page 28.)

F. L. Wightman and D. J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America*, 105(5):2841–2853, May 1999. (Cited on pages 11, 16, and 110.)

R. S. Woodworth and H. Schlosberg. *Experimental psychology*. Oxford and IBH Publishing, 1965. (Cited on page 10.)

D. Wright, J. H. Hebrank, and B. Wilson. Pinna reflections as cues for localization. *Journal of the Acoustical Society of America*, 56(3):957–962, 1974. (Cited on page 33.)

Z. Xie, J. Shen, Y. Liu, and D. Rao. Calibration of Headphones and Earphone with KEMAR. In Qiu, PH and Yiu, C and Zhang, H and Wen, XB, editor, *Proceedings of the 2nd international congress on image and signal processing*, pages 4553–4556, New York, NY, USA, 2009. IEEE. (Cited on page 21.)

S. Yairi, Y. Iwaya, and Y. Suzuki. Individualization feature of head-related transfer functions based on subjective evaluation. In *14th International Conference on Auditory Display*, Paris, France, 24-27 June 2008. (Cited on page 76.)

N. Zacharov and S. Bech. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, Chichester, England, 2006. (Cited on page 106.)

N. Zacharov and K. Koivuniemi. Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. In *Proceedings of the 111th Convention of the Audio Engineering Society*, 11 2001. (Cited on page 96.)

N. Zacharov and G. Lorho. What are the requirements of a listening panel for evaluating spatial audio quality? In *Spatial Audio and Sensory Evaluation Techniques Workshop*, Guildford, England, 2006. (Cited on pages xxv, 106, 107, and 123.)

P. Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002. (Cited on page 113.)

P Zahorik, DS Brungart, and AW Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United with Acustica*, 91(3):409–420, May-June 2005. (Cited on page 15.)

Xiang-Yang Zeng, Shu-Guang Wang, and Li-Ping Gao. A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures. *Journal of Sound and Vibration*, 329(19):4093–4106, 2010. (Cited on pages 41, 130, and 140.)

M. Zhang, K.C. Tan, and MH Er. Three-dimensional sound synthesis based on head-related transfer functions. *Journal of the Audio Engineering Society*, 46:836–844, 1998. (Cited on page 38.)

S. Zielinski, 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, 2008. (Cited on pages 24, 50, 78, 84, 97, 108, and 122.)

K. Zimmer, W. Ellermeier, and C. Schmid. Using probabilistic choice models to investigate auditory unpleasantness. *Acta Acustica United with Acustica*, 90(6):1019–1028, 2004. (Cited on page 85.)

D. N. Zotkin, J. Hwang, R. Duraiswami, and L. S. Davis. HRTF personalization using anthropometric measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics Proceedings*, pages 157–160, 2003. (Cited on pages 41 and 143.)

D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov. Fast head-related transfer function measurement via reciprocity. *Journal of the Acoustical Society of America*, 120(4):2202–2215, 2006. (Cited on page 37.)

J. Zwislocki and RS Feldman. Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28:860, 1956. (Cited on pages 10 and 12.)