Binaural Synthesis Individualization based on Listener Perceptual Feedback

Corentin Guezenoc

Binaural Synthesis Individualization based on Listener Perceptual Feedback

Corentin Guezenoc

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

THÈSE DE DOCTORAT DE

CENTRALESUPÉLEC
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Traitement du signal

Par

Corentin Guezenoc

Individualisation de la synthèse binaurale par retours perceptifs d'auditeur

Binaural Synthesis Individualization based on Listener Perceptual Feedback

Thèse présentée et soutenue à CentraleSupélec à Rennes, le 11 juin 2021
Unité de recherche : FAST / IETR, UMR CNRS 6164
Thèse N° : 2021CSUP0004

Rapporteurs avant soutenance :

Étienne Parizet	Professeur des Universités	INSA Lyon
Brian FG Katz	Direction de Recherche CNRS	Sorbonne Université, Paris

Composition du Jury :

Président :	**Étienne Parizet**	Professeur des Universités	INSA Lyon
Rapporteurs :	**Étienne Parizet**	Professeur des Universités	INSA Lyon
	Brian FG Katz	Directeur de Recherche CNRS	Sorbonne Université, Paris
Examinateurs :	**Nancy Bertin**	Chargée de Recherche CNRS	IRISA/INRIA Rennes
	Antoine Deleforge	Chargé de Recherche	INRIA Nancy
	Xavier Bonjour	Responsable produit	MICROOLED, Grenoble
Dir. de thèse :	**Renaud Séguier**	Professeur	CentraleSupélec, Rennes

– Voilà ! C'est tout ce qu'y a ! Unisson, quarte, quinte et c'est marre ! Tous les autres intervalles, c'est de la merde ! Le prochain que je chope en train de siffler un intervalle païen, je fais un rapport au pape !

Père Blaise, interprété par Jean-Robert Lombard,
Kaamelott, Livre II, Épisode 55, « La Quinte juste », par Alexandre Astier.

Scarlet sun, golden skies,
The scorching heat recedes.
Scarlet sun, golden skies,
The ground throbs beneath your feet.
The warm breeze ruffles the vultures feathers
While they fly for cover,
Scarlet sun, golden skies,
The scorching heat recedes.

[...]

Is it you, Electric Woman?
A being of power and steel...
Is it you, Electric Woman?
Oh god I cannot believe my eyes!

"Electric Woman", *Mind Trip EP*, 2021, by Electric Mistress.

Remerciements

Si le doctorat est une aventure très personnelle, le travail qui en résulte est un édifice qui repose sur de nombreuses fondations : les travaux des collègues, les publications scientifiques de divers chercheurs, les idées qui émergent au cours d'une conversation, le soutien moral de proches... Je ne suis et ne serai pas en mesure de remercier à leur juste mesure toutes les personnes qui ont contribué à l'aboutissement de cette thèse. C'est néanmoins ce que je vais tâcher de faire dans ces lignes. À ceux que je pourrais avoir omis, sachez que je vous suis infiniment reconnaissant.

En premier lieu, je tiens à remercier Renaud Séguier, mon directeur de thèse, pour son encadrement, pour nos nombreux échanges enrichissants, et pour son soutien tout au long de la thèse, y compris à travers la fermeture de 3D Sound Labs. Merci également à mes anciens patrons à 3D Sound Labs, Xavier Bonjour et Dimitri Singer, d'avoir accepté que je passe de mon rôle d'ingénieur R&D à celui de doctorant, toujours au sein de l'entreprise. En particulier, merci à Xavier pour nos discussions souvent passionnées et toujours fructueuses.

Merci aux rapporteurs de cette thèse, Brian FG Katz et Étienne Parizet pour leurs retours sur mon manuscrit. Leurs suggestions ont très certainement permis d'accroître la pertinence et la qualité de ces travaux. Merci une fois de plus aux rapporteurs, mais également aux examinateurs Nancy Bertin, Antoine Deleforge et Xavier Bonjour, pour la richesse et la qualité de la séance de questions lors de la soutenance.

Merci à mes collègues de l'équipe R&D de 3D Sound Labs, Adrien Leman, Pierre Berthet et Slim Ghorbal. Mes travaux de thèse reposent très largement sur notre travail d'équipe, que je tiens à saluer ici. Par ailleurs, merci à vous pour les nombreuses discussions enrichissantes qui m'ont permis d'orienter mes travaux de thèse. Enfin, merci à Adrien et Pierre qui ont, à l'occasion, gentiment pris du temps pour me donner un coup de main sur des développements spécifiques à la thèse. De manière générale, ce fut un plaisir de vous fréquenter, au travail et ailleurs.

Merci à la direction de la recherche de CentraleSupélec de m'avoir permis de poursuivre ma thèse dans de bonne conditions malgré la fermeture de 3D Sound Labs. En particulier, merci à Karine Bernard pour son soutien précieux durant cette période d'incertitude, et sans qui je n'aurai jamais trouvé ce financement. Merci à elle aussi pour l'assistance inestimable qu'elle fournit à tous les doctorants du campus rennais pour qu'ils trouvent leur chemin parmi les méandres des procédures administratives liées au doctorat.

Aux collègues, amis, parents, frère, sœur et beau-frère qui ont participé et supporté mes fastidieux tests d'écoute (souvent en plein confinement), je tiens à vous témoigner mon immense gratitude.

À mes chers collègues de CentraleSupélec, Adrien, Bastien, Morgane, Esteban, Eloïse et Lilian, merci à vous et kenavo ar wech all ! Sans vous, les pauses cafés, déjeuners, pauses jardinage et autres afterworks auraient été bien fades.

Merci également à mes amis, rennais ou autre, pour leur écoute et leur patience lorsque que je déblatérais encore et encore sur ma thèse et, plus généralement, pour leur amitié.

Merci à ma famille pour leur soutien dans une entreprise qui leur a probablement paru un peu floue, si ce n'est ésotérique. J'ai bon espoir qu'avoir assisté à ma soutenance (pour ceux qui l'ont pu) vous a permis d'y voir plus clair.

Ces années de thèse ont coïncidé avec une période fantastique pour moi en tant que musicien, et cela m'a grandement porté tout au long du doctorat. Elles seront pour moi toujours associées à mon groupe de stoner rock préféré, Electric Mistress, à ses innombrables répétitions, ses séances de composition, ses nombreux concerts et ses deux opus. Un énorme merci à vous les gars, Emmanuel, Julien et Alex, pour l'aventure musicale mais aussi pour votre amitié.

Enfin, il aurait été difficile de tenir la distance sans le soutien et la confiance indéfectibles de ma chère et tendre, Andréa. Je pense que tu connais l'étendue de ma gratitude :)

Abstract

In binaural synthesis, providing individual HRTFs (head-related transfer functions) to the end user is a key matter, which is addressed in this thesis. On the one hand, we propose a method that consists in the automatic tuning of the weights of a principal component analysis (PCA) statistical model of the HRTF set based on listener localization performance. After having examined the feasibility of the proposed approach under various settings by means of psycho-acoustic simulations of the listening tests, we test it on 12 listeners. We find that it allows considerable improvement in localization performance over non-individual conditions, up to a performance comparable to that reported in the literature for individual HRTF sets. On the other hand, we investigate an underlying question: the dimensionality reduction of HRTF sets. After having compared the PCA-based dimensionality reduction of 9 contemporary HRTF and PRTF (pinna-related transfer function) databases, we propose a dataset augmentation method that relies on randomly generating 3-D pinna meshes and calculating the corresponding PRTFs by means of the boundary element method.

Résumé

En synthèse binaurale, fournir à l'auditeur des HRTFs (fonctions de transfert relatives à la tête) personnalisées est un problème clef, traité dans cette thèse. D'une part, nous proposons une méthode d'individualisation qui consiste à régler automatiquement les poids d'un modèle statistique ACP (analyse en composantes principales) de jeu d'HRTF à partir des performances de localisation de l'auditeur. Nous examinons la faisabilité de l'approche proposée sous différentes configurations grâce à des simulations psychoacoustiques des tests d'écoute, puis la testons sur 12 auditeurs. Nous constatons qu'elle permet une amélioration considérable des performances de localisation comparé à des conditions d'écoute non-individuelles, atteignant des performances comparables à celles rapportées dans la littérature pour des HRTF individuelles. D'autre part, nous examinons une question sous-jacente : la réduction de dimensionnalité des jeux d'HRTF. Après avoir comparé la réduction de dimensionalité par ACP de 9 bases de données contemporaines d'HRTF et de PRTF (fonctions de transfert relatives au pavillon de l'oreille), nous proposons une méthode d'augmentation de données basée sur la génération aléatoire de formes d'oreilles 3D et sur la simulation des PRTF correspondantes par méthode des éléments frontières.

Diverradenn

Evit ar sintezenn divskouarnel, pourchas d'ar selaouer HRTF (*head-related transfer functions* e saozneg, da lavaret eo kevreizhennoù treuzdoug e diazalc'h ar penn) personelaet a zo ur gudenn a-ziazez, a zo kaoz outi en tezenn-mañ. Eus un tu, kinnig a reomp un hentenn personeladur, a dalvez da gefluniañ, en un doare emgefreek, pouezioù ur patrom statistikel PCA (*principal component analysis* e saozneg, da lavaret eo analizenn dre elfennoù pennañ) HRTF. Ensellet a reomp greadusted an hentenn-mañ e meur a gefluniadur a-drugarez da zrevezadennoù psiko-klevedoniel, hag he amprouiñ a reomp gant 12 selaouerien. Stadañ a reomp eo gwellaet kalz o barregezh war al lec'hiadur klevedoniel e-keñver doareoù selaou ha n'int ket hiniennek, betek barregezhioù damheñvel ouzh re danevellet el lennegezh evit doareoù selaou hiniennek. Eus un tu all, ensellet a reomp ar gudenn a-zindan-mañ : reduadur mentelezh ar strolloù HRTF. Da c'houde bezañ keñveriet ganeomp reduadur mentelezh dre PCA 9 stlennvonioù kempred HRTF ha PRTF (*pinna-related transfer functions*, da lavaret eo kevreizhennoù treuzdoug e diazalc'h ar skouarn), kinnig a reomp un hentenn evit pinvidikaat ar stlennoù hag a zo diazezet war ganedigezh dargouezhek stummoù skouarn 3D ha war drevezadur ar strolloù PRTF kenglot a-drugarez da hentenn an elfennoù bevenn (*boundary element method*, pe *BEM*, e saozneg).

Résumé substantiel

Ces travaux de thèse ont été réalisés à Rennes au sein de l'entreprise 3D Sound Labs et de l'équipe de recherche FAST (Facial Analysis, Synthesis and Tracking) de l'Institut d'Électronique et de Télécommunications de Rennes (IETR, UMR CNRS 6164), située à CentraleSupélec. Ces travaux s'inscrivent dans le projet principal de recherche et développement de cette première : apporter la synthèse binaurale individualisée au grand public. Quand l'aventure 3D Sound Labs prit fin en février 2019 (à mi-chemin du doctorat), les présents travaux de thèse furent poursuivis au sein de l'équipe FAST.

Notre système auditif nous permet de localiser les sources sonores environnantes grâce à seulement deux canaux audio, perçus aux tympans gauche et droit. Pour ce faire, le système auditif utilise divers indices de localisation : spectraux, temporels ou liés au niveau sonore, monauraux ou interauraux. Ces indices proviennent des réflexions et de la diffraction des ondes sonores entre leur émission et leur arrivée à nos tympans. Entre d'autres termes, notre tête, torse et pavillons d'oreille effectuent un filtrage directionnel des sons incidents. En reproduisant ces indices de manière adéquate dans les canaux droite et gauche d'un casque ou d'écouteurs, il est possible de donner l'illusion d'une scène sonore virtuelle (SSV) tri-dimensionnelle. Contrairement à la stéréo, cette technique, appelée reproduction binaurale, permet la perception de sons provenant de toutes les directions de l'espace, y compris en élévation.

D'autres techniques, telles que la synthèse de front d'onde ou l'ambisonie, permettent le rendu de SSV 3D grâce à des haut-parleurs. Cependant, elles en nécessitent un grand nombre, positionnés avec précision. De plus, comme pour toute technique de restitution basée sur des haut-parleurs, le rendu est souvent dégradé par les réverbérations dues à la salle environnante. En ce sens, la reproduction binaurale présente un avantage considérable : elle n'a besoin que d'équipement courant et peu coûteux pour fonctionner, c'est-à-dire un casque ou des écouteurs ordinaires. Ces derniers permettent de plus de s'affranchir de l'effet de salle.

L'approche historique à la restitution binaurale, toujours d'usage, est d'enregistrer une scène sonore au travers d'une paire de microphones placés dans les canaux auditifs d'une personne ou d'un mannequin. Le signal audio bicanal est ensuite rejoué au casque. La limitation majeure de cette technique est que le point de vue de l'auditeur sur la scène sonore est déterminé par la position ou trajectoire de la paire de microphones durant l'enregistrement, et ne peut être modifié après coup. Par exemple, lors de la restitution,

si l'auditeur tourne la tête, la SSV suit ce mouvement (alors qu'une scène fixe serait plus immersive).

Néanmoins, une autre approche, la synthèse binaurale, pallie à ce défaut en effectuant le rendu de la SSV au moment de la restitution. L'idée est, pour chaque source sonore virtuelle, de filtrer le signal mono par la paire de fonctions de transfert relatives à la tête (d'acronyme anglophone HRTF[1]) adéquate, qui contient les indices de localisation correspondant à la direction souhaitée. Grâce à cette technique, une SSV peut être adaptée en temps réel aux mouvements de l'auditeur par le biais d'un système de suivi de la tête. Mieux, une SSV complètement synthétique, c'est-à-dire constituée d'un certain nombre de sources sonores virtuelles en mouvement dans l'espace 3D, peut être l'objet d'une restitution binaurale. Cet aspect est primordial pour les jeux vidéos, et tout particulièrement adapté aux contextes de réalités virtuelle et augmentée, dans lesquelles l'utilisateur porte un casque et recherche l'immersion dans un environnement virtuel par la vision, le son et le mouvement.

Les HRTF, issues du filtrage acoustique effectué par la tête, le torse et les oreilles, dépendent non seulement de la position de la source sonore mais aussi de la morphologie de l'auditeur, ce qui leur confèrent un caractère individuel. Cependant, la synthèse binaurale est généralement effectuée à partir d'un jeu d'HRTF générique, donc non-individuel. Cela peut causer diverses dégradations dans la perception de la SSV, tels que des inversions avant-arrière, une perception erronée de l'élévation et/ou une faible impression d'externalisation (cf Section 1.3.2, [Wenzel93 ; Kim05]).

En effet, comme nous le verrons en Section 2.3 du Chapitre 2, l'obtention d'HRTF individuelles est loin d'être triviale. En particulier, la mesure acoustique, qui est la méthode historique et état-de-l'art, est fastidieuse, coûteuse et inappropriée pour le grand public. En effet, elle repose sur un dispositif de mesure coûteux et encombrant, installé en chambre anéchoïque quand c'est possible. Alternativement, il est possible de simuler numériquement ces sessions d'enregistrement à partir de scans 3D des pavillons d'oreille, de la tête et du torse. Bien que de qualité professionnelle, les scanners sont en général facilement transportables, et les sessions d'acquisition relativement courtes – de l'ordre de 15 minutes. Cependant, entre l'acquisition et le traitement des maillages 3D et la simulation numérique, le procédé dans son ensemble prend un certain temps (de l'ordre de plusieurs heures) et nécessite une puissance de calcul importante. De plus, la qualité objective et surtout perceptive d'HRTF calculées ainsi reste à démontrer.

[1]Head-related transfer function

Afin de proposer des solutions d'individualisation d'HRTF plus accessibles au grand public (*user-friendly* dans la langue de Shakespeare), des méthodes moins directes ont été proposées. Parmi celles-ci, deux catégories peuvent être distinguées : celles basées sur des données morphologiques, et celles basées sur des retours perceptifs de l'auditeur. Dans le cas du premier type de méthodes, un ou plusieurs clichés des pavillons d'oreilles, de la tête et du torse sont réalisés, puis des mesures anthropométriques en sont tirées. Un jeu d'HRTF personnalisé est ensuite déduit de ces mesures, la plupart du temps sur la base d'un jeu de données jointes d'HRTF et d'anthropométrie. Concernant le second type d'approches, l'auditeur est sollicité directement, soit en le faisant participer à des tests d'écoute dont les résultats servent à personnaliser le jeu d'HRTF, soit en lui proposant de régler lui-même les paramètres d'un modèle de jeu HRTF à l'oreille. Bien que l'approche basée anthropométrie réponde bien à notre contrainte d'accessibilité au public (il est aisé de prendre quelques photos à l'aide d'un smartphone), elle est basée sur des données morphologiques lacunaires et, malgré les nombreux travaux sur le sujet, la qualité perceptive de tels procédés d'individualisation reste à être démontrée (cf Chapitre 2, Section 2.3). D'autre part, l'approche basée sur des retours perceptifs a été sensiblement moins étudiée. Il convient de noter que ce type de procédé requiert l'attention de l'auditeur le temps d'une session de calibration des HRTF, ce qui est *a priori* plus exigeant pour l'utilisateur que de prendre quelques photos à l'aide d'un smartphone. Néanmoins, aucun équipement spécifique n'est nécessaire puisque le dispositif sur lequel est effectué le rendu binaural (smartphone, ordinateur ou tablette) est en général suffisant. Par ailleurs, ce type d'approche est guidé par une évaluation perceptive du jeu d'HRTF produit au fur et à mesure de la calibration, contrairement aux méthodes basées anthropométrie qui elles procèdent "à l'aveugle". Cela ouvre par ailleurs la possibilité d'un compromis entre durée de calibration et qualité perceptive du jeu d'HRTF proposé.

Pour les raisons évoquées ci-dessus, nous proposons donc en Chapitre 4 une méthode d'individualisation indirecte basée sur des retours perceptifs de l'auditeur. Cette dernière consiste à régler les poids d'un modèle statistique – d'analyse en composantes principales (ACP) – de jeu d'HRTF en magnitude à partir des performances de localisation de l'auditeur. Contrairement à de nombreuses approches concurrentes, ce réglage est effectué globalement, c'est à dire pour toutes les directions du jeu d'HRTF à la fois. Par ailleurs, l'auditeur est sollicité pour l'évaluation perceptive des divers jeux d'HRTF qui lui sont proposés au cours de la procédure, mais pas pour le réglage en lui-même des poids du modèle, qui est réalisé de manière automatisée par l'algorithme d'optimisation

de Nelder-Mead [Nelder65]. Dans les présents travaux, les tests d'écoute ont été restreints au plan médian, où les différences interaurales de temps et d'intensité (d'acronymes anglais respectifs ITD et ILD) sont proches de zéro, nous permettant de nous concentrer sur les indices spectraux monauraux, au cœur des problèmes perceptifs liés à l'absence d'individualisation.

Dans un premier temps, la simulation psycho-acoustique des tests d'écoute grâce au modèle auditif de Baumgartner *et al.* [Baumgartner14] nous a permis d'évaluer la faisabilité de la procédure sous diverses configurations : 3 bases de données d'entraînement pour l'ACP, et 5 nombres (compris entre 3 et 40) de composantes principales (CP) réglables. Dans toutes les conditions testées sauf une, le procédé d'optimisation a convergé vers un jeu d'HRTF qui donnait des erreurs de localisation significativement inférieures aux deux jeux d'HRTF non-individuels évalués, c'est-à-dire le jeu d'HRTF moyen de la base d'entraînement (condition initiale) et le jeu d'HRTF du mannequin Neumann KU-100. L'erreur de localisation finale tendait à décroître avec le nombre de CP, en particulier pour la base de données ARI, le taux d'erreur de quadrant (d'acronyme anglais QE) médian variant de 15 % à 7.5 %, pour des CP de 3 à 40. En comparaison, toujours pour la base ARI, les QE médians pour le jeu d'HRTF moyen et pour le KU-100 étaient respectivement de 23 % et 33 %, tandis qu'il était de seulement 6.3 % pour les jeux d'HRTF individuels. Bien que la durée estimée de la procédure pour un auditeur réel était prohibitive quand plus de 10 CP étaient utilisées, elle est apparue faisable (une ou deux heures environ) quand seulement 3 ou 5 PC étaient conservées, cela permettant une amélioration substantielle de la performance de localisation, quoique plus modeste qu'avec 10, 20 ou 40 PC.

Nous avons donc mis à l'épreuve cette faisabilité supposée en soumettant la procédure de réglage à 13 auditeurs réels. Tirant parti des enseignements des précédentes simulations, nous avons choisi d'utiliser le modèle d'HRTF entraîné sur la base ARI, limité à ses 5 premières CP. Les résultats ont excédé nos attentes, notre méthode ayant permis d'améliorer considérablement et significativement la performance de localisation par rapport aux deux conditions non-individuelles, jusqu'à une performance comparable à celles rapportées dans la littérature pour des jeux d'HRTF individuels [Middlebrooks99b ; Middlebrooks00 ; Baumgartner14]. En particulier, le QE médian pour les jeux d'HRTF customisés était de 6.2 %, tandis qu'il était de 31 % et 44 % pour les deux jeux non-individuels (moyen et KU-100, respectivement).

La méthode sus-mentionnée, ainsi que nombre de méthodes d'individualisation indirectes, reposent sur des bases de données d'HRTF, parfois couplées à des données morpho-

logiques. Cependant, les jeux d'HRTF sont une donnée de haute dimensionnalité (jusqu'à un demi million de degrés de liberté), alors que les jeux de données actuels n'incluent que peu de sujets en comparaison (un peu plus de deux cent au maximum avec la base ARI, cf Chapitre 2, Section 2.4). Il est donc souhaitable pour de telles applications de réduire la dimension du problème, c'est-à-dire la dimension de l'espace des variations inter-individuelles des jeux d'HRTF. C'est le problème que nous nous proposons d'examiner en Chapitre 3. En particulier, en Section 3.2, nous étudions la performance en réduction de dimensionnalité de l'analyse en composantes principales (ACP) sur les magnitude d'HRTF provenant de 9 jeux de données. Remarquons ici que nous avons privilégié l'ACP plutôt que d'autres techniques plus complexes d'apprentissage automatique. Ce choix est motivé par une volonté de focaliser l'analyse statistique sur les variations inter-individuelles des jeux d'HRTF, approche peu explorée jusqu'à présent dans la littérature. Puis, nous tournant vers la morphologie (dont sont issues les HRTF) en Section 3.3, nous avons constaté que la réduction de dimensionnalité par ACP fonctionne mieux sur 119 formes d'oreilles 3D que sur les 119 jeux de fonctions de transfert relatives à l'oreille (d'acronyme anglophone PRTF[2]) correspondants. En conséquence, et afin de parer au manque de bases de données d'HRTF de grande ampleur, nous proposons et implémentons en Section 3.4 une méthode d'augmentation de données qui repose sur la génération aléatoire de formes 3D d'oreilles et sur la simulation par méthode des éléments frontières des jeux de PRTF correspondants. Ces travaux ont donné lieu à la publication d'article dans le Journal of the Acoustical Society of America (JASA) [Guezenoc20a]. Le jeu de données résultant, comprenant un millier de maillages 3D d'oreille recalés et les jeux de PRTF correspondants, est public et disponible sur le site web Sofacoustics[3]. Enfin, nous nous intéressons en Section 3.5 à la performance en réduction de dimensionnalité de l'ACP lorsque entraînée sur les jeux de PRTF de WiDESPREaD. En particulier, en comparant cette performance en réduction de dimensionnalité avec celles obtenues pour d'autres bases de données d'HRTF, nous avons constaté de meilleurs résultats avec WiDESPREaD, notamment en terme de généralisation.

[2]Pinna-related transfer function

[3]`https://sofacoustics.org/data/database/widespread/`

Table of Contents

INTRODUCTION

This PhD was carried out in Rennes within the 3D Sound Labs company and the Facial Analysis, Synthesis and Tracking (FAST) research team of the Institute of Electronics and Telecommunications of Rennes (IETR, UMR CNRS 6164) located at CentraleSupélec. It falls within the principal research and development project of the former: providing individualized binaural synthesis to the public. After 3D Sound Labs closed its doors in February 2019 (halfway through the PhD), this work was carried on within the FAST team.

Our auditory system allows us to localize sound sources thanks to two audio signals perceived at the left and right ear drums. To achieve that, the human auditory system relies on monaural and interaural, spectrum-, time- and level-based auditory cues. These cues originate in the reflections and diffraction of sound on its path from the sound source to the ear drums. In other words, our head, torso and pinnae[4] perform a directional acoustic filtering of incoming sounds. By reproducing these cues appropriately in the left and right channel of a headphone or earbuds, the brain can be fooled into perceiving a three-dimensional virtual auditory scene (VAS). Unlike stereo, this technique, called binaural reproduction, allows the perception of sound from every direction in space, including along the vertical dimension.

Other techniques render 3-D VASs over loudspeakers, such as wave field synthesis or high-order Ambisonics [Furness90]. However, these require a large number of carefully positioned loudspeakers and, as any loudspeaker-based restitution, are often degraded by the surrounding room. In this regard, binaural rendition has a considerable advantage: it only requires a common and inexpensive piece of equipment to work, i.e. a standard pair of headphones or earbuds. Moreover, room effect is ruled out of the equation.

The historical approach to binaural reproduction, still well-used to this day, is to record a sound scene through a pair of microphones placed in a person or an anthropomorphic manikin's ear canals. The two-channel audio signal is then played-back through headphones. The main limitation of this technique is that the listener's point of view on the sound scene is determined by the position or trajectory of the pair of microphones at

[4] *Pinna*: latin for external ears.

recording time, and cannot be modified afterwards. For instance, at the time of play-back, if the listener turns his head, the VAS follows that movement (while a stationary scene would be more immersive).

However, another approach called binaural synthesis overcomes this limitation, by rendering the VAS at the time of play-back. The idea is, for every virtual sound source, to filter the mono signal by the adequate pair of head-related transfer functions (HRTFs) which include the localization cues that correspond to the desired sound direction. Using this technique, a VAS can be adapted in real time to the listener's orientation thanks to a head-tracker device. More importantly, a completely synthetic VAS, i.e. constituted of a number of virtual sound sources moving around the 3-D space, can be rendered binaurally. This aspect is essential for video games, and is particularly suited for virtual and augmented realities, contexts in which the user wears headphones and seeks 3-D immersion through vision, sound and movement.

HRTFs, deriving from the acoustic filtering effect of one's head, torso and pinnae, depend not only on sound source position but on morphology, which makes them specific to each listener. Nevertheless, binaural synthesis is generally performed using a generic (non-individualized) HRTF set, which can cause discrepancies such as front-back inversions, erroneous perception of the elevation and weak externalization (see Section 1.3.2, [Wenzel93; Kim05]).

Indeed, as we will see in Section 2.3 of Chapter 2, the obtention of individual HRTFs is far from trivial. For instance, the historical and state-of-the-art method to acquire individual HRTFs, acoustic measurement, is cumbersome and unsuitable for an end-user application. Indeed, it requires a heavy apparatus and an anechoic room, which makes the setup untransportable. As an alternative, it has been proposed to numerically simulate these measurement sessions from 3-D scans of the listener's pinnae, head and torso. While professional-grade, the scanning equipment is generally easily transportable, and the measurement session reasonably short – in the order of 15 minutes. However, between the scanning session, the processing of the 3-D meshes, and the simulation itself, the process in its entirety takes a long time (in the order of several hours) and requires considerable computing power. More importantly, the quality of such computed HRTFs is still to be demonstrated.

Focusing on the user-friendly aspect of HRTF individualization, less direct methods have been proposed to obtain individual HRTFs. Among these, two categories can be distinguished: those based on morphological information, and those based on subjective

feedback from the listener. In the first one, one or several pictures of the pinnae and/or head and torso are taken and anthropometric measurements derived from them. Then, a personalized HRTF set is inferred from the anthropometric data, most often based on a dataset of both HRTFs and anthropometry. In the second category, the listener either tunes the parameters of an HRTF set model while listening to it, or he participates in listening experiments whose outcomes serve to personalize the HRTF set. While the anthropometry-based approach answers well our constraint of user-friendliness – it is indeed easy to take a few pictures with a smartphone, it is based on sparse morphological information and, despite the quantity of work on the subject, the perceptual quality of such individualization processes remains to be established (see Chapter 2, Section 2.3). On the other hand, approaches based on perceptual feedback from the listener have been less studied. Such individualization processes require the listener to be attentive for the duration of a tuning session, which may be less practical than taking a few pictures with a smartphone. They however require no specific equipment (a smartphone, a PC, a tablet: any device on which the binaural synthesis is performed) and are actually based on a perceptual evaluation of the resulting HRTFs. In other words, this family of approaches do not go blindly about individualizing the HRTFs, they do it from some knowledge of the perceptual result. Furthermore, a trade-off is possible between the cumbersomeness of the process and the perceptual quality of the resulting HRTF set. In that sense, this less-explored approach is particularly interesting, which is why we propose and evaluate such a method in Chapter 4.

These user-friendly methods generally rely on databases of HRTFs, sometimes coupled with morphological data. For instance, in the approach that we present in Chapter 4, we propose to tune the parameters of a statistical model of HRTF set based on evaluations of the listener's localization performance. However, HRTF sets are a high-dimensional data (up to half a million degrees of freedom), whereas current datasets include few subjects in comparison (up to two hundred, see Chapter 2, Section 2.4). It is thus desirable for such applications to reduce the dimensionality of the problem – that is the variations of HRTF sets across individuals.

As a consequence, in Chapter 3, we explore the matter of reducing the dimensionality of magnitude HRTF sets. In particular, in Section 3.2, we investigate the dimensionality reduction performance of principal component analysis (PCA) on magnitude HRTFs from various datasets. Let us point out that we chose PCA over more complex techniques because we wanted to perform statistical modeling in a way that focuses on the inter-

subject variations of HRTF sets, which has barely been studied in the literature so far. In Section 3.3 we compare the dimensionality reduction performance of PCA on 119 pinna 3-D shapes with that of 119 matching sets of pinna-related transfer functions (PRTFs). In Section 3.4, in order to alleviate the lack of large-scale HRTF datasets, we propose and implement a data augmentation method that relies on random generations of ear shapes and numerical simulations of the matching PRTF sets. This work has been published in an article of the Journal of the Acoustical Society of America (JASA) [Guezenoc20a]. The resulting dataset, comprising over a thousand 3-D ear meshes and matching PRTF sets, was made available on-line on the Sofacoustics website[5]. In Section 3.5, we investigate the impact on dimensionality reduction performance of using this augmented PRTF dataset, which was published and presented at the 148$^{\text{th}}$ convention of the Audio Engineering Society (AES) [Guezenoc20b].

This manuscript is organized as follows. In Chapter 1 and Chapter 2, we cover background notions regarding binaural synthesis and establish a state-of-the-art of HRTF individualization techniques and databases. In Chapter 3, we deal with the statistical modeling and dimensionality reduction of magnitude HRTF sets. Contributions in this respect are five-fold. First, we present the constitution of a dataset of 119 3-D ear meshes and matching simulated PRTF sets, named FAST. Second, we look into the capacity of PCA to reduce the dimensionality of magnitude HRTF sets for FAST and 8 public datasets. Third, focusing on FAST, we compare the dimensionality reduction performance of PCA on its ear point clouds and on its matching magnitude PRTF sets. Fourth, based on the results of these two studies, we present a data augmentation method that relies on random generations of pinna meshes and numerical simulations of the corresponding PRTF sets. Fifth, we study the impact on dimensionality reduction performance of using this augmented PRTF dataset for training. Finally, in Chapter 4, we present a low-cost HRTF individualization method which consists in tuning the weights of a PCA model of magnitude HRTF set based on localization performance. First, we investigate its feasibility under various configurations by simulating the localization tasks thanks to an auditory model [Baumgartner14]. Second, the tuning procedure is submitted to 12 actual listeners.

[5] `https://sofacoustics.org/data/database/widespread/`

Chapter 1

Background

Thanks to only two audio signals perceived at the eardrums, the human brain is able to capture the spatial characteristics of surrounding sound sources. This psycho-acoustic process relies on auditory cues created by the alterations of sound on its acoustic path to the eardrums. Such cues depend not only on the room and the position of the acoustic source, but also on the listener's morphology. By reproducing them over headphones or ear-buds, it is possible, thanks to a process called binaural synthesis, to create a virtual auditory environment that imitates natural sound localization.

In this chapter, we go over the fundamentals of human auditory localization and binaural reproduction over headphone. First, we look into the mechanisms and auditory cues involved in sound localization. Second, we introduce signal processing concepts used to model these cues, namely the head-related transfer function (HRTF) and its derivatives, the pinna-related and directional transfer functions (PRTFs and DTFs, respectively). Third, we present binaural synthesis and discuss why it can and should be individualized. Finally, several important HRTF models are reviewed.

1.1 Human Auditory Localization

The human brain relies on various auditory cues to localize surrounding sound. After defining a listener-related coordinate system, we go over these interaural, monaural and dynamic cues. Finally, we discuss the sensitivity and accuracy of the human auditory localization system.

1.1.1 Coordinate System

Throughout this thesis, we will discuss the location of incoming sound sources relative to listener perception. Hence, before going on, let us introduce tools and terminology to describe spatial positions relative to the listener.

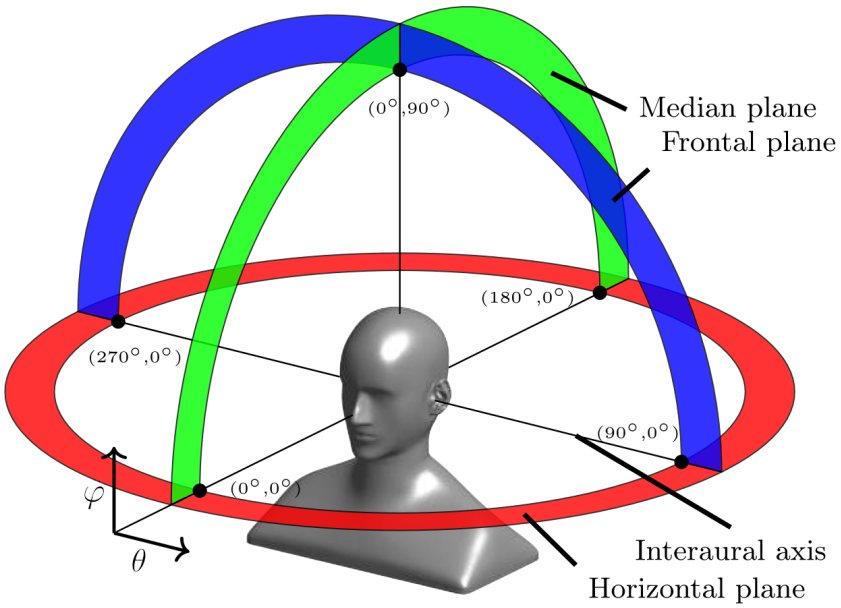


Figure 1.1 – The head-related coordinate system used throughout this thesis and the planes of interest named after standard anatomical terminology (source: [Richter19]). θ and φ denote the azimuth and elevation angles, respectively.

The axis that goes through both ears is referred to as the *interaural axis*. The center of the head and origin of the head-related coordinate system is usually defined as the middle point of the interaural segment. In coherence with the standard anatomical terms of location [Behnke12, Chap. 2], the vertical and horizontal planes that contain this axis are called the *frontal* and *horizontal* planes, respectively. The vertical plane, orthogonal to the interaural axis, that crosses it in the center of the head is called the *median plane*. A plane parallel to the median plane is called *sagittal plane*.

The Cartesian axes used throughout this thesis are the following. The x-axis stands for the front-back axis, defined by the intersection of the horizontal and median plane and oriented frontward. The y-axis is the interaural axis, oriented towards the listener's left. Finally, the z-axis represents the up-down direction and is orthogonal to the horizontal plane, oriented upward.

Several egocentric coordinate systems have been used in the literature that deals with auditory localization. The most widespread one is the spherical system, which uses azimuth and elevation angles θ and φ and a distance parameter r defined by the distance from sound source to origin. The convention adopted in this thesis is that azimuths range from -180° to 180° (back to back) and elevations from -90° to 90° (bottom to top). The

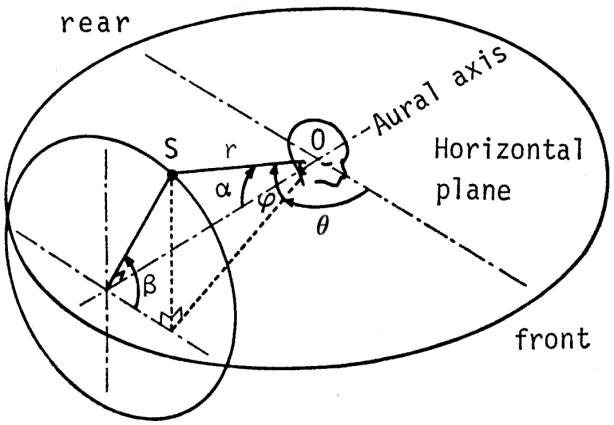


Figure 1.2 – The interaural-polar coordinate system (source: [Morimoto84]). S: sound source, O: center of the head / origin, r: distance between sound source and origin, θ: azimuth, φ: elevation, α: interaural angle, β: rising or polar angle.

direction of zero azimuth and zero elevation is located in front of the listener.

An alternative is the interaural-polar system introduced by Morimoto and Aokata [Morimoto84], deemed more adequate to sound localization. While the distance parameter is the same as in the spherical system, the rising or polar angle β is defined as the angle from the horizontal plane to the plane that contains the sound source and the interaural axis. As for the lateral angle α, it is defined as the angle from the median plane to the sagittal plane that contains the source.

1.1.2 Interaural Cues

Although early experiments on binaural hearing can be traced back to the late XVIII$^{\text{th}}$ century with Venturi (1796) and Wells (1792) [Wade08][1], Lord Rayleigh has arguably laid the foundations of our modern understanding of sound localization at the end of the XIX$^{\text{th}}$ century, with his "duplex" theory [Rayleigh07]. Experimenting with pure tones, he determined that left-right discrimination can be imputed to two types of cues: interaural time differences (ITDs) and interaural level differences (ILDs).

Interaural time difference For most directions, incoming sound waves reach one ear before the other due to the distance between both ears and head diffraction. ITD varies with source direction, starting at zero in the median plane area and reaching a maximum

[1]For a detailed account of the history of the study of binaural hearing, we advise to read Wade and Deutsch's work [Wade08].

on the left and right sides. This maximal value is of 709 µs on average, with a standard deviation of 32 µs, for a population of 33 adult subjects [Middlebrooks99a].

ITD can be well approximated using geometric models. One of the first well-known ones is the one by Woodworth [Woodworth54, Chap. 12]. Assuming a hard spherical head and a far sound source located in the horizontal plane, the ITD is modeled as

$$\mathrm{ITD}(\theta) = \frac{\Delta d(\theta)}{c} = \frac{r(\theta + \sin(\theta))}{c}, \tag{1.1}$$

where Δd is the path difference, r is the head radius, c the velocity of sound and $\theta \in [0, \frac{\pi}{2}]$ is the azimuth. Other models have been proposed in order to generalize the model to other frequency ranges [Kuhn77], sound source directions [Larcher97; Savioja99], or more complex geometrical models such as a variable position of the pinnae [Busson06; Ziegelwanger14a] and an ellipsoidal head shape [Bomhardt16c]. A more thorough state-of-the-art of ITD models can be found in Baumhardt's PhD thesis [Bomhardt17].

Interaural level difference For most incoming sound directions, acoustic pressure is greater at the ipsilateral[2] ear than at the contralateral[3] one. The phenomenon is mostly due to head diffraction. As the wavelength decreases (and frequency increases), the head is more and more of an obstacle to sound waves, leading to larger ILDs. ILD varies with sound direction, starting at zero in the median plane area and reaching maximal values in lateral positions. For instance, Middlebrooks and Green report a maximal ILD of 20 dB at 4 kHz and 35 dB at 10 kHz for an azimuth of $\theta = 90°$ [Middlebrooks90].

Perceptual importance of both cues The respective roles of ITD and ILD in lateral perception vary with frequency. For frequencies below approximately 1.5 kHz, i.e. wavelengths lower than the head width (14.5 cm on average[4]), ILDs are small and ITD is the predominant cue [Rayleigh07; Wightman92; Macpherson02]. Above 1.5 kHz, ILD becomes the predominant cue, as listener sensitivity to ITD decreases and ILD amplitude increases (diffraction is stronger for smaller wavelengths) [Rayleigh07; Kulkarni99; Macpherson02]. While the decrease in phase sensitivity is easily explainable in the case of pure tones, where the interaural phase difference is ambiguous for small wavelengths

[2]On the same side of the head as the incoming sound source.

[3]On the side of the head opposite to the incoming sound source.

[4]Source: The DINBelg 2005 campaign of anthropometric measurements of the Belgian population `http://dinbelg.be/anthropometrie.htm`.

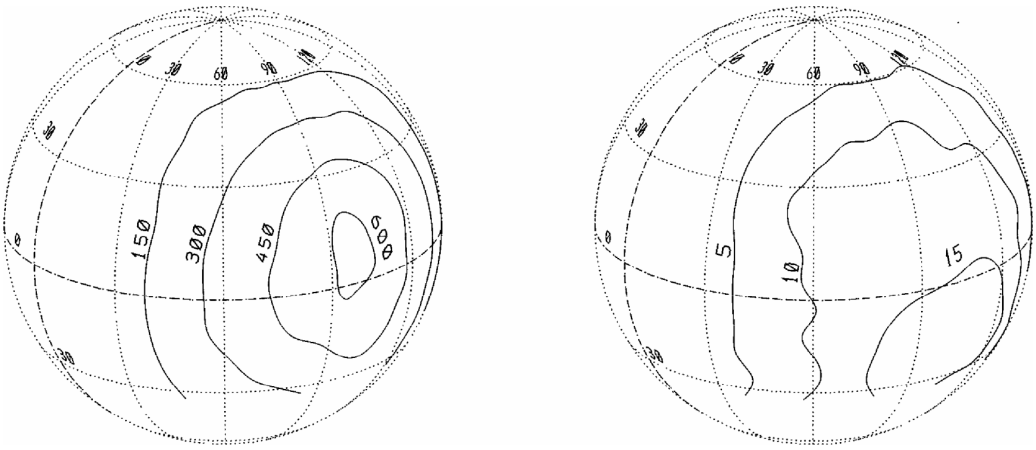


Figure 1.3 – Iso-ITD (in µs) and iso-ILD (in dB) contours of one human listener, on a globe that represents the directions of incidence (source: [Wightman99]). The direction of zero longitude/azimuth and zero latitude/elevation is faced by the listener, and the middle of the interaural axis coincides with the origin (same head-related coordinate system as in Figure 1.1).

[Rayleigh07], the psycho-acoustic mechanism remains unclear for signals with a larger band. However, ITD seems to be more important than ILD for localization as long as low-frequency phase information is present [Wightman92; Macpherson02].

1.1.3 Monaural Spectral Cues

While the perception of laterality is based on ITD and ILD, these cues are ambiguous in certain directions. As can be seen in Figure 1.3, iso-ITD and iso-ILD curves loosely correspond to circles contained by a sagittal plane, forming with the center of the head the so-called "cones of confusion" [Blauert97, Chap. 2, Sec. 5]. As a consequence, elevation and front-back discrimination can not be derived from ITD and ILD.

This information is provided to the human auditory system by monaural spectral cues. More particularly, high-frequency content (> 4 kHz) is critical for sound localization along the cones of confusion [Morimoto84; Hebrank74; Asano90].

At these frequencies, the peaks and notches caused by constructive and destructive interference in the external ear are predominant spectral features, and vary considerably with sound direction [Shaw68; Takemoto12] (see Figure 1.4) and pinna morphology. Using numerical simulations, Takemoto *et al.* [Takemoto12] establishes a thorough analysis of the link between resonances in the pinna and spectral patterns perceived at the ear canal entrance.

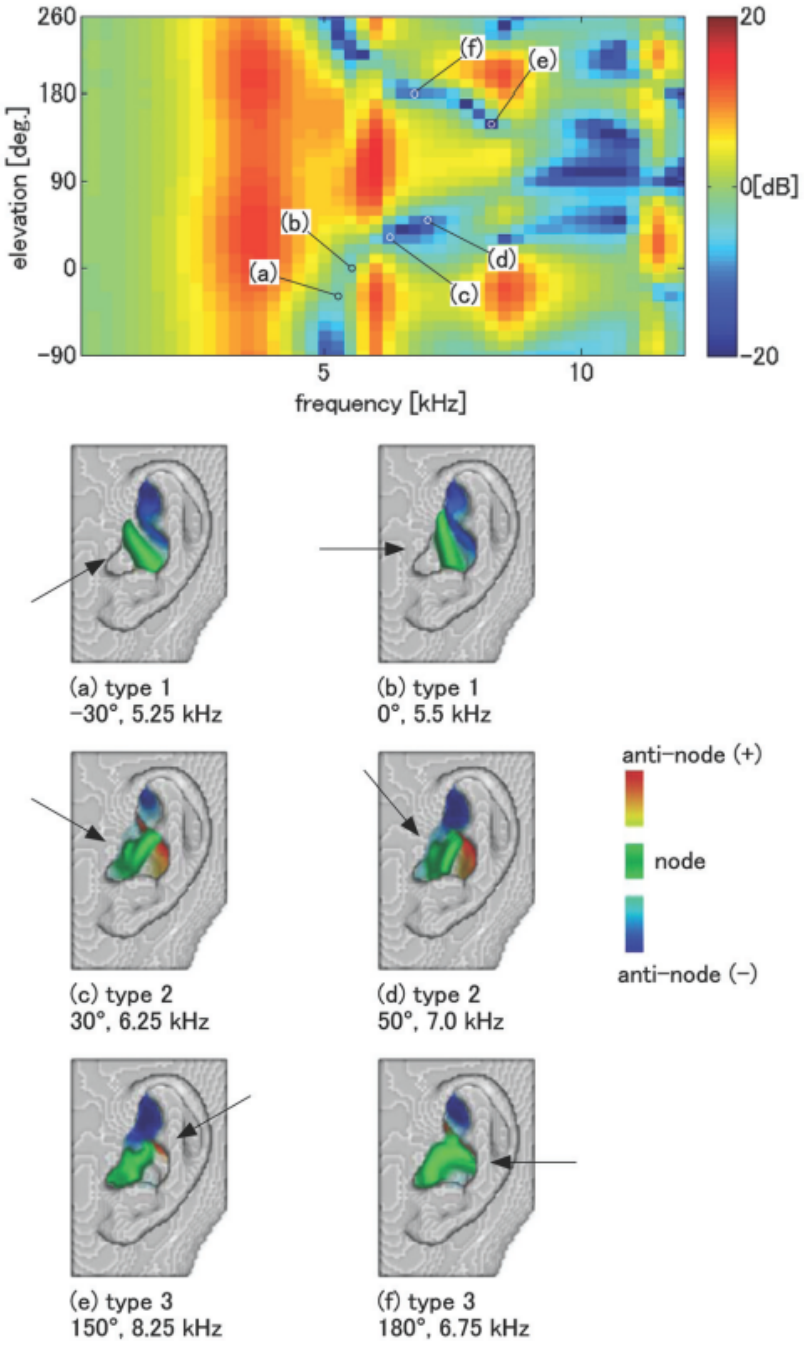


Figure 1.4 – Figure reproduced from [Takemoto12], illustrating resonances and anti-resonances in the pinna responsible for notches in the magnitude spectra of PRTFs, for an exemplary subject. The upper panel shows magnitude PRTFs in the median plane, in dB. The lower panels show the matching distribution patterns of pressure nodes and anti-nodes on the pinna. Arrows represent the source direction.

To a lesser extent, low-frequency features generated by the head and torso (< 3 kHz) can also sometimes convey useful cues for intra-conic localization [Asano90; Algazi01a].

1.1.4 Dynamic Cues

A complementary way to dispel the confusions that can occur on the sagittal planes is movement. Indeed, when the listener turns his head relatively to the sound source (or the other way around), the auditory cues are perceived for various subsequent positions, yielding precious additional information [Wallach40; Wightman99]. This is particularly useful to make up for poor spectral content or simply to improve localization (in a static set-up, front-back confusions sometimes occur even with broadband spectral cues [Bronkhorst95]). Furthermore, it would seem that dynamic cues override the monaural spectral ones [Blauert97, Chap. 2, Sec. 5].

1.1.5 Perceptual Sensitivity and Accuracy

Now that we have identified the mechanisms and cues used by the human auditory localization system, let us discuss its perceptual sensitivity and accuracy.

Interaural time difference In [Blauert97], Blauert summarizes the results of previous lateralization studies. He reports just noticeable difference (JND) ITD values between 2 and 62 µs, depending on the sound level, stimulus and experimental protocol. In addition, the JND in ITD has been found to increase with the azimuth. In a recent study, using a protocol carefully selected based on previous work [Simon16], Andreopoulou *et al.* [Andreopoulou17] report JND values ranging from 40 µs at an azimuth of 0° to 85 µs at an azimuth of 90°, in good agreement with previous research.

Interaural level difference In a study using pulse tones as stimuli, Mills [Mills60] reports median thresholds for ILD between 0.5 and 1 dB depending on the frequency (between 250 Hz and 10 kHz).

Spatial accuracy Many studies investigate the just noticeable difference in sound direction, or "localization blur", as summarized in [Blauert97, Chap. 2, Sec. 1].

In the horizontal plane, localization accuracy is best in the frontal position, steadily decreases exponentially towards the sides, and increases again towards the rear [Mills58;

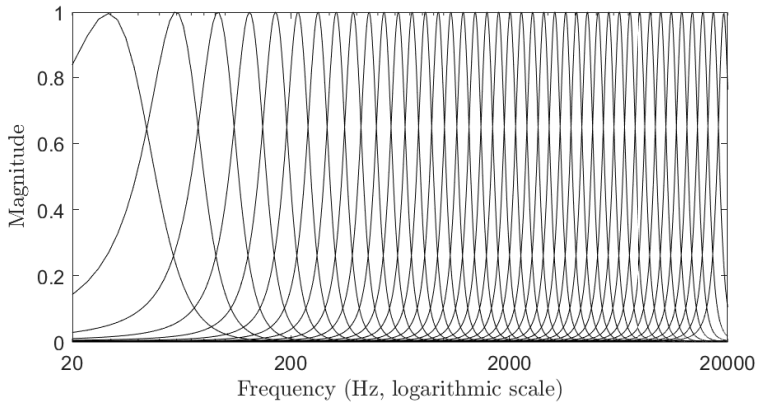


Figure 1.5 – Frequency response of a bank of 41 1-ERB-spaced 4^{th}-order gammatone filters between 20 Hz and 20 kHz.

Blauert97; Carlile97]. The order of magnitude of the localization blur in front, left-right and back is of 4°, 10° and 6° (according to Figure 2.2 of [Blauert97]).

In a study that includes various elevations, Carlile *et al.* report an average localization error of 3° in azimuth and 4° in elevation for short broadband stimuli [Carlile97]. They also notice that the errors are smaller in the anterior hemisphere.

Additionally, localization blur depends on frequency in both horizontal and median planes, as reported by Mills in the case of pure tones [Mills58]. More generally, it depends greatly on the stimulus: for instance, vertical imprecision in the frontal direction is reported in studies mentioned in [Blauert97, Chap. 2, Sec. 1] to increase from 4° to 17° by changing the stimulus from a white noise to an unfamiliar voice.

Frequency resolution Due to how the cochlea treats sound, the frequency resolution of the human auditory system is not uniform across the audible frequency range. Indeed, each hair cell along the organ of Corti is tuned to a certain frequency that depends on its location along the cochlea, resulting in higher sensitivity at low frequencies than at high ones [Ehret78].

This processing effect of the cochlea can be approximated by the so-called "Patterson-Holdsworth" filter bank [Patterson92], a bank of fourth-order gammatone filters whose bandwidths follow the *equivalent rectangular bandwidth* (ERB) scale introduced by Glasberg and Moore [Glasberg90]. This filter bank is plotted in Figure 1.5.

In the case of auditory localization, Breebaart and Kohlrausch [Breebaart01] report that smoothing non-individual spectral cues with a Patterson-Holdsworth filter bank does

not produce audible artifacts, even when using first-order gammatone filters (which are less selective than the fourth order ones). Furthermore, results from a study by Xie and Zhang [Xie10], in which the magnitude of individual spectral cues of six subjects at frequencies above 5 kHz is smoothed using a moving frequency window, suggest that a precision of 3.5 ERB for contralateral directions and 2 ERB elsewhere is sufficient. However, in their 2010 study, Breebaart and Nater argue that magnitude spectral cues smoothed using a bank of overlapping 1-ERB spaced filters are advisable as a safe frequency resolution for accurate sound localization [Breebaart10].

In the case of non-overlapping filters, the spacing must however be finer. Indeed, according to the same study by Breebart and Nater, using non-overlapping 1-ERB spaced filters instead of overlapping ones deteriorates the localization results. This is in accord with results from a study by Rugeles and Emerit [Rugeles Ospina14], in which non-individual magnitude spectral cues are filtered using a bank of non-overlapping filters. Indeed, the results of the subjective evaluation with 12 subjects suggest that the $\frac{1}{6}^{\text{th}}$-octave scale (roughly equivalent to 0.7 ERB) is too coarse. In contrast, a bank of non-overlapping $\frac{1}{12}^{\text{th}}$-octave filters seems not to produce audible alterations.

1.2 Modeling the Localization Cues

1.2.1 Head-Related Transfer Function

In the previous section, we presented different auditory cues used by the human auditory system to localize sound. These cues were identified in early experiments and associated to a corresponding spatial and/or frequency domain of perceptual influence. However, taking a step back, these cues can be viewed as the result of the alterations of sound on its path from the sound source to the left and right ear drums.

Under the traditional assumption of a linear and time-invariant system, these alterations can be described by a left and a right transfer function, commonly called head-related transfer functions (HRTFs) [Møller92, Chap. 2, Sec. 2]. A widely used definition is the one proposed by Blauert in the case of a free-field environment:

> "The *free-field transfer function* relates sound pressure at a point of measurement in the auditory canal of the experimental subject to the sound pressure that would be measured, using the same sound source, at a point corresponding to the center of the head (i.e. at the origin of the coordinate system) while the subject is not present." [Blauert97, Chap. 2, Sec. 2]

In the Fourier domain, this definition translates to the following equation:

$$HRTF_{\text{free-field}}(f) \triangleq \frac{P(f)}{P_{\text{ref}}(f)}, \tag{1.2}$$

where $P(f)$ refers to the Fourier transform of the sound pressure in the auditory canal, and $P_{\text{ref}}(f)$ refers to the Fourier transform of the reference pressure defined by Blauert i.e. the pressure at the origin in the absence of the head.

Throughout this thesis the term HRTF refers to this free-field definition. Its time-domain equivalent is referred to as the *head-related impulse response* (HRIR):

$$HRIR = F^{-1}(HRTF), \tag{1.3}$$

where F^{-1} denotes the inverse Fourier transform.

The fact that HRTFs are a function of frequency, sound source location, ear side and listener can be a source of ambiguity in what is meant by terms such as *HRTF*, *HRTFs* or *HRTF set*. Let us clarify the terminology employed in this thesis:

- *HRTF*: a filter, for a given sound source location, ear side and listener,
- *Pair of HRTFs / HRTF pair*: the left- and righ-ear filters for a given sound source location and listener,
- *Set of HRTFs / HRTF set*: a collection of filters for a given listener, for various sound source locations and ears.

The corresponding HRIR-related terms are to be understood in the same fashion.

Further on, HRTFs are denoted

$$H^{(\lambda)}(f, r, \theta, \varphi) \in \mathbb{C},$$

where $\lambda \in \{\text{L}, \text{R}\}$ denotes the left or right ear, $(r, \theta, \varphi) \in \mathbb{R}^+ \times [0, 2\pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the position of the sound source in the azimuth/elevation coordinate system, and $f \in \mathbb{R}^+$ is the frequency.

However, most often the dependency to distance r is not considered

$$H^{(\lambda)}(f, \theta, \varphi) = H^{(\lambda)}(f, r_0, \theta, \varphi).$$

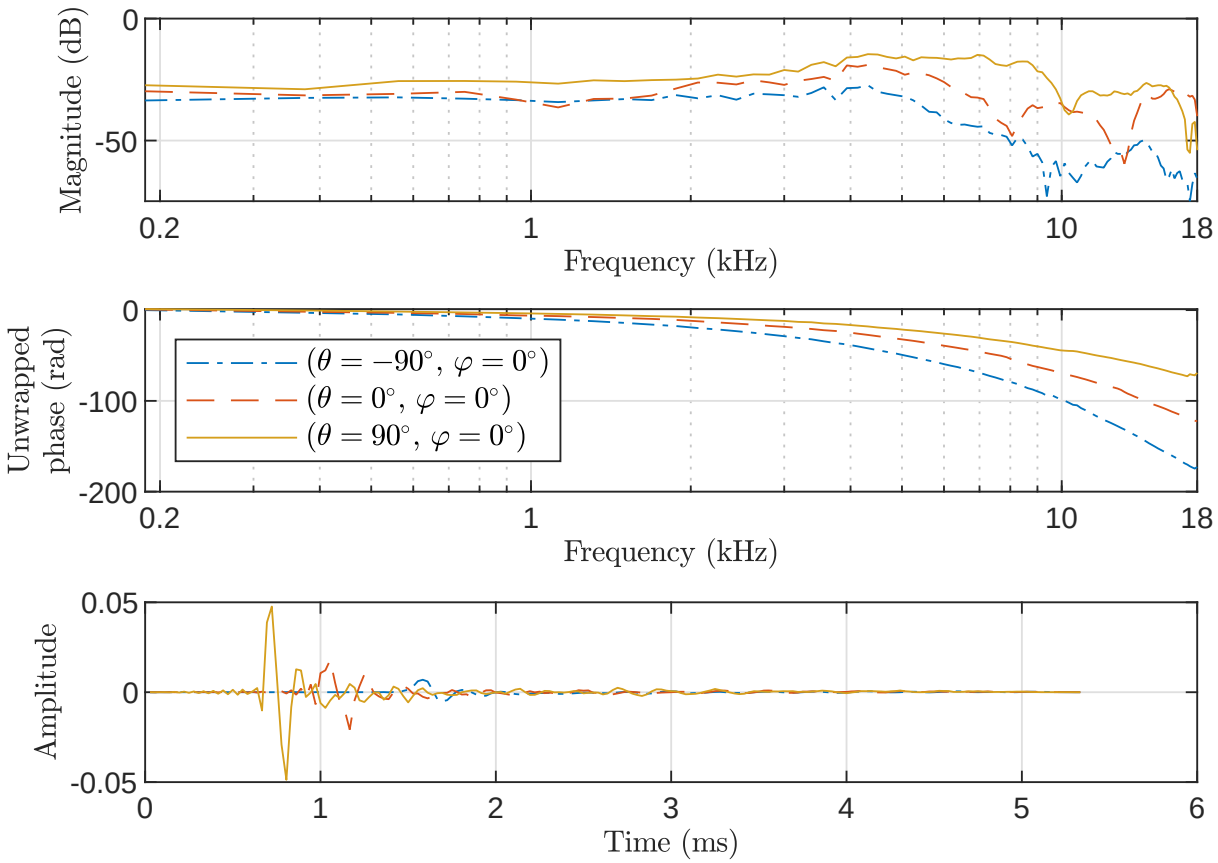


Figure 1.6 – Exemplary HRTFs and HRIRs. Magnitude (top) and phase (middle) of the HRTFs and corresponding HRIRs (bottom) of subject NH8 of the ARI dataset, for 3 horizontal directions of azimuths $-90°$, 0 and $90°$.

Indeed, while range dependency can be simulated thanks to reverberation and/or attenuation, rotations in a virtual acoustical space (VAS) rely completely on the directional variations of HRTFs. Furthermore, it is possible to extrapolate near-field HRTFs from far-field ($r_0 \gtrsim 1.5$ m) measurements [Pollow14].

For simplicity, when the ear side is irrelevant, an HRTF $H^{(\lambda)}(f, r, \theta, \varphi)$ is denoted $H(f, r, \theta, \varphi)$.

1.2.2 Pinna-Related Transfer Function

As discussed in Section 1.1.3, the pinna is at the origin of complex acoustic resonances at high frequencies that largely contribute to intra-conic[5] localization.

[5] *Intra-conic:* within a cone of confusion.

A number of studies have thus naturally focused on the component of HRTFs produced by the external ear. It is usually referred to as *pinna-related transfer functions* (PRTFs), or *pinna-related impulse responses* (PRIRs) in time-domain. Recorded or numerically simulated using the same processes as HRTFs, only the influence of the external ear is captured instead of that of a complete head and torso. PRTFs are defined and acquired in the same fashion as HRTFs. However, in contrast with the latter, only the influence of the external ear is captured, instead of that of a complete head and torso.

Methods to isolate the pinna vary. Although many studies use a mold of the pinna encased into a support for measurements [Shaw68; Hebrank74], some record real human ears after passing them through a hole in an isolation device [Spagnol11]. In the case of numerical simulations, the 3-D morphology of the pinna is easily separated from the rest of the body [Kahana06; Takemoto12; Bomhardt17].

1.2.3 Directional Transfer Function

A very widespread practice is to remove the diffuse component from the HRTFs, the *common transfer function* (CTF), and to retain only the so-called *directional transfer functions* (DTFs), as first proposed by Middlebrooks in 1990 [Middlebrooks90].

Commonly called *diffuse field equalization* (DFEQ), this process aims at uniformizing HRTF measurements while preserving auditory localization by removing the part of the HRTFs that does not vary with direction. Notably, it allows the removal of the ear canal resonance, which can vary between the left and right ears and between measurements sessions, seeing that it depends on the position of the microphone in the canal and/or the depth of the ear plug when the ear canal is blocked [Shaw68]. Furthermore, DFEQ can suppress undesired contributions from the measurement system (microphone, loudspeaker, recording amplifier, etc).

If DFEQ was initially proposed for the purpose of acoustic measurement, it is also useful for numerical simulations. While measurement imponderables are out of the picture in the latter case, the ear canal resonance is still an issue, fluctuating between the left and right ear and between subjects. Indeed, the depth at which the ear canal is blocked in the 3-D geometry and the position of the virtual microphone are both subject to variation.

DFEQ is written as follows:

$$\mathrm{DTF}(f,\theta,\varphi) = \frac{H(f,\theta,\varphi)}{\mathrm{CTF}(f)}. \tag{1.4}$$

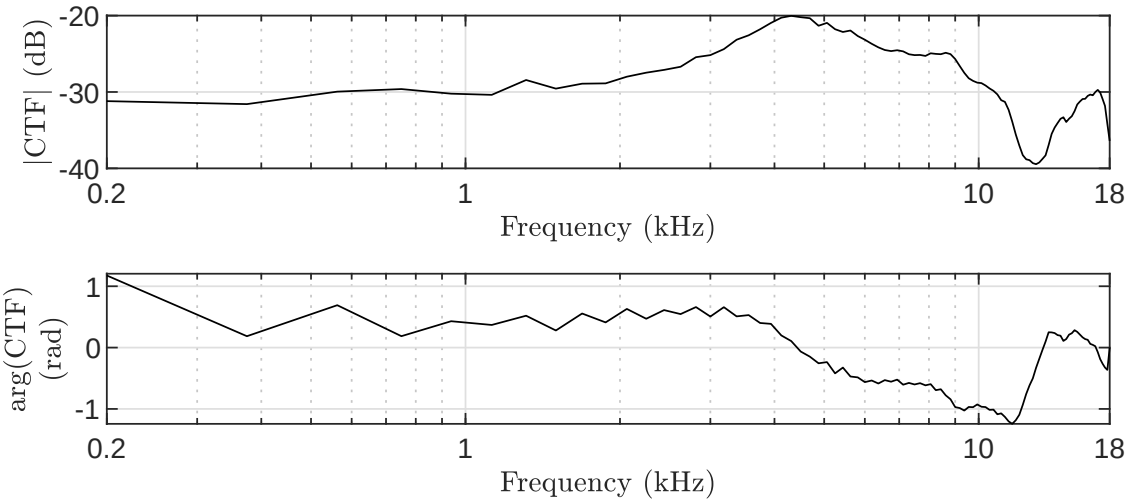


Figure 1.7 – Exemplary CTF, computed from the HRTF set of subject NH8 of the ARI dataset using the RMS averaging method.

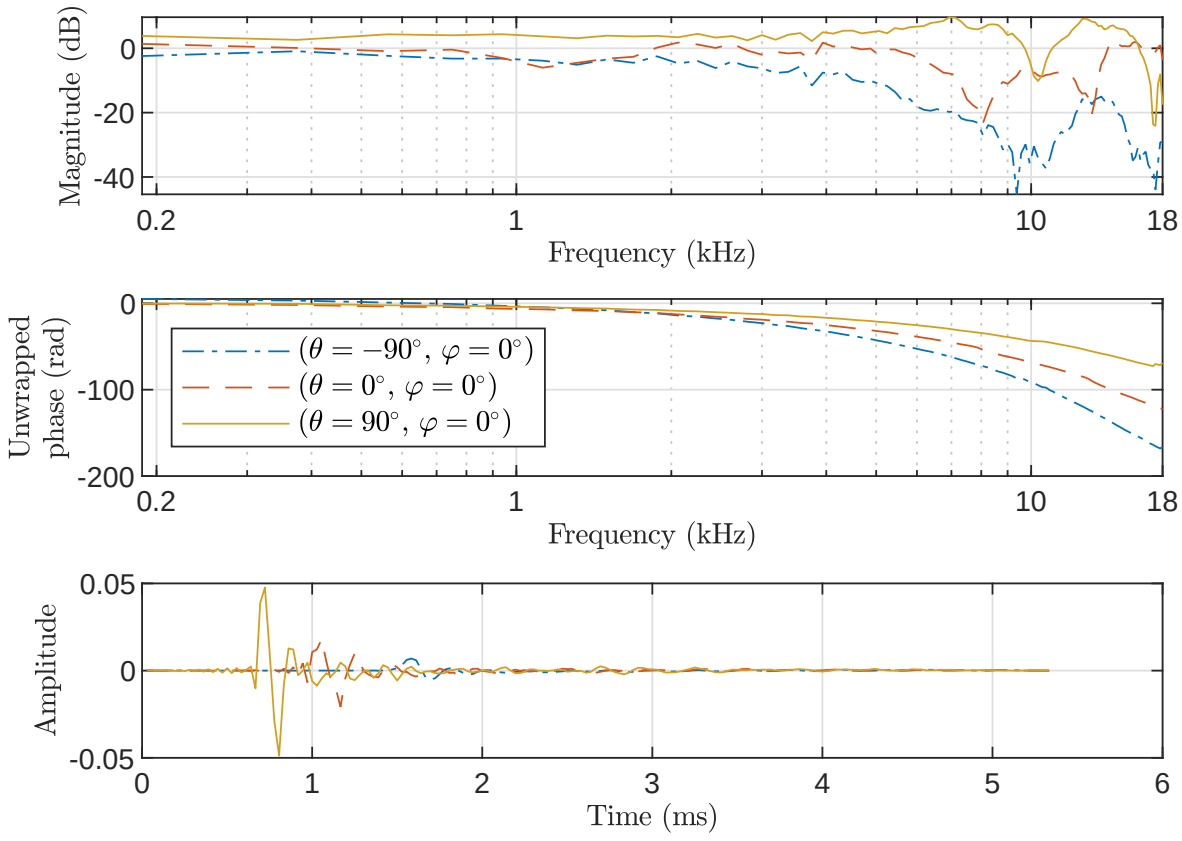


Figure 1.8 – Exemplary DTFs and DRIRs: diffuse-field equalized HRTF set of subject NH8 of the ARI dataset. As in Figure 1.6, the magnitude (top) and phase (middle) of the HRTFs and the corresponding HRIRs (bottom) are plotted for 3 horizontal directions of azimuths $-90°$, 0 and $90°$.

In [Middlebrooks90], Middlebrooks defines the diffuse field as "one in which the sound waves from all directions are of equal amplitude and in random phase", as initially proposed by Beranek [Beranek93]. Accordingly, the magnitude of the CTF is obtained by averaging the magnitude HRTFs from all n_d directions. It is often performed by computing the root-mean-square (RMS) of the magnitude spectra [Middlebrooks99a; Rugeles Ospina15]

$$|\mathrm{CTF}_{\mathrm{rms}}(f)| = \sqrt{\frac{1}{n_d}\sum_{d=1}^{n_d}|H(f,\theta_d,\varphi_d)|^2}. \tag{1.5}$$

Sometimes the averaging is performed in the log-magnitude domain [Majdak10; Baumgartner13; Guezenoc20a], which is equivalent to geometric averaging:

$$|\mathrm{CTF}_{\log}(f)| = 10^{\frac{1}{n_d}\sum_{d=1}^{n_d}\log_{10}|H(f,\theta_d,\varphi_d)|} = \sqrt[n_d]{\prod_{d=1}^{n_d}|H(f,\theta_d,\varphi_d)|}. \tag{1.6}$$

Additionally, it can be desirable to weight the average in order to give a lesser weight to measurement directions located in densely sampled areas and vice versa. A Voronoi diagram [Augenbaum85] can be used to compute the weights.

While the definition of the magnitude spectrum of the CTF stems rather clearly from the concept of diffuse field, the definition of its phase spectrum is less limpid. Indeed, according to Beranek's definition, the phase should be left indeterminate. To alleviate this, Middlebrooks [Middlebrooks90] proposes to design the CTF as a minimum phase filter – after having unsuccessfully tried to average HRIRs spatially, which resulted in ripples corresponding to a delay-and-add spectral pattern. Considering the system as linear and time invariant (a hypothesis that underlies the concept of HRTFs), they decompose the phase of the diffuse field as a sum of a minimum-phase component and an all-pass component [Oppenheim09, Chap. 5, Sec. 6]. They argue that the latter is the pure delay from the entrance of the ear canal to the recording microphone, and that it should cancel between the two ears when computing the ITD within a negligible uncertainty below 15 µs. They thus choose to set this delay to zero for convenience in computation. Using this approach to DFEQ, they verify experimentally that the phase difference between DTFs computed from measurements at two points of the same ear canal is limited to a pure delay (which corresponds to the distance between the two points), and that the difference in magnitude is close to zero.

Widely used in the community, this CTF phase design allows an easy computation of

the phase spectrum of the CTF from its magnitude by means of the Hilbert transform $\mathcal{H}$:

$$\arg\left(\mathrm{CTF}(f)\right) = \mathcal{H}\left(-\ln\left|\mathrm{CTF}(f)\right|\right). \tag{1.7}$$

1.3 Binaural Synthesis

1.3.1 Binaural Reproduction Techniques

As we have seen above, certain auditory cues allow the listener to localize sound. By incorporating these cues into the audio signals perceived at his ear drums, a two-channel audio system is able to generate the illusion of a spatial sound scene.

Binaural recording and play-back The most direct manner to achieve this is *binaural recording and play-back*: a sound scene is recorded through a pair of microphones placed inside the ear canals of a person or of an artificial head. Later on, the recording is played back through headphones or ear-buds. First experiments with binaural play-back date back to as early as the late XIX$^{\text{th}}$ century. Nowadays, the process is used in a variety of applications such as radio-phonic documentaries[6], music recordings [7] or experimental musical creations[8].

Such recordings naturally include the spatial cues due to the propagation of sound from its points of emission to the ear drums. However, the trajectory and orientation of the listener in his virtual environment is immutable. Worse, if he rotates his head while listening, the virtual auditory scene rotates with it, which is a major drawback in terms of immersion (see Section 1.1.4). Furthermore, the auditory cues are tailored to the head used for measurement whose morphology can be quite different from the listener's. This is cause to perceptual discrepancies, as we will see in Section 1.3.2.

Binaural synthesis An alternative approach made possible by last century's technological advances is to incorporate the spatial auditory cues into the binaural signals not at the time of recording but at the time of play-back, thus opening a new world of possibilities. This process, called *binaural synthesis* [Wightman89b; Møller92], consists in filtering

[6]Example of audio documentary: [Casadamont18].
[7]Example of binaural music recording: [Rueff20].
[8]Example of experimental music creation: [KRoll18].

the sound emitted by a given virtual sound source with the pair of HRTFs that corresponds to its position. This allows the synthesis of whole audio scenes by placing various sound sources at different locations in a virtual environment. This is an indispensable quality for video games, and virtual and augmented reality applications, for instance.

In contrast with binaural recording, the HRTFs and thus the spatial auditory cues can be adapted to the context of play-back. The HRTFs at play can be adjusted to the listener's position in real time, thus providing precious dynamic cues (see Section 1.1.4) and/or individual HRTFs can be used instead of an artificial head's (see Section 1.3.2 on the importance of individualization).

Extension to loudspeakers: transaural Both binaural techniques can be adapted for broadcast on loudspeakers thanks to *transaural* corrections. First proposed by Schroeder in 1970 [Schroeder70] and refined later by Cooper and Bauck [Cooper89], the fundamental principle is to cancel the cross-talk between the loudspeakers so that each ear drum receives its own spatial cues without interference from the opposite ear's. However, the spatial auditory image is very sensitive to the listener's position and orientation relatively to the loudspeakers. Corrective strategies have been developed such as using more than two loudspeakers [Baskind12] and/or adapt to the listener's position *via* head-tracking [Gardner97]. Transaural reproduction is out of the scope of this thesis.

1.3.2 Individualization - Impact on Perception

By definition (see Section 1.2.1), HRTFs describe the transformation of a sound wave on its path from the free field to the ear drums. In free field, this transformation is due to the interaction of the sound wave with the listener's pinnae, head and torso. Hence, HRTFs are in principle specific to each individual, due to their morphological origin.

In practice, using non-individual HRTFs instead of individual ones in binaural synthesis has indeed adverse effects on the perceptual quality of a VAS. In particular, localization within cones of confusion – based on monaural spectral cues – is subject to deterioration, whereas lateral localization – based on ILD and ITD – is less affected.

Indeed, in a study where 16 subjects participated in localization tests with non-individual static and free-field binaural synthesis, Wenzel *et al.* [Wenzel93] report a deterioration in the capacity to resolve location along the cones of confusion, with higher front-back and up-down confusion. In contrast, they note that lateral perception is more robust to non-individual cues.

Similar observations are made by Møller *et al.* [Møller96] when studying the localization performance of 8 subjects with real sound sources and both individual and non-individual binaural recordings. They observe increased median-plane errors for non-individual reproduction in comparison with individual reproduction – the latter being reported to be on a par with real life. In particular, they identify a general trend for frontal sources to be heard in the rear.

While the aforementioned work studied the impact of using either an individual or non-individual HRTF set on localization performance, it did not attempt to isolate the various localization cues involved. Romigh *et al.* [Romigh14] thus propose to decompose the HRTFs into an ITD component and average, lateral and intraconic spectral components. One by one, they replace each component of an individual HRTF set with its match from a non-individual HRTF set (that of the KEMAR manikin) and study the resulting localization performances. 9 subjects participated in the listening experiments. They find that the intraconic spectral component encodes the most important cues for HRTF individualization. In contrast, localization is only minimally affected by introducing non-individualized cues into the other HRTF components.

Besides front-back confusions and erroneous elevation perception, non-individual binaural synthesis can also cause discrepancies in the perception of externalization. In a study where 5 subjects were asked to report the perceived direction and distance of a virtual sound source synthesized (in the horizontal plane) thanks to both non-individual and individual binaural synthesis, Kim *et al.* reports higher front-back confusion rate and intra-cranial perception when using non-individual HRTFs [Kim05]. In contrast, no intra-cranial perception is observed by Møller *et al.* in [Møller96]. However, these experiments differ in the stimuli used for the listening experiments: Kim *et al.* use a wide-band white noise whereas Møller *et al.* use a female voice. Indeed, using a narrow-band signal is known to deteriorate localization performance compared to a wider-band signal, seeing that the monaural spectral cues are then restricted to its frequency range.

Chapter 2

State of the Art

2.1 HRTF Modeling

In this section, we review various ways of modeling HRTFs, distinguishing three categories. The first category concerns spectral models i.e. models related to the representation of HRTFs as filters. In addition to understanding which spectral features are useful for sound localization, these models were generally motivated by a concern for the reduction of the computational load and latency of binaural engines. Indeed, generating a convincing VAS potentially requires a large number of virtual sources, everyone of which needs to be convoluted with a pair of HRTFs. In that context, reducing the size of the finite impulse responses (FIRs) is critical.

The second category of models are related to the representation of HRTFs as frequency-dependent directivity responses, typically called spatial frequency response surfaces (SFRSs) [Guillon08]. Rather than the variations of HRTFs along the frequency axis, it is their variations with sound source position that are modeled. Such models are typically motivated by the need to generate continuously moving virtual sound sources, while HRTF measurement grids are discrete and their resolution often below human localization accuracy. Another motivation is to be able to recover a spatially dense HRTF set from a sparse one, thus facilitating the acquisition of individual HRTF sets by means of acoustic measurement.

The third category is statistical modeling. In particular, PCA has been widely used in the community, although other machine learning techniques have been used as well. As we will see, statistical modeling has been used as an alternative to more conventional techniques in both cases reviewed above, modeling HRTFs as filters or as SFRSs. In addition, statistical modeling can be used to learn the inter-individual variations of HRTFs, which is particularly relevant in a context of HRTF individualization.

2.1.1 Filters

Minimum-phase filter and interaural time delay

A widespread and key HRTF model is the combination of a minimum-phase filter and a pure delay often called time of arrival (TOA).

Principle For linear time-invariant systems – a hypothesis that underlies the concept of HRTF, a transfer function can be decomposed into a minimum-phase and a unitary-gain excess-phase component [Oppenheim09, Chap. 5, Sec. 6]. The latter can be decomposed further into two unitary gain components: a linear phase one i.e. a pure delay, and an all-pass one that contains the remaining phase information.

$$H = H_{\text{min phase}} \cdot H_{\text{exc phase}} \tag{2.1}$$

$$= H_{\text{min phase}} \cdot H_{\text{lin phase}} \cdot H_{\text{all-pass}}. \tag{2.2}$$

The minimum phase component $H_{\text{min phase}}$ is determined by the magnitude spectrum of the all-phase filter H. While, by construction, its magnitude spectrum is that of the original filter, its phase can conveniently be derived from the magnitude spectrum by computing the Hilbert transform of the additive inverse of its logarithm [Smith07]:

$$\begin{cases} |H_{\text{min phase}}| & = |H|, \\ \arg\left(H_{\text{min phase}}\right) & = \mathcal{H}\left(-\ln\left(|H|\right)\right). \end{cases} \tag{2.3}$$

The so-called minimum-phase processing concentrates a filter's energy into the early part of its impulse response (see Figure 2.1) while faithfully preserving the magnitude spectral response.

In an objective study of measured HRIRs of 20 subjects in 30 directions of the horizontal and median planes, Mehrgardt and Mellert [Mehrgardt77] observe that HRTFs are nearly minimum phase up to 10 kHz. Following that early work, it has been very common in the literature to approximate HRTFs as a combination of minimum-phase filters and pure delays, hence neglecting the all-pass component. According to that approximation, an HRTF $H(f, \theta, \varphi)$ is decomposed as follows:

$$H(f, \theta, \varphi) = H_{\text{mp}}(f, \theta, \varphi) \cdot \exp\left[-2\pi j f \cdot \tau(\theta, \varphi)\right], \tag{2.4}$$

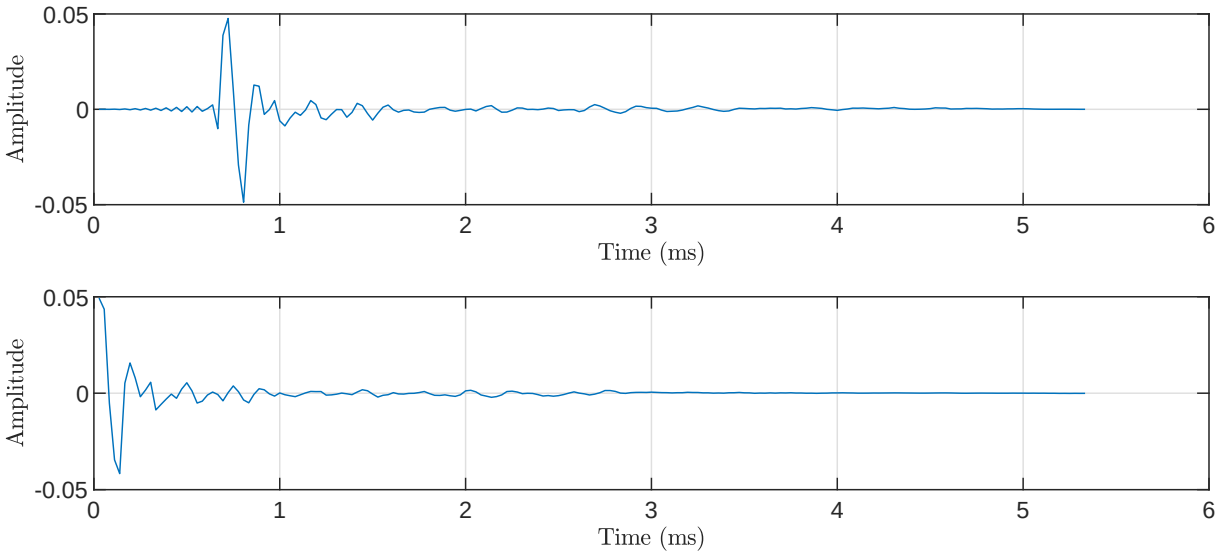


Figure 2.1 – Exemplary HRIR (above) and matching minimum-phase impulse response (below). The exemplary HRIR is that of the left ear of subject NH8 of the ARI dataset, in the ipsilateral direction of 90° azimuth and 0 elevation).

where $H_{\text{mp}}(f, \theta, \varphi)$ is the minimum-phase filter and $\tau(\theta, \varphi)$ is a pure delay, for all frequencies $f \in \mathbb{R}^+$, azimuths $\theta \in [0, 2\pi]$ and elevations $\varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

This approximation presents two major advantages. First, it permits a compact representation of HRIR data as a combination of short finite impulse responses (FIRs) and pure delays, thus reducing the computational load of binaural rendering. Second, it is highly convenient on a psychoacoustic level: the magnitude spectra and the pure delays respectively correspond to the spectral and ITD localization cues, allowing for independent analysis and manipulation of both types of cues [Hoffmann08].

Perceptual relevance The perceptual relevance of this model is investigated in several studies. Kistler *et al.* [Kistler92] compare the localization performances of 5 listeners with their own HRTF sets in 36 virtual source directions, with or without minimum-phase-plus-delay approximation. The similarity of the localization results between both conditions lead them to conclude that the approximation is perceptually valid.

Rather than performing localization experiments, Hammershøi *et al.* [Sandvad94] study directly the ability of 7 listeners to detect differences between measured and minimum-phase-plus-delay HRTFs in a multiple choice experiment in 17 virtual source directions. Their results show that some minimum-phase HRTFs are detected by some listeners, without further insight about dependency on direction or listener.

In another multiple choice experiment, Kulkarni *et al.* [Kulkarni99] compare the capacity of 4 listeners to hear the difference between measured and minimum-phase-plus-delay HRTFs, at 4 horizontal positions of azimuths 0, ±90° and 180°. In agreement with the results of [Kistler92], they find that some of the listeners were able to hear the difference between some of the HRTFs, in particular in lateral directions. Indeed, while no subject was able to discriminate between both HRTF sets at 0 and 180°, the discrimination rate was significantly greater than chance at ±90° for 2 of the 4 listeners.

Let us note that this is coherent with the observation by several authors [Algazi02; Katz14] that HRIRs are bimodal in contralateral positions close to the interaural axis, due to multiple-path propagation around the head. Such HRIRs are thus, to some extent, non-causal and contradict the minimum-phase assumption.

Plogsties *et al.* [Plogsties00] confirm, by means of a multiple choice experiment with 12 listeners, that the removal of the all-pass component is inaudible for most HRTFs but can be detected for some, in particular those that correspond to lateral directions. However, with further scrutiny and work on the way the ITD is derived from the excess-phase component, they show that the minimum-phase-plus-delay approximation is perceptually transparent for every HRTF, provided that the ITD is properly calculated.

ITD estimation ITD estimation from measured HRIRs has indeed been the subject of much work in the literature. While we shall not delve further into this question here, we encourage the curious reader to refer to [Katz14] for a thorough comparative study of the degree of variability between many of the most common ITD estimation methods, and to [Andreopoulou17] for a perceptual assessment of which of these methods are the most relevant for use in the minimum-phase-plus-delay approximation.

Further on, for simplicity, we use the terms “mag-HRTF”, “mag-DTF” or “mag-PRTF” and to refer to the magnitude spectrum of an HRTF, DTF or PRTF, respectively.

Pole-zero modeling

In coherence with the physical interpretation of HRTFs as containing resonances and reflections, the overall structure of mag-HRTFs exhibits several narrow-band peaks and notches. Based on this observation, it has been proposed to approximate HRTFs using pole-zero parameterizations [Asano90; Blommer97; Haneda99].

In particular, Haneda *et al.* approximate the horizontal HRTFs of a dummy-head using the so-called common acoustical poles and zeros (CAPZ) approach: each HRTF is modeled thanks to 20 direction-independent poles and 40 zeros [Haneda99]. All-pole or all-zero modeling have been used as well to avoid mutual cancellations of poles and zeros [Sandvad94; AlSheikh09].

Spectral smoothing

Based on prior knowledge of the frequency resolution of the human auditory localization system (see Section 1.1.5 for more details), the magnitude spectrum of an HRTF can in principle be represented thanks to magnitude coefficients distributed in a logarithmic manner along frequencies [Breebaart10].

Although most studies suggest that a magnitude value per ERB band is sufficient (i.e. about 30 magnitude coefficients) [Breebaart10], other results tend to indicate that when using non-overlapping filters the frequency scale should be as fine as a $\frac{1}{12}^{\text{th}}$ of octave which is equivalent to about a third of ERB, i.e. about 120 magnitude coefficients.

2.1.2 Spatial Frequency Response Surfaces

So far, we have reviewed approaches to model magnitude HRTFs as filters, without looking at their directional variations. An HRTF set, i.e. the set of HRTFs from all directions, can however be viewed as a directivity response that depends on frequency. This mode of representation is referred to as spatial frequency response surface (SFRS) [Guillon08]:

$$\text{SFRS}(f) = \left\{ H(f,\theta,\varphi) \mid \theta \in [0, 2\pi] \,,\ \varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \right\}. \tag{2.5}$$

Spherical harmonics

Spherical harmonics decomposition (SHD) is a popular method to model and approximate SFRSs. Similarly to the Fourier transform in 1-D, SHD expands a function $g : [0, 2\pi] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \mapsto \mathbb{C}$ into an infinite sum of weighted orthonormal basis functions called spherical harmonics.

Continous SHD The SHD of SFRSs relies on the fact that, by applying the reciprocity principle, HRTFs can be formulated as an acoustical radiation problem [Pollow14]. Assuming that the Sommerfeld radiation condition is satisfied, the solution of the Helmholtz

equation in spherical coordinates results in the expansion of the acoustic pressure field, and thus of the HRTF $H(f, r, \theta, \varphi)$ at frequency f and position (r, θ, φ) as follows.

$$H(f, r, \theta, \varphi) = \sum_{p=1}^{+\infty} \sum_{q=-p}^{p} a_{pq}(r, k) Y_p^q(\theta, \varphi) \tag{2.6}$$

where $k = \frac{2\pi f}{c_0}$ is the wavenumber. Y_p^q denotes the complex SH function of order p and degree q, defined as

$$Y_p^q(\theta, \varphi) = (-1)^q \sqrt{\frac{(2p+1)}{4\pi} \frac{(p-|q|)!}{(p+|q|)!}} P_p^{|q|}(\cos\theta) e^{jq\varphi}, \tag{2.7}$$

where $P_p^{|q|}$ is the associated Legendre polynomial. a_{pq} denotes the spherical expansion coefficients

$$a_{pq}(r, k) = b_{pq}(k) h_p(kr), \tag{2.8}$$

where $h_p(kr)$ is the spherical Hankel function of the first kind.

h_p can be used for range extrapolation which allows to derive HRTFs for any given distance r from measurements at a fixed distance r_0 (usually $r_0 \gtrsim 1.5$ m) [Pollow14]. The first spherical harmonic functions are plotted in Figure 2.2.

In practical applications, the infinite sum is truncated to a finite number of SHs $n_{\text{sh}} \in \mathbb{N}^*$:

$$H(f, r, \theta, \varphi) \simeq \sum_{p=1}^{n_{\text{sh}}} \sum_{q=-p}^{p} a_{pq}(r, k) Y_p^q(\theta, \varphi). \tag{2.9}$$

In order to avoid spatial aliasing, n_{sh} must be limited to an upper bound $n_{\text{sh}_{\max}}$, determined by the number n_d and distribution of the measurement points on the sphere of possible directions:

$$n_{\text{sh}_{\max}} = \left\lfloor \sqrt{\frac{n_d}{\gamma}} - 1 \right\rfloor, \tag{2.10}$$

where $\gamma = 4$ for an equiangular spatial sampling, $\gamma = 2$ for a Gaussian one and $\gamma = 1$ for an hyper-interpolation one [Bomhardt17, Chap. 2, Sec. 5].

Due to the analytical definition of the SHs, SHD provides a continuous representation of the HRTFs on the sphere. This characteristic facilitates real-time rotation of the VAS is binaural synthesis. Additionally, it permits the spatial interpolation of sparsely measured HRTF sets [Duraiswami04; Pollow14].

Depending on n_{sh}, SHD can be used provide a compact representation of the spatial

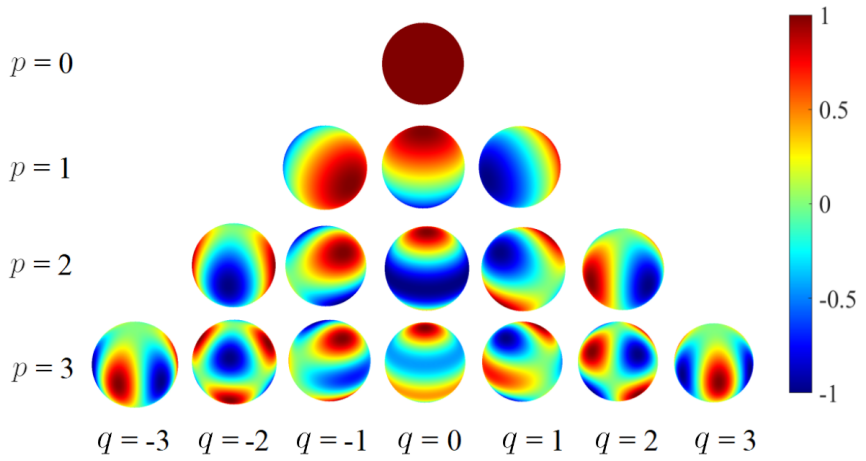


Figure 2.2 – Spherical harmonic functions Y_p^q of order $p = 0, \ldots 3$ and degree $q = -p, \ldots p$. Source: [Liu19a].

variations of an HRTF set. However, the lower n_{sh}, the more smoothed the SFRSs.

The HRIR sets of the HUTUBS dataset, for instance, are included as 35-order SHDs of the complex HRTF sets [Brinkmann19]. The $n_d = 440$ directions of an HRTF set are thus represented using $2 \cdot n_{\text{sh}} + 1 = 2 \cdot 35 + 1 = 71$ spherical harmonics coefficients a_{pq}.

Spherical wavelets

One of the limitations of SHD is that the basis functions are global, i.e. they take significant values over the whole sphere. However, magnitude HRTFs typically include sharp peaks and notches, important for intra-conic [1] localization. Accurately modeling these local features implies using spherical harmonics up to a high order.

In order to provide more efficient SFRS modeling and compression, Hu *et al.* [Hu16] proposed in 2016 to use local basis functions for spatial decomposition, in a fashion inspired by the wavelet transform. More recently, they have further improved their spatial decomposition scheme by using spherical wavelets based on the lifting scheme [Hu19]. The first analysis functions of the spherical wavelets decomposition (SWD) are displayed in Figure 2.3.

[1] Within a cone of confusion.

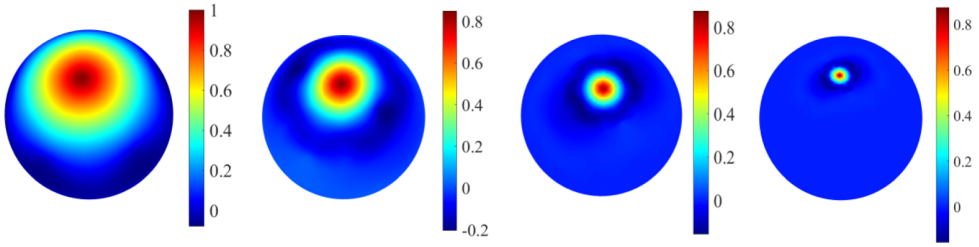


Figure 2.3 – Analysis functions for the spherical wavelet transform based on the lifting scheme [Hu19]. Left to right: scaling function of scale level 1, wavelets of scale level 1, 2 and 3. Source: [Liu19a].

2.1.3 Statistical Modeling

Statistical approaches can also be used to model HRTFs. Seeing that HRTFs are functions of frequency, sound direction, ear side and subject, depending on how the data is presented to the statistical analysis, a different kind of modeling is achieved.

Similarly to the aforementioned methods, HRTFs can be modeled as filters or as SFRSs. In addition, machine learning algorithms can be used to model the variations of HRTFs between subjects, a particularly interesting feature in a context of HRTF individualization.

Principal component analysis

Principal component analysis (PCA) [Jolliffe02] has been particularly widely used to model HRTFs in the literature, likely because of its low computational and algorithmic complexity. Moreover, PCA is a direct competitor to techniques such as pole-zero modeling or SHD and SHW, as it can be used to decompose SFRSs or filters onto a basis of orthogonal functions.

PCA is a statistical method that uses an orthogonal transformation to convert the input data into a set of uncorrelated variables called principal components (PCs). The transformation is defined so that the PCs are ordered by decreasing order of variance. The first PC thus represents the most variability in the data, then each succeeding PCs represents the most variability, under the constraint that it is orthogonal to the previous PCs. The resulting vectors form a set of orthogonal basis functions.

PCA can thus be used to decompose HRTFs onto a set of orthogonal basis functions. Furthermore, dimensionality reduction can be achieved by only retaining the first p PCs.

Let us consider a dataset of HRTFs of n_s subjects, n_f frequency bins and n_d directions.

For the sake of simplicity, we only consider the HRTFs from one ear. Let $\mathbf{X} \in \mathbb{C}^{N \times M}$ be the data matrix, where $N \in \mathbb{N}$ is the number of examples and $M \in \mathbb{N}$ the data dimension.

Spectral The most widespread approach is to perform a spectral decomposition of the HRTFs. In this case, HRTFs are viewed as filters, whose variability is learned across directions and subjects (when several are available) i.e. $N = n_d n_s$ and $M = n_f$ [Kistler92; Middlebrooks92; Hu06; Fink15; Bomhardt16a; Mokhtari19]:

$$\overset{n_d n_s}{\left[\begin{array}{c} \\ \mathbf{X} \\ \\ \end{array}\right]}\overset{}{.}_{n_f}$$

Performing PCA, Hwang *et al.* [Hwang08b] and Hugeng *et al.* [Hugeng10] have used the same data formatting approach for HRIRs and minimum-phase HRIRs, respectively, thus yielding a decomposition onto basis impulse responses rather than transfer functions.

Spatial As an alternative to spherical harmonics and wavelets decompositions (see Section 2.1.2), statistical analysis can be used to model and provide a compact representation of the spatial variations of HRTFs. Here, HRTFs are viewed as SFRSs whose variations are learned across frequency bins and subjects (when there are several) i.e. $N = n_f n_s$ and $M = n_d$:

$$\overset{n_f n_s}{\left[\begin{array}{c} \\ \mathbf{X} \\ \\ \end{array}\right]}\overset{}{.}_{n_d}$$

PCA performed in this context [Larcher00; Xie12; Takane15; Zhang20] is generally referred to as spatial PCA or SPCA.

Inter-individual A third and less explored way of performing statistical analysis is to focus on the inter-individual variations of HRTFs. Indeed, while the aforementioned approaches include contributions from various subjects (when available) in the statistical analysis, the inter-individual variability is mixed with the spatial one (when modeling HRTFs as filters) or with the spectral one (when modeling the HRTFs as SFRSs).

In particular, the set of HRTFs from all sound directions can be seen as a whole. In this case, $N = n_s$ and $M = n_f n_d$, that is each sample of the data actually corresponds to

a subject:

$$\overset{n_s}{\left[\begin{array}{c} \\ \mathbf{X} \\ \\ \end{array}\right]}_{n_f n_d}.$$

To the best of our knowledge, there is little work in the literature in which PCA is performed in the *inter-individual* fashion. In a quite extensive study [Hölzl14, Chap. 5], Hölzl compares various manners of formatting HRTF data prior to PCA and their impact on the number of PCs needed to retain a certain amount of information, including the *inter-individual* approach presented above. In [Schönstein10] and [Schönstein12a], Schönstein *et al.* perform *inter-individual* PCA on HRTF sets from the LISTEN database in order to reduce their dimensionality, as part of a method that aims at selecting a best-fit non-individual HRTF set among a database based on anthropometric measurements (see Section 2.3.3). Finally, Hold *et al.* [Hold17] perform PCA on log-magnitude HRTFs of 40 subjects, in order to study directional and frequencial areas of inter-subject variability. In addition, they emphasize that using PCA as a dimensionality reduction technique can contribute to de-noising HRTF data. However, they do not perform inter-individual PCA on complete HRTF sets, but on horizontal- and median- planes subsets.

Direction-by-direction inter-individual Alternatively, inter-subject variations of HRTFs can be studied direction by direction [Nishino07; Xu08]. In this case, $N = n_s$ and $M = n_f$:

$$\left(\overset{n_s}{\left[\begin{array}{c} \\ \mathbf{X} \\ \\ \end{array}\right]}_{n_f}\right) \times n_d \text{ times.}$$

This allows the number of examples to be of the same order as the dimension of the data. However, the critical downfall of this approach is that a different PCA must be performed at each direction, resulting in n_d (i.e. hundreds or thousands) statistical models, which is hardly practical in most problems.

Other statistical modeling techniques

Other machine learning techniques have been used to model and reduce the dimensionality of HRTFs. As in the case of PCA, statistical modeling can be performed on HRTFs seen

either as filters, SFRSs or HRTF sets.

Independent component analysis Among linear techniques, let us mention independent component analysis (ICA) [Larcher00; Liu19b] or high-order singular value decomposition [Li13]. While Larcher *et al.* use it to model complex SFRSs [Larcher00], both Liu *et al.* [Liu19b] and Li *et al.* [Li13] use it to reduce the dimensionality of HRTFs – seen as HRTF sets – in a context of HRTF individualization based on regression from anthropometric measurements.

Non-linear techniques Regarding non-linear approaches, Isomap [Grijalva16; Kapralos08] and locally linear embedding (LLE) [Duraiswami05; Kapralos08] have been applied to SFRSs.

Local tangent space alignment (LTSA) has been used as well on binaural acoustic data, with the aim of retrieving the latent two-dimensional manifold that corresponds to sound source direction and/or head orientation. In that way, Aytekin *et al.* [Aytekin08] reduce the dimensionality of HRTFs of human beings[2] and echolocating bats, in order to simulate the process of learning auditory localization for a living organism. Cues based on head movement are also used. With the aim of providing a means for sound localization for a two-ear robot, Deleforge *et al.* [Deleforge15], perform LTSA on a dataset of binaural recordings of a human-like manikin head. Based on interaural level and phase differences, they identify a two-dimensional non-linear manifold that corresponds to the head's orientation (or sound source direction, conversely).

Neural networks have come up recently in unsupervised HRTF modeling. In [Yamamoto17] , Yamamoto *et al.* train a variational autoencoder on HRTFs seen as filters, associated with directional information and personalization weights. For compression purposes, Chen *et al.* [Chen20] propose to use a different approach and to train a convolutional network on median-plane HRTF subsets, thus focusing on inter-individual variations.

Overall, as in the case of PCA, other machine learning techniques have rarely been performed in a way that focuses on the inter-individual variations of HRTF sets. Here as well, a likely cause is the small number of examples in currently available datasets compared to the dimensionality of a whole HRTF set. The scarcity of data may even be more of a problem with these more complex techniques.

[2]The 45 HRTF sets from the CIPIC dataset (see Section 2.4, [Algazi01c]).

2.2 Evaluation of HRTF Sets

In a context of HRTF individualization, we seek to improve the quality of binaural synthesis by modifying the HRTF set used for rendering. It is thus desirable to be able to evaluate and compare HRTF sets objectively and subjectively.

2.2.1 Objective Metrics

Due to the time and cost of performing perceptual evaluations, objective metrics are a necessity. They are however as diverse as there are ways of representing the signal (time- or frequency-domain, magnitude or complex spectra, linear or logarithmic scale, cepstral coefficients...). We herein present two metrics that have been commonly used in the literature. For an extensive review of HRTF metrics, we advise the curious reader to refer to [Bahu16a, Chap. 5].

Spectral distortion A rather widespread metric, the spectral distortion (SD) [Inoue05] is the RMS of the difference between log-magnitude HRTFs and is expressed in dB. This metric is sometimes referred to as spectral distance [Inoue05].

Let there be H_B and H_A two HRTF sets, and $\Delta G_\mathrm{dB} = 20\log_{10}\left(\frac{|H_\mathrm{B}|}{|H_\mathrm{A}|}\right)$ the difference between the corresponding log-magnitudes.

$$\mathrm{SD}(\theta,\varphi) = \sqrt{\frac{1}{N_f}\sum_{k=1}^{N_f} |\Delta G_\mathrm{dB}(f_k,\theta,\varphi)|^2}, \tag{2.11}$$

where (θ,φ) is a direction, designated here by its azimuth and elevation.

In order to compare two HRTF sets, the SD is typically extended to all directions by computing its RMS across directions

$$\mathrm{SD}_\mathrm{global} = \sqrt{\frac{1}{N_d}\sum_{d=1}^{N_d} \mathrm{SD}(\theta_d,\varphi_d)^2} = \sqrt{\frac{1}{N_d}\frac{1}{N_f}\sum_{d=1}^{N_d}\sum_{k=1}^{N_f} |\Delta G_\mathrm{dB}(f_k,\theta_d,\varphi_d)|^2}. \tag{2.12}$$

Prior to computing the SD, some [Huopaniemi99] resample the HRTFs to a logarithmic frequency scale in order to better fit human perception.

Inter-subject spectral difference In order not to account for gain differences, Middlebrooks *et al.* [Middlebrooks99a] proposed a metric termed inter-subject spectral dif-

ference (ISSD), expressed in dB^2. At each direction (θ, φ), the HRTF is passed through a filter-bank of $N_b = 64$ bands ranging from 3.7 to 12.9 kHz whose center frequencies are logarithmically distributed. The ISSD is then computed as the variance of the difference between the log-magnitude HRTFs in this logarithmic frequency scale:

$$\mathrm{ISSD}(\theta, \varphi) = \frac{1}{N_b} \sum_{b=1}^{N_b} \left[\Delta G_{\mathrm{dB}}(b, \theta, \varphi) - \frac{1}{N_b} \sum_{b=1}^{N_b} \Delta G_{\mathrm{dB}}(b, \theta, \varphi) \right]^2 . \tag{2.13}$$

The ISSD is typically extended to all directions by averaging the local ISSD

$$\mathrm{ISSD}_{\mathrm{global}} = \frac{1}{N_d} \sum_{d=1}^{N_d} \mathrm{ISSD}(\theta_d, \varphi_d). \tag{2.14}$$

There are some variants of the ISSD in the literature. For instance, it can be computed from a linear frequency scale [Guillon08]. Additionally, a weighting of the contributions of each frequency can be applied [Durant02].

2.2.2 Subjective Evaluation

Objective metrics can however not account for the full complexity of human auditory localization, and perceptual experiments are the ultimate test of binaural rendering quality.

Perceptual experiments are however far from trivial to implement. For instance, subjective judgments are subject to inter- and intra-subject variability [Schönstein12b; Andreopoulou16]. As a result, subjective evaluations generally include many repetitions of the same stimuli in order to be able to extract statistically significant information.

On another level, there is no absolute answer as to which type of criterion is to be used to evaluate the perceptual quality of a binaural reproduction system in a given context. The criteria found in the literature can however be divided into two categories: *spatial* ones such as localization accuracy or sensation of externalization, and *timbral* ones such as coloration or naturalness [Le Bagousse10].

We hereon provide a summarized state-of-the-art of the two main types of subjective evaluations found in the literature: judgment and localization experiments. For an extensive discussion on perceptual assessment of the quality of binaural spatial reproduction, we encourage the reader to read [Katz19].

Judgment experiments

A first type of approach is judgment experiments. Presented with one or several VASs, the listener is asked to rate the renderings [Katz12; Brinkmann19] or to indicate a preference [Yamamoto17] in an A/B comparison, based on attributes defined by the experimenters. These attributes can be global, or related to one of two categories: spectral content (i.e. timbre, coloration) or sound source location [Le Bagousse10]. Brinkmann *et al.* [Brinkmann19], for instance, asks 46 subjects to rate their measured and numerical computed HRTF sets according to 12 criteria such as difference, low-, mid- and high-frequency coloration, crispness, horizontal and vertical direction, distance, externalization and source extension. It is not trivial, however, to define the attributes so that the underlying concepts are understood by the listener as the experimenters intended. Much research has been carried out to establish a set of such attributes, which is reviewed extensively in [Katz19, pp. 380-386].

Sometimes, the test is simply a discrimination task: the listener is asked to indicate if he is able to hear a difference between two VASs. This is the case for instance in [Langendijk99], where binaural synthesis is compared with real sound sources.

Overall, this approach has the advantage of being able to explore various perceptual dimensions in spatial audio rendering quality. Moreover, the experiments can be shorter and less tiring than localization ones. However, they are highly dependent on the definition of the attributes and of the rating scales.

Localization experiments

In a context of spatial audio, it is only natural to test for localization accuracy i.e. the accuracy of the perceived position of a given sound source. In that case, the listener is presented with one or several sound sources in a virtual environment and is asked to report the position at which he perceives them. In most cases, only the direction of the sound source is evaluated, although sometimes it is rather the distance to the listener that is under test [Kim05]. The historical and perhaps most widespread manner in which binaural rendering has been evaluated is localization experiments, especially in work related to HRTF individualization [Wightman89a; Mokhtari08; Middlebrooks00; Seeber03; Shin08; Majdak10; Fink15; Liu19b].

This approach allows for a quantified and absolute evaluation of the perceptual results. Furthermore, unlike some of the criteria used in judgment experiments, there is little

ambiguity in what the listener is asked to judge.

Localization metrics Thanks to the quantitative nature of the results of localization experiments, a number of localization metrics have been proposed. Some [Wightman89a; Carlile97; Jin00] use the spherical correlation coefficient (SCC), a form of correlation between actual and perceived sound source locations on the sphere of possible directions. Others simply report a percentage of correct answers [Hu08]. However, most studies use metrics based on the angular difference between actual and perceived sound source direction.

Independently of the metric chosen, it is generally supplemented with the percentage of front-back, up-down or hemispherical inversions, due to the recurrence of such phenomena [Asano90]. These confusions are often removed or corrected prior to the computation of the main metric [Carlile97; Middlebrooks99b; Martin01; Middlebrooks99b; Zhang20].

For instance, Middlebrooks *et al.*'s [Middlebrooks99b] set of metrics, is composed of a quadrant error (QE), a lateral angle error (LE) and a local polar angle error (PE). Used by a number of other studies since [Majdak10], and in particular by Baumgartner *et al.* in their auditory model for localization prediction [Baumgartner14]. These metrics are based on the lateral-polar coordinate system introduced by Morimoto *et al.* [Morimoto84] (see Chapter 1, Section 1.1.1) and are defined as follows.

Let $\alpha_d^{(\mathrm{req})}$ and $\beta_d^{(\mathrm{req})}$ be the requested lateral and polar angles and $\alpha_{d,r}^{(\mathrm{ans})}$ $\beta_{d,r}^{(\mathrm{ans})}$ be the corresponding answers, for all tested sound directions $d = 1, \ldots D$ and all repetitions $r = 1, \ldots R$. QE is a percentage that accounts for intraconic errors of more than 90°:

$$\mathrm{QE} = 100 \cdot \frac{\mathrm{card}(\mathcal{Q})}{D \cdot R}, \tag{2.15}$$

where

$$\mathcal{Q} = \left\{ (d,r) \in \{1, \ldots D\} \times \{1, \ldots R\} \mid \left| \beta_{d,r}^{(\mathrm{ans})} - \beta_d^{(\mathrm{req})} \right| > 90° \right\}. \tag{2.16}$$

PE is the RMS of the local polar angular error:

$$\mathrm{PE} = \sqrt{\frac{1}{\mathrm{card}\left(\breve{\mathcal{Q}}\right)} \sum_{(d,r)\in\breve{\mathcal{Q}}} \left| \beta_{d,r}^{(\mathrm{ans})} - \beta_d^{(\mathrm{req})} \right|^2}, \tag{2.17}$$

with $\breve{\mathcal{Q}} = \{1, \ldots D\} \times \{1, \ldots R\} \setminus \mathcal{Q}$.

As to the LE, it accounts for errors along the lateral dimension:

$$\mathrm{LE} = \sqrt{\frac{1}{D \cdot R} \sum_{d=1}^{D} \sum_{r=1}^{R} \left(\alpha_{d,r}^{(\mathrm{ans})} - \alpha_{d}^{(\mathrm{req})}\right)^2}. \tag{2.18}$$

As a variant, the polar error can be computed without excluding the intraconic errors *a priori* – which we do in Chapter 4. Implemented by Baumgartner *et al.* [Baumgartner14] as part of their auditory modeling toolbox [Søndergaard13], the absolute polar error (APE) is defined as follows:

$$\mathrm{APE} = \sqrt{\frac{1}{D \cdot R} \sum_{d=1}^{D} \sum_{r=1}^{R} \left|\beta_{d,r}^{(\mathrm{ans})} - \beta_{d}^{(\mathrm{req})}\right|^2}. \tag{2.19}$$

Localization versus judgment

In [Zagala20], Zagala *et al.* point out that there is a considerable lack of cross-comparison of localization and judgment tasks in the literature. In order to alleviate that, they study the link between rankings of 8 representative HRTF sets from the LISTEN database (previously identified in [Katz12]) according to two different types of perceptual evaluations: a localization task, and a judgment task – similar to that of [Katz12] – in which the listeners evaluate the overall rendering quality of two virtual trajectories, horizontal and vertical. For each type of test, various metrics are covered. 28 subjects participated in the experiment.

Overall, they observe that localization performances across HRTF sets are correlated to overall quality of experience judgments. Notably, the best HRTF set selected according to perceptual metrics for one given method exhibit a rating score better than a random selection in the alternate method.

Looking into the various metrics related to each task, they report that some of the metrics from the localization method correlated better to metrics from the quality evaluation method than others: metrics such as the *mean great circle error* and *mean unsigned polar error* should be preferred over the *confusion rate* or *mean unsigned lateral error* to predict overall quality of experience.

Finally, studying the repeatability of the listeners' answers, they find that raters who were consistent in one task tended to be consistent in the other. What is more, consistent raters tended to score best with the same HRTF sets in both methods, whereas inconsistent raters were more likely to score differently with each HRTF set depending on the

method.

2.2.3 Localization Prediction

As we have seen, although subjective experiments are indispensable when evaluating an HRTF set, they are delicate to implement due to several problems such as headphone calibration, listener fatigue and variability of subjective answers. Furthermore, many repetitions are needed to establish statistical significance, which is costly in time and money. As to objective metrics, they allow for an inexpensive comparison of HRTF sets but cannot account for the complexity of the human auditory system.

A compromise is reached thanks to auditory models that mimic the mechanisms of sound localization to predict localization performance. While there were previous attempts at modeling sound localization [Middlebrooks92; Langendijk02], we herein focus on the widely popular Baumgartner model [Baumgartner14] which we use intensively in Chapter 4.

The Baumgartner model

The Baumgartner model aims at predicting localization performance inside a sagittal plane. It has been used in a large number of studies by different research teams [Geronazzo18; Brinkmann17; Braren19; Spagnol20; Zhang20]. One of the reasons for this popularity is the fact that its Matlab code is freely available online in the Auditory Modeling Toolbox[3] (AMT) [Søndergaard13]. Another one is the fact that the results of this auditory model have been verified against real localization experiments.

This function model is based on the hypothesis that a listener constructs an internal template of his own HRTFs as the result of a lifelong learning process. The structure of the model is displayed in Figure 2.4. For a given sagittal plane, the internal representation of the spectral features is associated to matching corresponding polar angles. When listening to a sound signal, its internal spectral representation is compared to the internal template. The more similar the input signal is to the cues associated with a given direction, the higher is the probability of perceiving sound coming from that direction. When listening to a new *target* HRTF set, the input signal is created by convoluting a reference stimulus (impulse) with the target HRTF.

[3] `https://amtoolbox.sourceforge.net/`

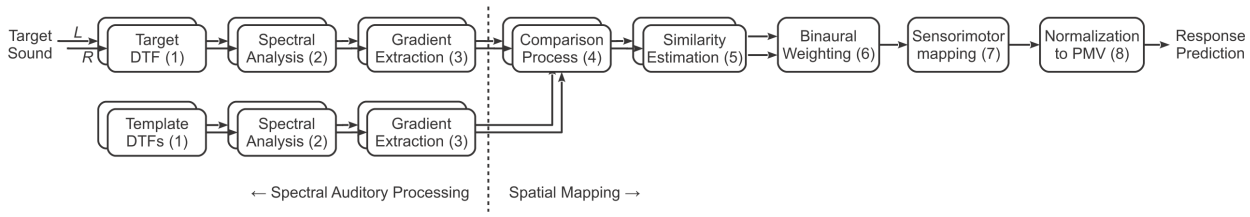


Figure 2.4 – Structure of the Baumgartner sagittal-plane localization model (reproduced from [Baumgartner14]).

Internal representation The internal representation is derived from the HRTFs as follows. First, a DFEQ of the HRTFs is performed by geometric averaging (see Equations (1.4) and (1.6) and [Majdak10]). The resulting DTFs are then filtered by a 1-ERB-bandwidth gammatone filter bank, aimed at simulating the frequency resolution of the cochlea (see Chapter 1, Section 1.1.5), for frequencies ranging from 0.7 kHz to 18 kHz.

In order to simulate the effect of the dorsal cochlear nucleus (DCN), a positive gradient is then extracted from the log-magnitude spectra. This model of the DCN derives from a study on cats by Reiss and Young [Reiss05].

$$\mathrm{PG}(D, b, \beta) = \max\left[20\log_{10}\left(\frac{|D(b,\beta)|}{|D(b-1,\beta)|}\right), 0\right], \tag{2.20}$$

where $b = 2, \ldots N_b$ denotes the frequency band, $\beta \in [-90°, 270°]$ the polar angle and $D(b, \beta) \in \mathbb{C}$ the corresponding DTF value.

Comparison Given a target sound signal emitted at polar angle β_0, its internal representation is compared to all templates (each associated with a polar angles β). The underlying idea is that the listener will perceive the sound source at the angle associated with the template representation closest to the target one.

The distance metric is computed by averaging across frequencies the absolute difference between positive gradients:

$$\mathrm{dist}(\beta, \beta_0) = \frac{1}{N_b - 1}\sum_{b=2}^{N_b} \left|\mathrm{PG}(D_{\mathrm{temp}}, b, \beta_0) - \mathrm{PG}(D_{\mathrm{targ}}, b, \beta)\right|, \tag{2.21}$$

where D_{temp} and D_{targ} denote the template and target DTFs respectively.

Similarity estimation The distance is then translated into a similarity index (SI) SI in a non-linear fashion by means of a sigmoid function:

$$\mathrm{SI}(\beta, \beta_0) = 1 - \frac{1}{1 + \exp\left(-\Gamma\left[\mathrm{dist}(\beta, \beta_0) - S_l\right]\right)}, \tag{2.22}$$

where Γ denote the degree of selectivity and S_l the sensitivity. These two parameters are later tuned based on real localization results. The sensitivity parameter, in particular, is designed to be individual, and accounts for inter-subject variability in localization performance. The lower Γ and the higher S_l, the more sensitive the listener to spectral variations and the more precise his localization.

Binaural weighting At this point of the process, spectral features were compared independently for the right and left pinnae. Then, left and right similarity indices are combined with binaural weighting. The weights vary with the lateral angle $\alpha \in [-90°, 90°]$ according to sigmoid functions, based empirically on two studies [Morimoto01; Macpherson07]:

$$\begin{cases} w_{\mathrm{L}}(\alpha) = \frac{1}{1+e^{-\frac{\alpha}{\Omega}}}, \\ w_{\mathrm{R}}(\alpha) = 1 - w_{\mathrm{L}}(\alpha), \end{cases} \tag{2.23}$$

where Ω is a parameter set to 13° in order to fit the experimental results of the aforementioned studies.

The SIs are then interpolated to match a regular sampling of the polar angles.

Sensorimotor mapping Between auditory perception and source source pointing, a complex sensorimotor process takes place, which results in pointing hazards [Bahu16b]. These pointing hazards are modeled by Baumgartner *et al.* as a centered Gaussian scatter which "smears" the answers. This scattering effect is defined in the elevation dimension (coherent with the body frame) with a constant concentration. Projected into the polar dimension, the concentration depends on the interaural angle and is expressed as follows:

$$\kappa(\alpha) = \frac{\cos^2 \alpha}{\epsilon^2}, \tag{2.24}$$

where $\alpha \in [-90°, 90°]$ is the lateral angle and ϵ is the scatter parameter defined in the elevation dimension.

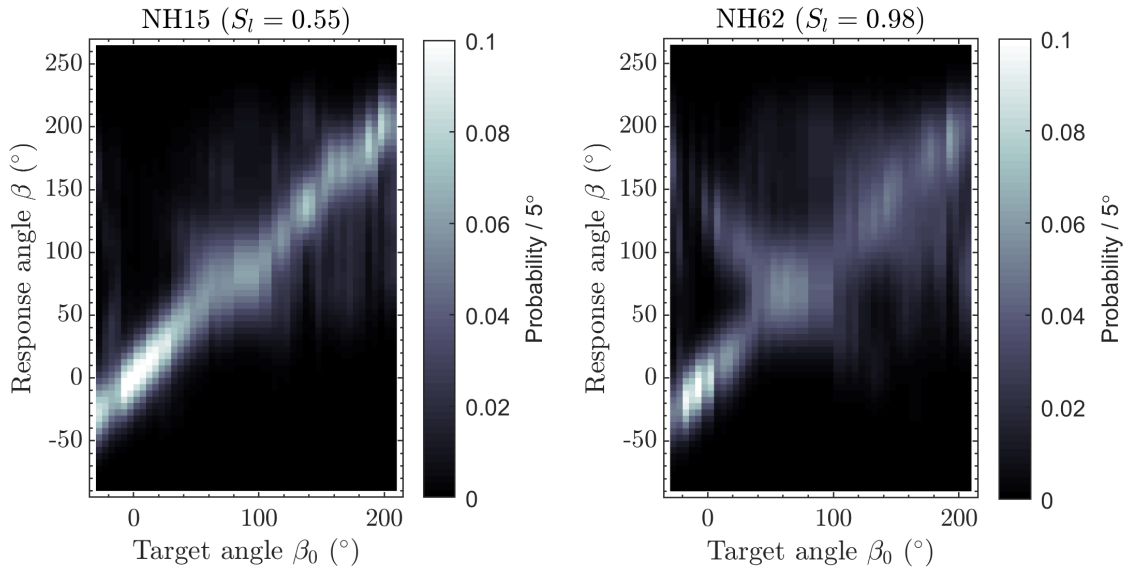


Figure 2.5 – Probability mass vectors output by the Baumgartner model for subjects NH15 (left) and NH62 (right) of the ARI database, when predicting localization performance with their own HRTF set. Their own personal selectivity parameters S_l (as reported in [Baumgartner14]) are entered in the model and indicated above each plot.

Probability mass vector Finally, in order to obtain probabilistic predictions, the similarity indices are normalized by the sum of the SIs on the sagittal plane:

$$\mathrm{P}(\beta|\beta_0) = \frac{\mathrm{SI}(\beta, \beta_0)}{\sum_{\beta} \mathrm{SI}(\beta, \beta_0)}. \tag{2.25}$$

An example of probability mass vector is shown in Figure 2.5.

Alternatives and extensions

Recently, Barumerli *et al.* [Barumerli20] have proposed an extension of the Baumgartner model to both lateral and polar dimensions, which is to be added to the AMT.

In her PhD thesis [Bahu16a, Chap. 6-7], Bahu proposes an alternative auditory model. Indeed, the Baumgartner model is calibrated for each listener based on their localization performance with their own HRTFs. When individual HRTFs are not available, the individual sensitivity parameter can thus not be tuned and a generic setting must be used. To alleviate this, Bahu's model aims at predicting localization performances with non-

individual HRTFs as well as individual ones, without having to tune the model parameters individually. Furthermore, the model handles both angular dimensions.

2.3 HRTF Individualization Techniques

As we have seen in Chapter 1, using individual HRTFs in binaural synthesis is key to reproducing accurate localization cues. Nevertheless, in most current applications a generic HRTF set is used. Indeed, the historical and state-of-the-art method to capture individual HRTFs, i.e. acoustic measurement, is cumbersome and inaccessible to the public. Hence, a lot of work has been done over the course of the last decades to provide an alternative to acoustic measurement.

In this section, we provide a survey of the various ways of obtaining individualized HRTFs. Four categories are distinguished: acoustic measurement, numerical simulation and indirect methods either based on morphological data or perceptual feedback. We pay attention, in particular, to the perceptual assessment of the methods (see Table 2.1 for an overview) and their user-friendliness, according to criteria such as user comfort, required equipment and process duration.

2.3.1 Acoustic Measurement

As mentioned above, acoustic measurement is the historical and most straightforward method to acquire HRTFs. It consists in placing microphones in the subject's ear canals and to record impulse responses from every direction of interest. Ideally, the measurements are performed in an anechoic or semi-anechoic environment in order to acquire free-field auditory cues. Indeed, HRTFs are by definition free-field transfer functions (see Chapter 1). Furthermore, it is easier to control room reverberation *a posteriori* in a VAS if the HRTFs are anechoic in the first place.

Measurement setup

A state-of-the-art measurement setup [Bomhardt16b; Rugeles Ospina16; Carpentier14; Enzner08; Mokhtari08] typically features loudspeakers on one or several vertical arcs and a turntable on which the subject stands or sits, though a variety of measurement setups can be read of in the literature such as one or several loudspeakers moving around a still subject [Langendijk99]. This is the main shortcoming of the method: the equipment

	Eval. type	Baseline	N_{subj}	Results
Acoustic measurement	Localization [Wightman89a; Bronkhorst95; Møller96; Blauert98; Carlile98; Martin01; Majdak10]	RS	3-10	Variable between studies, often degraded compared to RS.
	Preference [Langendijk99]	RS	6	
Numerical simulation	Localization [Ziegelwanger15b]	IA	3	Too few subjects / studies.
	Rating [Brinkmann19]	IA	42	Marked audible differences with IA, loc. discrepancies.
Indirect indiv. from anthropometric data				
Selection	Localization [Zotkin02]	NIA	6	Not conclusive.
	Rating [Yao17]	NIA	30	Retrieval of preferred HRTF set for 40 % of the subjects.
Adaptation	Localization [Middlebrooks00]	NIA, IA	9	Variable: notably less confusions for 7/9 subjects, increase for 2/9.
Regression	Localization [Hu06; Hu08; Liu19b; Zhang20]	NIA	5-6	Rare perceptual experiments (in 4/15 papers). Some improvement over NIA for 3 studies [Hu06; Hu08; Liu19b]. Stat. significant improvement for [Zhang20].
Indirect indiv. from perceptual feedback				
Selection	Localization [Seeber03; Iwaya06; Katz12; Zagala20]	wNIA, IA, RS	7-25	Modest improvement over NIA, variable between studies.
	Rating [Schönstein10]	Chance	37	Better than random ranking for 26/37 subjects.
Adaptation	Localization [Tan98; Middlebrooks00]	NIA	9-10	Modest improvement over NIA for [Tan98]. Same as for anthropometry-based adaptation for [Middlebrooks00].
Synthesis	Localization [Shin08; Hwang08a; Fink15]	IA, NIA	1-6	Notable improvement for [Hwang08a]. Less clear for [Shin08; Fink15].
	Rating [Yamamoto17]	bNIA	20	Stat. significant preference over bNIA for 18/20 subjects.

Table 2.1 – Overview of perceptual evaluations for the major HRTF individualization approaches.
Eval. type: type of subjective evaluation. Baseline: baseline condition(s). Acronyms RS, IA, NIA, bNIA and wNIA stand respectively for real sound sources, individual and non-individual HRTF sets, and best- and worst-fit non-individual HRTF sets (selected among a database). N_{subj}: number of subjects. Results: overview of the perceptual studies' results.

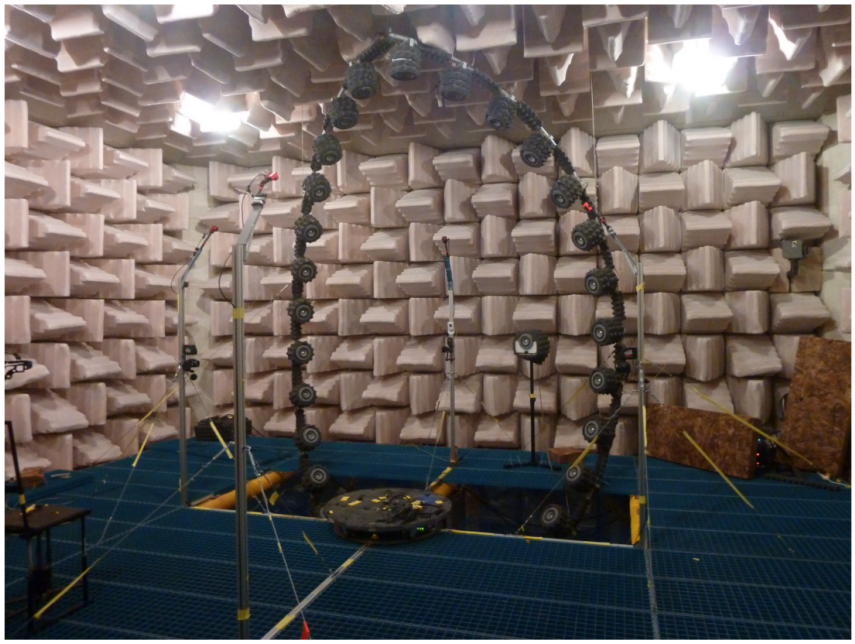


Figure 2.6 – HRTF measurement setup used at Orange Labs: two vertical arcs and a turntable on which the subject is seated. Picture reproduced from [Rugeles Ospina16].

is expensive and hardly transportable. A more detailed presentation of measurement setups and their respective benefits and constraints can be found in Rugeles's PhD Thesis [Rugeles Ospina16, Chap. 3, Sec. 1].

Measurement time

Another major disadvantage of the method is the time needed to measure the HRTFs for thousands of directions. Indeed, between a few minutes and a couple of hours depending on the method, the subject is supposed to remain still for that duration, which is difficult and highly uncomfortable.

The historical approach, which consists in measuring the HRIRs one direction at a time, takes up to 1h45 on a modern setup such as the IRCAM's [Carpentier14]. It is however often sped up by means of interleaved multiple sweep sines, as proposed by Majdak *et al.* in 2007 [Majdak07]. Using this method, Rugeles [Rugeles Ospina16] reports a recording duration of 20 min on Orange Labs' setup.

To further reduce the measurement time, Zotkin *et al.* [Zotkin06] propose in 2006 to swap microphones and loudspeakers based on the acoustic reciprocity principle in order to speed up the measurement session. Although this approach shows good agreement with

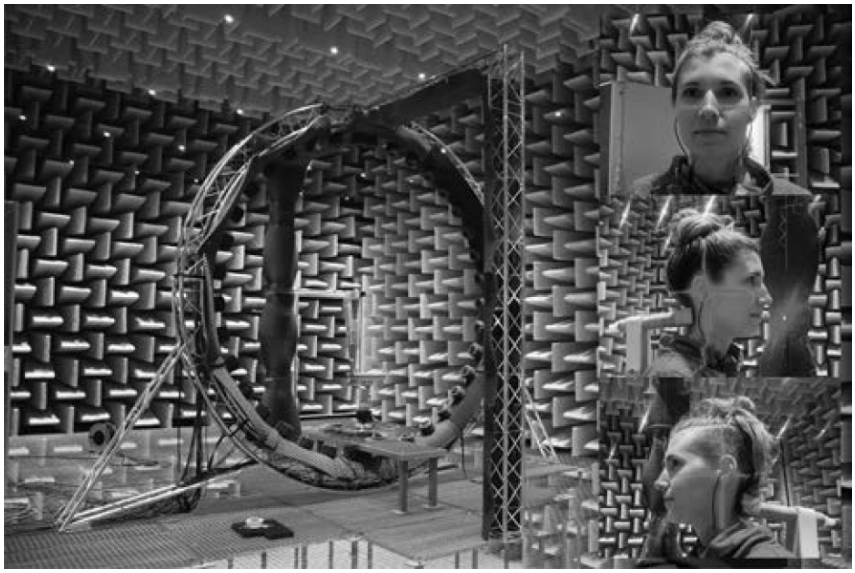


Figure 2.7 – HRTF measurement setup used at the Technical University of Berlin by Brinkmann *et al.* for the constitution of the HUTUBS database [Brinkmann19]. Picture reproduced from [Brinkmann19].

more conventional measurements, it has disadvantages that are inescapable. Indeed, the fact that the loudspeakers are near the subject's ear drums leads to major constraints. First, the size of the in-ear loudspeakers is highly constrained. Second, the sound level of the impulses have to be kept low to preserve the subject's audition, resulting in poor signal-to-noise ratio, particularly at low frequencies. As a consequence, these disadvantages outweigh the benefits of this approach [Matsunaga10]. Let us point out, however, that this method has proven very useful in the context of numerical simulations, as we will see in Section 2.3.2.

An alternative to conventional HRIR measurement is proposed in 2008 by Enzner [Enzner08]. By means of adaptive filtering and a continuous azimuth-wise rotation of the subject, this new paradigm allows the measurement time to be reduced down to a few minutes (2 and 5 min for Rothbucher *et al.* [Rothbucher13] and Brinkmann *et al.* [Brinkmann19], respectively). In an objective and subjective comparison with conventional measurements, Rothbucher *et al.* [Rothbucher13] confirm the quality of such measurements, reporting only a slight degradation in the signal-to-noise ratio, not audible according to the subjective evaluation. This method was recently used by Brinkmann *et al.* [Brinkmann19] to measure the HRIRs of 96 subjects for the HUTUBS database.

Directional imprecision due to subject movement

Measurement time exacerbates another issue: as reported in 2010 by Hirahara *et al.* [Hirahara10] the subject cannot stay completely still all the way through the measurement session, which is a source of errors about the actual direction of the measured HRTFs.

Nevertheless, studies from 2010 and 2017 [Majdak10; Denk17] seem to have successfully limited the subject's movements by giving him a visual feedback. Denk *et al.* [Denk17], in particular, report the directional error to be imperceptible with HRTFs measured using their setup.

Using the same principle of adaptive filtering as the one used for continuous-azimuth HRIR measurements, Ranjan *et al.* [Ranjan16] propose an experimental method that aims at avoiding this issue altogether by recording the HRIRs in a context of unconstrained head rotations. However, the method was only tested on synthetic data derived from the CIPIC dataset.

Reproducibility

Although acoustic measurement is the state-of-the-art method, it should not be considered as perfectly accurate. Indeed, potential inaccuracies become apparent when looking into the reproducibility of HRIR measurements.

Intra-database Measurements are subject to variations from one occurrence to the other, even when the setup and the subject stay the same. In [Riederer98], Riederer investigates thoroughly the influence of various factors on the repeatability of HRIR measurements in a well-controlled environment. The factors under test include reflections from the equipment, microphone placement, head position, clothes and hair. In ideal conditions, i.e. a dummy head with built-in microphones, the author reports an excellent agreement between two independent measurements (spectral differences below 1 dB). In contrast, the factors under study are reported to induce non-negligible variations: up to 2 dB below 6 kHz and between 3 and 5 dB below 10 kHz. Moreover, as this factors were studied one by one, larger variations are to be expected when they combine in real measurement sessions.

Inter-database Much larger variations are observed between databases. For instance, as part of the "Club Fritz" project, Andreopoulou *et al.* [Andreopoulou15] compare 12 different measurements from 10 laboratories of the HRTF set of the Neumann KU-100

manikin. The same pair of microphones, built in the artificial head, was used in all the measurements. Looking at the ITD, they report worrisome variations of up to 235 µs, well above the JND (about 10 µs, see Chapter 1, Section 1.1.5). As to the magnitude spectrum, considerable variations are observed: between 1.4 and 22 dB for the frontal position, and between 2.5 and 19 dB for the rear one. Additionally, left-right asymmetries are noted as well.

Perceptual assessment

For the last 30 years binaural synthesis with individual measured HRTFs has been extensively compared with real free-field sound sources. The vast majority of studies on the subject consist in localization experiments.

While some of such studies [Møller96; Langendijk99; Martin01] report equivalent localization performances, a number of others [Wightman89a; Bronkhorst95; Blauert98; Carlile98] report worse localization performance with virtual sources than with real ones. First, the confusion rate increases by a factor 2 with virtual sources [Wightman89a; Bronkhorst95; Blauert98; Carlile98]. For instance, it goes from 6 % to 11 % for Wightman *et al.* [Wightman89a], and from 21 % to 41 % for Bronkhorst [Bronkhorst95]. Second, somewhat poorer vertical localization is observed with virtual sources than real ones [Bronkhorst95; Blauert98]. For instance, Bronkhorst [Bronkhorst95] reports the vertical variability to have increased from 8° to 13°. In contrast, provided that confusions are resolved, the horizontal accuracy – only related to the ITD – is equivalent with virtual and real sources [Wightman89a; Bronkhorst95].

As to the cause of the observed degradations, no definite answer was found. As Wightman *et al.* [Wightman89a] suggest, small dynamic clues (absent from their binaural synthesis condition) could impact the real-source condition favorably. Bronkhorst [Bronkhorst95], on the other hand, tends to attribute it to microphone positioning and sound source position variability. Indeed, as we have seen in the two previous sections, some inaccuracies are inevitable when measuring HRTFs, due to various factors including microphone positioning or accidental movement from the subject.

2.3.2 Numerical Simulation

An alternative to measurements is to simulate numerically the propagation of acoustic waves. Its main advantages over HRTF measurement are mobility and user comfort. In-

deed, only a 3D scan of the listener is needed for individualization, which results in a much less tedious acquisition session than acoustic measurement. Moreover, once the 3D geometry is acquired, the simulation procedure is completely repeatable and free of measurement noise. Thus, it holds a large potential to better understand the inter-individual variations in HRTFs. Furthermore, a low-cost version can be made available to the end user by using 2D-to-3D reconstruction techniques, thus reducing the acquisition requirements to a set of consumer-grade 2-D pictures [Kaneko16b; Ghorbal16; Mäkivirta20].

Methods

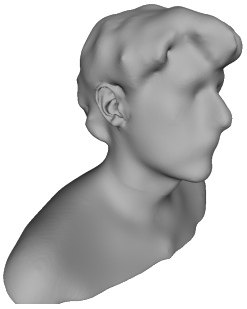
Thanks to the technological advances in terms of computing power, several research teams proposed to numerically simulate HRTFs in the early 2000s. Three approaches can be distinguished: the boundary element method (BEM) [Kahana99; Katz01; Otani03; Greff07] and the finite element method (FEM) [Kahana99; Huttunen07; Farahikia17] in the harmonic domain, and the finite difference time domain method (FDTD) [Xiao03; Mokhtari07; Prepeliță16] in the time domain.

To this day, the most popular technique is the fast-multipole-accelerated boundary element method (FM-BEM) [Gumerov07; Kreuzer09; Huttunen13; Rui13; Jin14; Ghorbal17]. Introduced in 2007 by Gumerov *et al.* [Gumerov07], it owes its popularity to competitive computing times and to the release in 2015 of the *Mesh2HRTF* open-source simulation software by the Acoustics Research Institute (ARI) [Ziegelwanger15a]. This is the technique used for numerical simulations in the present thesis (see Chapter 3, Section 3.1).

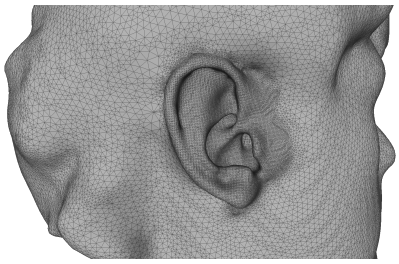
Alternative approaches include a sped-up version of the FDTD called the adaptive rectangular decomposition (ARD) [Meshram14], and the more exotic differential pressure synthesis (DPS) [Tao03] and ray-tracing techniques [Röber06].

3D geometry acquisition

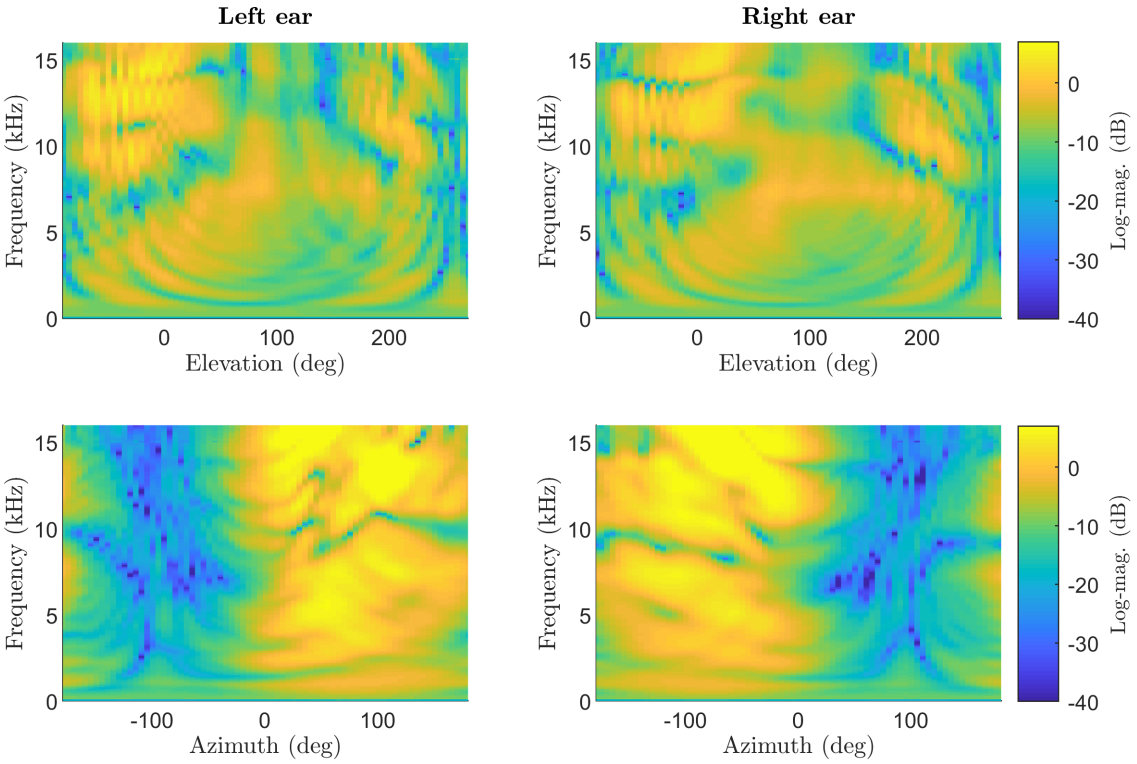
A major topic of interest for HRTF calculation is the accuracy of the 3D geometry of the head, pinnae and torso, starting with acquisition. Let us note that the problem of 3-D surface accuracy lies mostly in the pinnae. Indeed, their shape is complex – with various convolutions and occlusions, and have an important impact on perceptually-sensitive high-frequency HRTF content. In contrast, the head and torso are much simpler shapes and are easier to acquire.



(a) 3-D scan of a human subject (wide shot).



(b) Close shot of the same mesh, edges apparent.



(c) Log-magnitude HRTFs of the left and right ears, in the median (top) and horizontal (bottom) planes.

Figure 2.8 – Exemplary 3-D scan of a human subject acquired at 3D Sound Labs (a, b) and corresponding HRTFs calculated by means of FM-BEM using the *Mesh2HRTF* software (c).
The head and torso were acquired by means of a Microsoft Kinect, the pinnae by means of the United Sciences eFit Scanner for a higher resolution (visible in (b)). Both meshes were then merged.

Often, the pinnae are scanned separately and more precisely than the rest of the morphology, then combined with a rougher scan of the head and/or torso by a human operator [Ziegelwanger14b; Kaneko16a; Brinkmann19]. In our experience, this step can take up to dozens of minutes of manual labor.

MRI & CT Magnetic resonance imagery (MRI) [Mokhtari07; Jin14] and computerized tomography (CT) scan [Turku08] have often been used to acquire pinnae, head and torso geometries for HRTF calculation, especially in early work. While these methods have the advantage of not being sensitive to occlusions, the 3-D surface is deduced from the data by means of a segmentation process which may be a source of errors. In order to attain better accuracy, some [Reichinger13; Ziegelwanger14b; Kaneko16a] have performed CT scans of negative impressions of the pinnae. These silicone or plaster molds being constituted of high-contrast material, it is then easier to extract an accurate 3-D surface. Reichinger *et al.* [Reichinger13] use this method as ground truth in their comparison of various scanning methods.

Although interesting for research purposes, these hospital-grade scanning methods are not suited for an end-user purpose, for obvious reasons of cost and accessibility.

Structured light & laser scanners Structured light- or laser- based devices are a good alternative. Indeed, they are much more practical, some of them being hand-held. Among the numerous commercial options that exist, let us quote the eFit Scanner by United Sciences[4] – which we used in this thesis (see Chapter 3) – and the Artec Space Spider Scanner[5] (used to build the HUTUBS dataset [Brinkmann19]) regarding hand-held devices, and the GOM ATOS-I[6] Scanner (used to build the FABIAN dataset [Brinkmann17]) regarding stationary ones.

Photogrammetry Finally, benefiting from technical advances in the domain of photogrammetry, a recent trend has been to reconstruct 3-D morphology from 2-D pictures. Although, as we will see below, this technique is not very inaccurate, its holds an inescapable potential in its practicality: being able to acquire one's 3-D shape thanks to a few pictures or a video clip taken by means of a smartphone. Commercial applications have already emerged at Genelec [Mäkivirta20] and 3D Sound Labs [Ghorbal20],

[4] `http://www.unitedsciences.com/efit-scanner/`
[5] `https://www.artec3d.com/portable-3d-scanners/artec-spider`
[6] `https://www.gom.com/`

both proposing to reconstruct 3-D morphology from 2-D pictures (a video for the former and a few pictures for the latter) then calculating the corresponding HRTF set. Some use rather conventional photogrammetry methods [Reichinger13; Brinkmann17] such as structure-from-motion (SFM) [Mäkivirta20], whereas some rely on statistical modeling. For instance, Ghorbal [Ghorbal20] fits a PCA model of 3-D ear shape onto a set of pictures, while Kaneko *et al.* [Kaneko16b] perform a non-linear regression between 2-D pictures and 3-D ear shape PC weights by means of a convolutional neural network.

Accuracy All these methods provide 3-D morphological scans, with different accuracies and various impacts on the resulting HRTFs.

In [Reichinger13], Reichinger *et al.* compare the geometric accuracy of 6 scanning approaches on the left and right pinnae of 3 human subjects and on plaster molds of them. The 6 approaches under study are 2 hand-held laser scanners, a hand-held laser scanner coupled to a depth sensor, a stationary and a hand-held structured light scanners, and a photogrammetry commercial software[7]. In addition, CT-scanning of a silicone mold of the pinna is considered as ground truth.

The authors report that the lowest deviations were achieved with two of the hand-held laser scanners and the stationary structured-light one, and that photogrammetry performed worse than all other scanners. In particular, large deviations tend to occur in the narrow cavities of the pinnae. This is problematic, knowing the impact of resonances in such cavities on the resulting PRTFs and HRTFs [Takemoto12]. On another note, considerably lower variations from the ground truth are reported with plaster molds, highlighting the challenge of scanning *in vivo* pinnae. Finally, the authors point out that the scanning results depend on many factors such as the skill of the scanning operator, and that the reliability of the processes ought to be further studied by repeating them. However, that work does not study the impact of the geometrics inaccuracies on the resulting HRTFs. It should be kept in mind that 3-D scanning and photogrammetry technologies are subject to a rapid evolution, and that some of these results may be outdated.

In a recent study, Dinakaran *et al.* [Dinakaran18] compare three state-of-the-art structured light scanning devices, including the Artec Space Spider and GOM ATOS-I, and three low-cost alternatives: the Microsoft Kinect depth sensor, and two photogrammetric methods, different from the one studied in [Reichinger13]. For each method, the FABIAN

[7] Agisoft PhotoScan 0.8.5 Build 1423: `https://www.agisoft.com/`

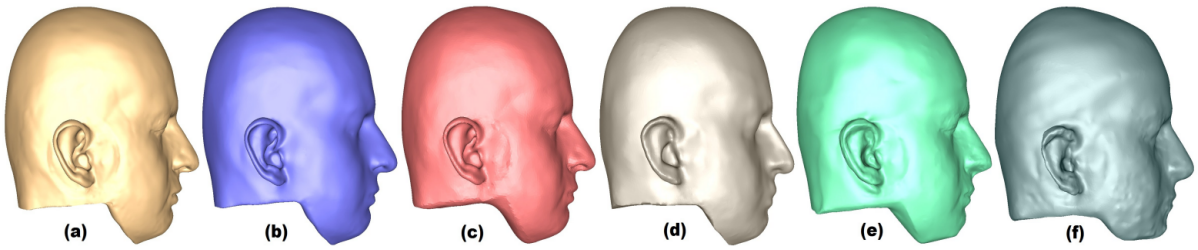


Figure 2.9 – 3-D meshes of a human subject acquired with 6 different scanning methods by Dinakaran *et al.*. (a) GOM ATOS-I Scanner, (b) Artec Space Spider Scanner, (c) Canfield Vectra M3 scanner, (d) Microsoft Kinect scanner, (e) Autodesk 123D (photogrammetry smartphone app), and (f) Python Photogrammetry Tool. Reproduced from [Dinakaran18].

dummy head [Brinkmann17] is scanned, then the corresponding HRTF set is computed. The comparison is then done on three levels: geometric, acoustic, and perceptual, i.e. on the meshes, on the HRTFs and on localization results predicted thanks to the Baumgartner auditory model (see Section 2.2.2 and [Baumgartner14]). The authors report outstanding agreement between the three structured-light methods on all three levels of comparison, including the pseudo-perceptual one, with differences of PEs and QEs below 0.4 % and 0.7°, respectively. In particular, it is worth noting that the hand-held device is on a par with the stationary ones. With the other methods, a notable loss of details is observed, particularly in the fine structure of the pinnae. While this has only a minor influence on the overall spectral shape of the HRTFs, this degradation has a strong impact on the predicted localization performance: differences of PEs and QEs (see Section 2.2.2) between 6° and 12° and between 4 % and 6 %, are reported, respectively.

Overall, although no conclusion can be drawn as to a potential absolute reference, Dinakaran *et al.* demonstrates that 3 different structured light-based methods are in excellent agreement. In particular, a hand-held device is shown to be as accurate as stationary ones, which is a great point for practicality. According to both [Reichinger13] and [Dinakaran18], photogrammetry methods seem to deviate considerably from other scanning methods. In particular, Dinakaran *et al.* show that the geometric inaccuracies result in high perceptual deviations. Photogrammetry is however an interesting low-cost and user-friendly approach which may well improve in the future with technical advances.

Mesh grading Another major matter in HRTF calculation concerns the re-sampling – also called mesh grading – of the 3-D geometry prior to simulation. Regarding BEM, in

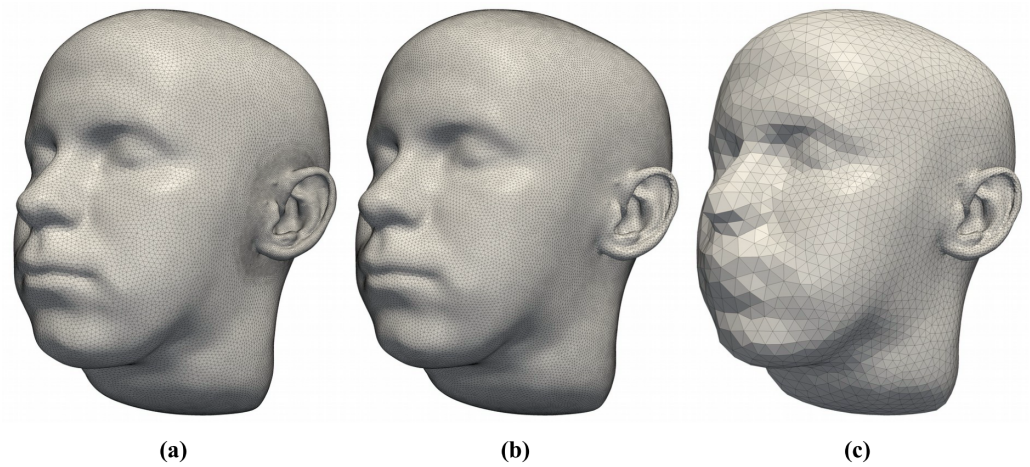


Figure 2.10 – Reference 3-D mesh (a), uniform grading with an average edge length of 2 mm (b), and progressive grading with an edge length ranging from 1 to 25 mm (c). Reproduced from [Ziegelwanger14c].

particular, the surface mesh must be re-arranged so that it is regular enough and so the edge lengths are small enough in regard to the simulation's wavelength. As computing time increases considerably with the number of mesh elements, the re-meshing resolution is a trade-off between numerical accuracy and computing time.

Although the use of the six-elements-per-wavelength rule [Marburg02] has been widespread, the Acoustics Research Institute has recently well contributed to the subject. Indeed, by studying the effect of various average edge lengths (AEL) on the resulting HRTFs, objectively and subjectively, Ziegelwanger *et al.* [Ziegelwanger15b] determine that the optimal resolution for uniform re-meshing is an AEL of 1 mm.

Going further, in their 2016 study [Ziegelwanger16], they implement and compare various re-meshing methods, demonstrating that a progressive approach is appropriate and desirable. Indeed, making the mesh fine (AEL $\simeq$ 1 mm) near the ear canal and coarser the further away from it allows a factor-10 decrease in the computing cost of FM-BEM simulation while maintaining HRTF accuracy. Their code was made available on-line along with their HRTF simulation software *Mesh2HRTF*[8] [Ziegelwanger15a]. The effect of both uniform and progressive mesh gradings are shown in Figure 2.10.

In the case of FDTD simulation, similar work has been carried out through the study of the impact of the voxelization of a subject's volumetric geometry on the resulting HRTFs [Prepeliță16].

[8] `https://sourceforge.net/projects/mesh2hrtf/`.

Computing time

Computing time used to be the main drawback of HRTF calculation. Indeed, up until 2007 [Huttunen07; Mokhtari07], computations of HRTFs on the whole audible frequency range were scarce. For instance, in pioneering work by Katz *et al.* [Katz01] in 2001, the BEM calculation of HRTF set is limited to a frequency range of 1 kHz to 5.4 kHz, and took 28 hours for 54 regularly-spaced frequencies and a single ear. The author extrapolates that, using his setup, it would take more than 5 years to compute an HRTF set for frequencies up to 20 kHz for both ears.

Computing times have however greatly been reduced since then. While the exponential decrease in the cost of CPU power and RAM is certainly a major factor, several technical advances have had a major part in this reduction.

One of these advances was the introduction by Gumerov *et al.* of FM-BEM in 2007 [Gumerov07]. In their 2010 study, Gumerov *et al.* [Gumerov10] report that the FM-BEM computation of a single-ear HRTF set of a mesh that includes the torso takes 30 h for 117 frequencies ranging from 172 Hz to 20.155 kHz.

The work by Ziegelwanger *et al.* [Ziegelwanger16] on progressive mesh grading constitutes another major step forward, as it reported to permit a factor-10 decrease in the computing load (see previous paragraph).

The democratization of distributed computing on clusters over the last decade and the constant increase in available computing power have further decreased the computing times. Indeed, simulations in the harmonic domain such as the FEM or BEM are highly distributable, as each frequency is simulated independently. Although in theory hundreds of frequencies could be computed simultaneously, parallelization is generally limited by high memory requirements (especially at high frequencies), as the memory is often shared by the parallel threads. As early as 2007, Huttunen *et al.* [Huttunen07] distribute FEM simulations on a PC cluster of 22 CPU cores and 44 GB total RAM. They report computing times ranging from a few tens of seconds at 20 Hz to 2.5 h at 20 kHz, and extrapolate that a complete HRTF set with relatively low high-frequency resolution (500 Hz steps for frequencies above 13 kHz) could be computed in a few days. By distributing FM-BEM computations on a cluster of 5 PCs with 2-core CPUs, Kreuzer *et al.* [Kreuzer09] report in 2009 to have computed a complete single-ear HRTF set in 5 hours for 100 frequencies ranging from 200 Hz to 20 kHz. Recently, Fan *et al.* [Fan19] have implemented a GPU-distributed version of conventional BEM and used it to compute HRTFs. Due to limitations in global GPU memory, computations for a mesh with torso were limited to

an upper frequency of 12 kHz. The authors report computation times of 5 to 7.5 seconds per frequency and of 12.8 to 21.5 seconds per frequency for a mesh without and with torso, respectively. It is however unclear whether these per-frequency computation times are averaged over all frequencies, or if they correspond to a specific frequency.

With our own numerical simulation setup (see Chapter 3), the calculation of a complete PRTF set by means of FM-BEM was distributed over 10 cores of a desktop workstation of 12 CPU cores and 32 GB of RAM. Not all 12 cores could be exploited due to limitations in the memory, shared by all threads. The computation of one PRTF set from a pinna mesh (up to about 55000 triangular faces at the highest frequency i.e. 16 kHz) was achieved in 1 hour, with computing times ranging from 4 s to 5 min per frequency. For a complete torso – a substantially larger mesh (up to about 110000 triangular faces at the highest frequency i.e. 16 kHz), using the same setup, the calculation of an HRTF set is distributed on only 5 CPU cores and takes about 10 hours, with computing times ranging from 40 s to 45 min per frequency.

Comparison with measurements

Several studies compare calculated HRTF sets to acoustically measured ones [Greff07; Kreuzer09; Gumerov10; Ziegelwanger13; Brinkmann19] and agree on the following. While the shapes of the spectral patterns are overall coherent and while there is good agreement below 5 to 7 kHz, large mismatches are observed at higher frequencies. In particular, local spatial-spectral features such as notches and peaks – known to be important features for elevation perception – are impacted, being displaced (in space and/or frequency), attenuated and sometimes absent [Greff07; Kreuzer09; Gumerov10].

Greff *et al.* [Greff07] compare two different calculations (carried out by different teams) and a measurement of the HRTF set of a dummy-head manikin. They find that, in the frontal position, the two calculated HRTF show minimal spectral variations between each other, but both exhibit a frequency shift above 5 kHz compared to the measurement. In terms of ITD, good agreement is obtained between all three methods.

Such deviations are also reported by Brinkmann *et al.* [Brinkmann19] in a larger-scale study, in which the calculated and measured HRTF sets of 96 human subjects of the HUTUBS database are compared. The authors report an average spectral difference of less than 1 dB below 5 kHz and of up to 7 dB at 17.1 kHz. Differences in ITD are reported to be lower than the JND of 20 µs for most sound source positions and subjects. Going further, the authors asked 46 subjects to participate in a rating experiment which aimed

at comparing both types of HRTF sets based on 12 criteria. The listeners were generally were able to discriminate computed and measured HRTFs. In particular, large differences in coloration were perceived, with emphasized high- and attenuated low-frequencies for computed HRTFs. The authors note that, indeed, simulated HRTFs contain on average more high-frequency energy than measured ones. Regarding localization, an elevation shift of 12° upwards and an azimuth shift of 2° clockwise were reported. According to the authors, the former might be partially explained by the high-frequency boost.

Beside [Brinkmann19], a few studies evaluate computed HRTFs perceptually [Turku08; Ziegelwanger15b; Fan19]. However, among them, only one [Ziegelwanger15b] concerns individual HRTF sets of human subjects. In that study, Ziegelwanger *et al.* study various simulation settings such as mesh grading or source position. They evaluate the localization performance in the horizontal and median planes of 3 subjects presented with their computed own HRTF set and their own acoustically measured one (from the ARI database). With the setting that performed best, they report localization performances with computed HRTF sets to be on a par with measured ones. However, these results should be taken with caution seeing that only 3 subjects participated in the study.

Overall, the perceptual relevance of computed HRTFs remains to be demonstrated. Indeed, notable spectral mismatches are generally observed between computed and measured HRTFs, potentially affecting features that are useful for vertical localization. Nevertheless, acoustic measurement is no absolute reference (see Section 2.3.1) and perceptual assessment ought to be the ultimate criteria. However, perceptual studies are conspicuous by their scarcity and present mitigated results. While Ziegelwanger *et al.* [Ziegelwanger15b] report localization performances with computed HRTFs to be as good as with measured ones, the study includes too few subjects to be really conclusive. Furthermore, according to Brinkmann *et al.*'s [Brinkmann19] 46-subject rating experiment, listeners consistently discriminate computed and measured HRTF, localize differently and report a different timbre colorations.

2.3.3 Indirect Individualization based on Morphological Data

Though more convenient than acoustic measurement, HRTF calculation still requires specialized equipment and non-negligible mesh processing and computing time. Hence, based on the idea that the individual character of HRTFs derives from morphological differences, many studies have explored the idea of a low-cost HRTF individualization based on simple morphological data such as anthropometric measurements.

Selection

One way to tackle the problem is to select the most suited non-individual HRTF set among a database.

Using the 46-subject CIPIC database (see Section 2.4, [Algazi01c]), Zotkin *et al.* [Zotkin02] propose to select the HRTF set associated with the anthropometric nearest neighbor. The latter is determined based on 7 morphological parameters measured on a picture of the pinna. According to their 6-subject localization experiment, for 4 of the subjects, the elevation error is lower by 15-20 % with the best-fit HRTF set than with a generic one (the HRTF set of a listener who did not participate in the experiment). However, for the 2 remaining subjects, the error is either lower by only 5 % or considerably higher (by 73 %), highlighting highly variable performances. Averaging the results over the 6 subjects ourselves, we find an elevation error decrease of only 0.7° with a standard deviation of 3.1°. Regarding the azimuthal dimension, we find a notable degradation, with an average error increase of 3.3° (with a standard deviation of 3.3°). The latter is not commented by the authors but is somewhat expected, seeing that the HRTF selection process relies on the dimensions of the pinna but not of the head. Overall, the localization results are hardly conclusive and, as the authors point out, a larger-scale perceptual study would be needed.

In [Schönstein10] Schönstein *et al.* propose to select an HRTF set among 37 from the LISTEN database (see Section 2.4, [Warusfel03]) based on a set of 5 anthropometric parameters. To do so, a multilinear regression is performed between the 37 sets of morphological measurements and a compact representation of the matching HRTF sets. Two methods are considered to create this compact representation: PCA of linear magnitude HRTF sets performed in the inter-individual fashion (see Section 2.1.3), and MDS (multidimensional scaling) of global frequency scaling factors which are used to characterize spectral dissimilarity between HRTF sets as in [Middlebrooks99a]. To evaluate their method, they compare its rankings of HRTF sets to a rating (*bad*/*ok*/*excellent*) of the 46 LISTEN HRTF sets established in a previous study [Katz12] by 45 of the subjects by means of listening tests. In particular, they look at the proportion of *excellent* ratings in the HRTF sets ranked among the first 10 by the method. With regard to that metric, they find that their method outperforms the random selection of 10 HRTF sets for 26 out of 37 subjects.

Recently, Yao *et al.* [Yao17] have proposed a concurrent method that relies on a neural network trained to predict a perceptual score from anthropometric measurements.

30 subjects were asked to rate 18 HRTF sets of the CIPIC database on a scale from 1 to 5 according to two criteria: front-back and elevation discrimination. The two rating scores are then combined into one by computing the mean, thus giving a perceptual score for each subject and HRTF set. A single- or double-hidden-layer neural network was then trained to predict this localization score from 10 anthropometric measurements of the head and pinnae. Thus, when presenting the neural network with a new set of anthropometric measurements, one or several best-fit non-individual HRTF sets are presented to the user based on the predicted perceptual score. Evaluating the method by means of a leave-one-out cross-validation, they compare the performance of both neural networks with that of Zotkin *et al.*'s approach. They find that the two former outperform the latter. In particular, the "target" HRTF set (i.e. the one with the best perceptual score) is found to be among the 3 predicted best-fit HRTF sets for 40 % of the 18 subjects for both neural network methods, against 23.3 % for Zotkin *et al.*'s.

Adaptation

Complementary to the selection of a best-fit non-individual HRTF set in a database, a generic HRTF set can be adapted to the user by means of rudimentary transformations.

Based on the idea that a variation in pinna size results in a frequency scaling of the corresponding spectral features in the HRTFs from all directions, Middlebrooks *et al.* [Middlebrooks99a; Middlebrooks00] propose a rough adaptation of a generic HRTF set by means of a global frequency scaling. Three methods of determining the optimal scaling factor are compared: best spectral match in terms of ISSD (see Section 2.2.1), linear regression from 9 morphological measurements of the head and pinna, and tuning by the listener. In [Middlebrooks99a], an objective comparison between the two former is presented, for 33 subjects of a proprietary database. The authors report that the acoustic optimal scaling factor could be retrieved from only pinna height and head width with a correlation factor of 0.89 and RMS error of 0.069. In addition to a frequency scaling, the ISSD can be further reduced by applying a head tilt to the HRTF set. On average over all 990 pairwise comparisons, average ISSDs of 8.29 dB^2, 6.18 dB^2 and 5.37 dB^2 are reported for HRTF sets without adaptation, with scaling, and with scaling and head tilt, respectively.

In the companion paper [Middlebrooks00], the listener-driven method is presented and compared to the two former by means of localization experiments with 5 subjects. Three non-individual HRTF sets are used for this comparison, chosen so as to span the range of

optimal frequency scalings observed in their dataset of 33 measured HRTF sets. Regarding the subjective tuning procedure, the listeners were randomly presented with virtual sound sources located in the median plane. During each of 240 trials, knowing the sound source's position, they elected in an A/B comparison one of two scaled HRTFs according to criteria of front-back discrimination, primarily, and elevation accuracy, secondarily. The process lasted about one hour. A good agreement was obtained between the scaling factors obtained *via* tuning and the acoustic ones, with a correlation of 0.89 and a RMS error of 0.069. Regarding the perceptual assessment of scaled generic HRTF sets, a notable decrease in quadrant error (QE, see Section 2.2.2) of more than half the difference between own and raw generic is reported in 7 cases out of 9 (each of the 5 listeners listened to one or two non-individual HRTF sets for a total of 9 cases). However, for the two remaining subjects, the QE increases. Local angular accuracy is not evaluated.

Later on, other researchers [Maki05; Guillon08] also propose to apply a combination of frequency scaling and rotation to adapt a generic HRTF set. In particular, Guillon *et al.* [Guillon08] derives the frequency scaling and rotation parameters from 3-D scans of the head and pinnae for 6 subjects. However, neither study include a perceptual evaluation of the resulting HRTF sets.

Regression

The methods reviewed above aim at reducing perceptual discrepancies due to non-individual HRTF sets by rudimentary means which do not embrace the full complexity of the inter-individual variations of HRTFs. They thus cannot pretend to provide an HRTF set whose perceptual quality would come close to individual conditions. Hence, a lot of work has relied on statistical modeling and regression to synthesize individualized HRTF sets from anthropometric measurements.

To this end, an approach that has widely been used since the early 2000s is to perform a regression between a set of 8 to 93 heuristically-chosen morphological measurements and the corresponding HRTF sets.

While most – especially early – work use multiple linear regression [Jin00; Hu06; Huang09b; Hugeng10; Bomhardt16a; Liu19b] or other linear methods [Bilinski14] to link the anthropometric and acoustic spaces, others use non-linear techniques such as support vector regression (SVR) [Huang09b] and neural networks [Hu08; Li13; Grijalva14; Fayek17; Qi18; Zhang20]. Due to their high dimensionality, the HRTF sets are typically "compressed" prior to regression, by means of PCA [Jin00; Hu06; Hu08; Huang09a;

Hugeng10; Bomhardt16a; Fayek17; Zhang20], independent component analysis (ICA) [Huang09b; Liu19b], sparsity-constrained weight mapping (SWM) [Qi18], high-order singular value decomposition [Li13] or Isomap [Grijalva14].

Often, a subset of "key" morphological measurements is selected based on their statistical relevance [Hu08; Huang09a; Zhang20] prior to regression. Sometimes, they are represented as weights of a statistical model such as PCA [Jin00] or factor analysis [Liu19b]. However, the number and choice of the parameters is limited by the dataset. In the vast majority of cases, the dataset is CIPIC, which includes 27 heuristically-defined anthropometric features, measured from a 2-D picture. It thus seems legitimate to question the accuracy of these measurements and the choice of only 27 parameters, in particular for a complex 3-D shape such as the pinna. This issue is however barely addressed in the literature, although sometimes mentioned in the few studies in which a different dataset is used: [Bilinski14] and [Bomhardt16a] and their respective 96 and 12 measurements made from 3-D meshes, and [Jin00] and their 20 measurements made with a 3D stylus pen.

Regarding perceptual assessment, among the 15 aforementioned studies, 4 provide localization experiments. In their first study based on multiple linear regression, Hu *et al.* [Hu06] compare the localization performance of 5 subjects presented with their customized HRTF set and with a non-individual HRTF set, that of CIPIC's subject 003. The results show a modest advantage for the customized condition, with an average rate of correct answers of 79.2 % against 61 % and an average rate of front-back confusions of 10.8 % and 11.7 %. The variance of these results is however not reported and only horizontal positions are under test. In their later study based on a three-layer neural network, Hu *et al.* [Hu08] perform a similar 5-subject localization experiment. They report a slightly better result than the previous study, with average rates of correct answers of 75.2 % and 56.1 %, and front-back confusion rates of 9.7 % and 12.2 % for the customized and CIPIC 003 HRTF sets, respectively. In this study as well, the variance is not reported and only horizontal positions are under test. Liu *et al.* [Liu19b] perform a 6-subject localization experiment in 6 directions of the median plane for the customized and the KEMAR HRTF sets. They report an improvement in localization with the customized condition over the KEMAR one, with respective front-back confusion rates of 5.1 % and 10.6 % and respective up-down confusion rates of 6.9 % and 10.2 %. Finally, Zhang *et al.* [Zhang20] perform a localization experiment with 5 subjects for 3 directions of the median plane. For all 3 directions, a statistically significant decrease in angular error is observed between KEMAR and customized HRTF sets. A statistically significant difference in front-back

confusions is observed only for one of the elevations (at 22.5°), at the advantage of the customized condition. Additionally, simulating localization experiments for all directions of the median plane thanks to the Baumgartner auditory model, they report a statistically significant improvement in angular error for all directions. As an order of magnitude, the average front-back confusion rate and angular error drop from 16.32 % to 11.22 % and from 17.71° to 13.93°.

2.3.4 Indirect Individualization based on Perceptual Feedback

If methods for indirect individualization based on morphological data are practical for the end user, it is doubtful that a few dozens heuristically-defined anthropometric measurements can account for the full complexity of inter-individual HRTF variations. Furthermore, in practical applications, the acquisition of pictures or direct measurements of the subject's morphology is likely to be entrusted to the user, which is an additional source of errors. As subjective perception is the ultimate judge of HRTF quality, an alternative approach is to provide low-cost individualization based on the listener's perceptual feedback.

Selection

A quite straightforward low-cost strategy that has been well explored in the literature since the late 1990s is to help the listener select the best non-individual HRTF set among a database [Seeber03; Iwaya06; Katz12; Zagala20]. While in these three approaches the listener is presented with a sound source moving according to a known trajectory, they differ on several aspects.

First, the selection processes are quite diverse. Seeber *et al.* [Seeber03], for instance, present a 2-step selection of a best-fit non-individual HRTF set: the listener first selects a subset of 5 HRTF sets among 12 according to a broad criterion of "spaciousness", then he chooses the best among the 5 according to criteria of "localization variance" and "externalization". On the other hand, Iwaya *et al.* [Iwaya06] propose a tournament-style selection among 32 non-individual HRTF sets according to a criterion of accuracy of the perceived sound source trajectory. Regarding Katz *et al.*'s study [Katz12], the approach is more holistic and aimed at guiding further work on HRTF selection: 45 subjects were asked to rate 46 HRTF sets from the LISTEN database [Warusfel03] (including their own) as *ok*, *bad* or *excellent*. As in Iwaya *et al.*'s study, the rating criterion was the fidelity

of the virtual sound source trajectory. Best-fit non-individual HRTF sets are thus the ones rated as *excellent*. Tuning times for the procedures ranged from 15 min [Seeber03; Iwaya06] to 35 min [Katz12].

Second, Seeber *et al.* and Iwaya *et al.* [Seeber03; Iwaya06] limit their studies to the horizontal plane, where individualization is less important. Indeed, the lateral localization cues are ITD and ILD, which are more robust to a lack of individualization (see Chapter 1, Section 1.3.2). In contrast, in [Katz12] both vertical and horizontal trajectories are presented to the listener.

Regarding perceptual assessment, all three studies perform localization experiments. In a 10-subject experiment with sources on the frontal horizontal arc, Seeber *et al.* [Seeber03] report an average azimuth error close to that observed with real sound sources (difference of 1 %). In their evaluation with 7 subjects, Iwaya *et al.* [Iwaya06] compare individual, best-fit non-individual, and worst-fit individual HRTF set. They report front-back confusion rates of about 5 %, 7 % and 12 %, respectively, the difference between best- and worst-fit being statistically significant. Regarding Katz *et al.*'s [Katz12] 7-subject localization experiment for the individual, best- and worst-fit HRTF sets, they report respective average front-back and up-down confusion rates of 20, 32 and 35 %, and 13, 15 and 19 %. While there is still an improvement from worst- to best-fit, unlike in Iwaya *et al.*'s study, the best-fit performance is closer to the worst-fit one than to the individual one. This difference might be partially explained by the fact that the individualization problem is harder when the vertical dimension is included. Also, Katz *et al.*'s three-degree rating process might be less selective than a tournament approach.

Conjointly, in order to improve the relevance and duration of selection procedures, it has been proposed to cluster *a priori* the database based on either objective [Xie15] or perceptual [Katz12] criteria. In particular, Katz *et al.* [Katz12] show that for a particular subset of 9 HRTF sets (out of 46), 89 % of the subjects would find at least one HRTF set that he had rated as *excellent*.

Following up Katz *et al.*'s study, Zagala *et al.* [Zagala20] propose and compare two different methods of subjective evaluation to rank the 8 representative HRTF sets previously identified in [Katz12]. The first method is a localization task while the second is a judgment task similar to the one employed in [Katz12], which consists in rating global preference of renderings of horizontal and vertical virtual trajectories. 26 listeners participated in the experiments. As discussed in more details in Section 2.2.2, they find that good agreement is obtained between both methods of ranking. The focus of the study

is not on the perceptual performance of the top-ranked HRTF sets. Nonetheless, they report in Appendix B that the best scoring HRTF sets of most subjects yield median unsigned polar errors comparable to those obtained in another study with individual HRTFs [Stitt19], and that the difference between worst- and best-fit non-individual HRTF sets seem to be substantial. They also report that the mean unsigned lateral errors appear to be generally comparable for the worst- and best-fit HRTF sets and for the results from [Stitt19]. However, the statistics behind these statements are not provided, although localization errors for each subject are summarized in Fig. 7. The duration of both ranking procedures were about 25 min.

Adaptation

Complementarily to the selection of a best-fit non-individual HRTF set among a database, a non-individual HRTF set can be roughly adapted in the hope of reducing perceptual discrepancies related to a lack of individualization.

Frequency scaling For instance, a generic HRTF set can be modified by means of a global (i.e. identical for all directions) frequency scaling, as proposed by Middlebrooks *et al.* [Middlebrooks99a; Middlebrooks00]. Three methods are proposed in [Middlebrooks00] to determine the scaling factor: minimal spectral difference, regression from anthropometry and a procedure in which the listener tunes the scaling parameter by ear in about one hour. The scaling factors obtained by all three methods are in good agreement and localization experiments were performed, whose results are somewhat mitigated. For further details on this study and its results, please refer to Section 2.3.3 where it is well covered.

Filter-design-based adaptation Other work [Tan98; Runkle00] have relied on the tuning of filters to further adapt a generic HRTF set previously selected among a database. For instance, Tan *et al.* [Tan98] asked 10 subjects to tune a 5-band filter applied to a generic HRTF with instructions to reduce front-back confusions and elevation mismatch. They report that, only 4 subjects out of 10 experienced front-back confusion after the procedure, against 8/10 initially. The study presents several obvious limitations, starting with a limitation to the frontal position which raises a major question: would the tuning procedure need to be performed for each direction? Furthermore, very little information is given on the tuning procedure (tuning time for instance) and on the localization experiment.

On another note, Runkle *et al.* [Runkle00] propose a framework in which a generic HRTF set is adapted through filtering by a low-order pole-zero filter whose 16 parameters are tuned by a generic algorithm based on perceptual feedback. The perceptual feedback is the result of a subjective evalution performed at each iteration in which the listener rates 8 HRTF sets. However, very little detail is given on the tuning procedure. For instance, it is unclear how different directions are handled: are they tuned globally or one by one? Which directions is the listener presented with during the tuning procedure? Furthermore, results only concern the convergence of the algorithm. No objective or subjective assessment of the produced HRTs is presented and no information is given regarding tuning time. To the best of our knowledge, there is no follow-up publication that would answer these questions.

Synthesis

Although they are able to somewhat reduce the perceptual discrepancies caused by a lack of individualization, the aforementioned approaches are rudimentary and cannot claim to embrace the full complexity of the inter-individual variability of HRTF sets. In contrast, more ambitious approaches propose to synthesize an HRTF set from a statistical model, whose parameters are tuned based on perceptual feedback from the listener.

Statistical-model-based tuning Among these, many consist in a tuning procedure in which the listener is asked to tune by ear the weights of a PCA model of HRTFs [Shin08; Fink15] or HRIRs [Hwang08a]. Only the first 3 to 5 PCs of the model are tuned in order to limit tuning time. However, the duration of the tuning procedure is not reported in any of the three studies. Let us note that, most likely because of the small size of the 46-subject CIPIC dataset (used by all three studies for training), the PCA is performed in the *spectral* fashion defined in Section 2.1.3. As a consequence, a set of PCWs corresponds to one transfer function (or impulse response), and thus the tuning must be performed independently for every direction of interest. For Shin *et al.* and Hwang *et al.* [Shin08; Hwang08a] these directions are in the median plane, whereas for Fink *et al.* [Fink15] they are in the horizontal one. For the latter, a parameter controlling ITD amplitude is tuned in addition to the 5 magnitude HRTF PC weights.

Regarding perceptual assessment, the three studies provide a localization experiment. In Fink *et al.* [Fink15], only one subject participated in the subjective procedure and subsequent localization experiment. They report the front-back confusion rate to be

notably better with the customized HRTF set than with the average HRTF set of the database: 36.25 % against 16.25 %. However, it should be noted that the average HRTF set is somewhat unrealistic: the peaks and notches are smoothed out compared to a "real" HRTF set, likely degrading useful spectral localization cues. Regarding Shin *et al.* and Hwang *et al.* [Shin08; Hwang08a], the customized HRTF set is compared to the individual one and to that of the KEMAR manikin. In the former study, for two subjects out of 4, the front-back confusion rate is notably lower with the two former HRTF conditions (between 6 % and 14 %) than with the latter (between 29 % and 43 %). However, for the 2 remaining subjects, there is no clear trend. The latter study by Hwang *et al.* shows a clearer trend regarding the front-back confusion rates of the KEMAR and individual conditions. Indeed, for all 3 subjects, they are in the order of 20 % and 0-1 % respectively. Regarding the customized condition, they report front-back confusion rates close to the individual condition for two subjects (in the order of 1 % to 3 %), whereas for the remaining subject it is quite high (13.3 %).

In these studies, the tuning is local, in the sense that it is performed independently at each direction of interest. This poses a problem of tuning time, seeing that a high-resolution HRTF set typically contains HRTFs for several hundreds of directions. To alleviate this, in his Master's thesis, Hölzl [Hölzl14] proposes a method to tune an HRTF set globally. Like in the three aforementioned studies, the listener tunes by ear the weights of a PCA model – built in the *spectral* fashion – of magnitude HRTFs. However, instead of tuning the PCWs directly, the listener is asked to tune the coefficients of a spherical harmonics representation (see Section 2.1.2) of the PCWs. Three training sets are used in turn to build the PCA model: LISTEN, CIPIC and ARI.

This global approach was however not put in practice and thus there is no perceptual assessment of the method. Let us note that, although this approach allows a global tuning of an HRTF set, there is no guarantee that these tuning parameters (i.e. SHs of PCWs) result in plausible HRTF spatial patterns. As mentioned by the author, if the SH coefficients are tuned by the user with regard to certain directions, it is unknown whether the tuning will be appropriate for other areas of the sphere.

Recently, in 2017, Yamamoto and Igarashi [Yamamoto17] propose a method that relies on the modeling of HRTF sets thanks to a variational autoencoder neural network. The tuning procedure consists in a gradient descent optimization of the network's weights where, at every iteration, the cost is derived from the user's A/B rating of two HRTF sets presented to him by the algorithm. Here as well, the database used to train the statistical

model is CIPIC. In contrast with the aforementioned approaches, the parameters that are tuned correspond to a complete HRTF set (all directions). The optimization thus explores the space of the inter-individual variations of HRTF sets. The tuning procedure is reported to last 20 to 35 min with about 100-200 pairwise comparisons.

In guise of perceptual assessment, after tuning, the 20 participants are asked to rate HRTF sets pair by pair in a double-blind manner. The baseline condition is a best fit non-individual HRTF set for each participant, selected among the database by means of a previous rating test procedure. The authors report a statistically significant preference of the customized HRTF set over the best-fit non-individual HRTF set for 18 participants out of 20.

Conclusion

Overall, acoustic measurement remains the reference in individual HRTFs acquisition. Indeed, it is the historical approach and the resulting HRTFs have been well compared to real-life sound localization over the years. However, HRTF measurement is far from being flawless. Indeed, it suffers from a lack of reproducibility which translates to large variations in both ITD and magnitude spectra between different measurements setups, but also between repetitions of the same measurements. In particular, when evaluating a VAS generated thanks to individual measured HRTFs, a number of studies observe a degradation of the localization performance compared to a real auditory environment. In these studies, confusion rates are reported to increase by a factor 2 and elevation accuracy to be somewhat degraded. Furthermore, this approach can not be proposed to the end user: besides the uncomfortable nature of the acquisition process for the subject, the measurement setup is delicate, expensive and, most of all, untransportable.

As an alternative, individual HRTFs can be computed from 3-D scans of the listener's pinnae, head and torso by means of numerical simulations of acoustic propagation. Unlike measurement, the data acquisition step can be performed anywhere, in particular when reconstructing 3-D morphology from 2-D pictures. Moreover, it allows to work around, or at least to displace, the reproducibility issue: once the 3-D mesh is acquired, the rest of the simulation process is deterministic. Be that as it may, the quality of computed HRTFs remains to be demonstrated. Indeed, perceptual studies have been scarce and mismatches have been reported in objective comparisons with measured HRTFs. Furthermore, between acquisition, 3-D shape preparation and the simulation itself, the process takes a considerable amount of time (in the order of hours), which may be a serious limitation in

user-friendly applications.

Focusing on the user-friendly constraint, less direct approaches to HRTF individualization have been proposed as well. Many of these approaches rely on anthropometric measurements, either performed manually or derived from one or several 2-D pictures. These morphological parameters can then be used to derive a personalized HRTF set in a variety of ways, such as the selection of a best fit among a database, rough adaption of a generic HRTF set, and linear or non-linear regression. While they have the merit of proposing user-friendly HRTF individualization – taking a few pictures with a smartphone is indeed easy – the quality of the resulting HRTFs can be questioned. Indeed, it is somewhat doubtful that a few dozen measurements can account for the full complexity of the 3-D shape of the pinna and of its directional acoustic filtering effect. This seems to be corroborated by the scarcity of perceptual evaluations of the more ambitious regression-based methods. This scarcity could be partially explained by the fact that regression methods rely on databases which, as we will see in Section 2.4, are small compared to the dimensionality of HRTF sets.

Another family of user-friendly methods rely instead on perceptual feedback from the listener: the user participates in subjective evaluations whose outcomes serve to provide a personalized HRTF set. Methods to achieve this include selection of a best fit among a database, rough adaptation of a generic HRTF set, and tuning of an HRTF model. The two former are basic approaches that cannot claim to provide realistic individual HRTFs, but have shown some perceptual improvement over non-individual conditions. In contrast, the latter are more ambitious and propose to adapt models that embrace the complexity of HRTF variations. Less explored, they often rely on statistical modeling and thus on HRTF databases, whose small size may be an issue (see Section 2.4).

Percept-based methods may be a little less practical for the listener as they require his attention and possibly more of his time. However, it requires little to no specific equipment: the device on which the VAS is rendered (PC, tablet, smartphone *etc.*) is enough in most cases. Furthermore, unlike other approaches, a perceptual assessment of the produced HRTFs is performed throughout the process and even guides it. What is more, a trade-off is thus possible between tuning time and perceptual quality. Hence, in Chapter 4, we propose an HRTF individualization method which consists in tuning the parameters of an HRTF statistical model based on the results of localization experiments.

2.4 HRTF Databases

As we have seen in Section 2.3, many user-friendly HRTF individualization approaches rely on HRTF statistical modeling and thus on databases. In this section, we review the major HRTF databases. While we take a particular interest into the number of subjects available, we review other important characteristics as well, such as the spatial resolution of the measurements and the morphological data included. First, datasets of acoustically measured HRIRs are presented. Then, datasets of numerically simulated ones are reviewed. Finally, these surveys are discussed.

2.4.1 Acoustically Measured

Most HRIR datasets were built thanks to acoustic measurements (see Section 2.3.1 for more details on the technique). In the following, we go over ten of them and their characteristics, such as the number of subjects, their spatial resolution and the type of morphological data included (if present).

In the early 2000s, one of the first freely available HRIR datasets was created by the Center for Image Processing and Integrated Computing (CIPIC)[9] [Algazi01c]. It features HRIRs of 45 human subjects, measured in a regular room whose walls were covered with absorbing materials. The spatial resolution of the measurements is of 5.6° in elevation, and 5° in azimuth for azimuth ranging from −45° to 45° and from 135° to 225° and of 10, 15 or 20° for more lateral positions. The dataset innovated by including 27 anthropometric measurements of the pinnae, head and torso for 43 subjects, measured from pictures. Consequently, since then this dataset has been used in a wide variety of work on HRTF individualization, particularly in the context of morphology-based low-cost personalization processes. Subsequent anthropometric datasets have for the major part followed the lead, using a set of measurements identical or similar to the one proposed in CIPIC.

In the same period, another HRTF database named LISTEN was built at the Institut de Recherche et Coordination Acoustique/Musique (IRCAM), [Warusfel03]. It comprises HRIRs of 51 subjects that were recorded in a fully anechoic room with a lesser spatial resolution of about 15° both in azimuth and elevation.

More recently, i.e. during the last half-decade, a number of datasets of measured HRTF

[9]The CIPIC dataset is available at `https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/`.

Name	Subjects	Spatial sampling					Room	Anthropometry	3-D Meshes		
		N_{dirs}	$\Delta\phi\,(^\circ)$	$\Delta\theta\,(^\circ)$	$\theta_{\text{min}}\,(^\circ)$	$r\,(\text{m})$			Pinnae	Head	Torso
CIPIC [Algazi01c]	45	1226	[5, 20]	5.6	-45	1	Some acoustic treatment	37 subj.	-	-	-
LISTEN [Warusfel03]	51	187	15	15	-45	1.95	Anechoic	Yes	-	-	-
RIEC [Watanabe14]	105	865	5	10	-30	1.5	Anechoic	-	LR[a], 39 subj.	39 subj.	39 subj.
ARI [Majdak10]	201	1550	[1, 7] (GCD[b])	5, 10	-30	1.2	Semi-anechoic	60 subj.	-	-	-
BiLi [Carpentier14]	55	1680	6	$\simeq 6$	-62	1	Anechoic	-	-	-	-
ITA [Bomhardt16b]	46	2233	5	5	-66	1.2	Semi-anechoic	Yes	Yes	-	-
SADIE [Kearney15]	18	170	[7, 180]	[10, 60]	-75	1.5	Anechoic	-	-	-	-
SADIE II [Armstrong18]	18	2818 (2114)	5 (10)	15	-81	1.5	Anechoic, reverberant	-	LR	Yes	-
SYMARE [Jin14]	10	393	$\simeq 10$ (GCD)	10	-45	1	Anechoic	-	Yes	Yes	Yes
HUTUBS [Brinkmann19]	96	440	$\simeq 10$ (GCD)	10	-90	1.47	Anechoic	Yes	Yes	Yes	Yes

Table 2.2 – Public measured HRIR databases. N_{dirs}, $\Delta\phi$, $\Delta\theta$, ϕ_{min} and r denote the number of directions in the spatial sampling grid, the resolution along azimuths, the resolution along elevations, the lowest elevation in the grid, and the grid's radius, respectively.

a: Low-resolution.

b: Great circle distance (GCD) between two neigboring points of identical elevation, as in [Brinkmann19].

sets have been issued. For instance, RIEC, a database with twice as many subjects (105) as CIPIC and ARI, was published in 2014 by the Advanced Acoustic Information Systems Laboratory at Tohoku University, and features HRIRs measured for 865 directions at a distance of 1.5 m with azimuth and elevation resolutions of 5 and 10°, respectively. The dataset includes anthropometric measurements for 39 subjects as well as scans of the head and torso. However, no detailed scans of the pinnae were provided.

The same year, a database of HRIR sets measured in a semi-anechoic room was introduced by the Acoustics Research Institute (ARI) database [Majdak10], featuring a higher spatial resolution than RIEC (azimuth resolution between 2.5° and 5° and elevation resolution of 5°) and a comparable initial number of subjects. It has however been supplied with new subjects ever since, reaching 201 in December 2019, thus making it the largest HRTF database available to this day. CIPIC-like anthropometric measurements of 60 subjects are provided.

In 2014 as well, another high-resolution HRTF database, named BiLi[10] (Binaural Listening) [Carpentier14; Rugeles Ospina15], was released as the result of a collaboration between IRCAM and Orange. It features HRIRs measured for 54 human subjects in an anechoic chamber on a 1680-point Gaussian grid of radius 2.06 m. The Gaussian grid was chosen for its convenience for measurements (practical with a vertical ark of loudspeakers and an azimuth-wise rotating subject) and its adequateness to high-order Ambisonics, i.e. SHD (see Section 2.1.2). Using that setup, the measurement of a complete HRIR set took about 20 minutes, thanks to the use of overlapping exponential sweeps [Majdak07].

Another database of high-resolution HRIR sets, ITA [Bomhardt16b], was published in 2016 by a team from the University of Aachen. HRIRs were measured in a semi-anechoic environment for 2304 points of a 1.2-meter-radius spherical azimuth/elevation equiangular grid whose resolution was 5°. The dataset includes high-resolution 3-D scans of the pinna (obtained by MRI), 4 measurements of the head and 8 CIPIC-like pinna anthropometric measurements made on the scans.

The SADIE dataset[11] [Kearney15] includes HRIR sets of 18 subjects, measured in an anechoic room at the University of York. However, as these measurements were intended for the specific needs of 5$^{\text{th}}$-order Ambisonics, the spatial resolution is quite low, with only 170 directions across the 1.5 m-radius sphere.

More recently, in 2018, a new iteration was issued, the SADIE II dataset[12] [Arm-

[10]The BiLi dataset is available at `http://bili2.ircam.fr`.
[11]The SADIE dataset is available at `https://www.york.ac.uk/sadie-project/database_old.html`.
[12]The SADIE II dataset is available at `https://www.york.ac.uk/sadie-project/`.

Name	Subjects	$f_{\max}$ (kHz)	Method	3-D Meshes			Measured HRTFs	Public
				Pinnae	Head	Torso		
SCUT [Rui13]	56	16	FM-BEM	Yes	Yes	Yes	-	-
SYMARE [Jin14]	61	16	FM-BEM	Yes	Yes	Yes	Yes	10/61
Yamaha [Kaneko16a]	30	24	FM-BEM	Yes	Yes	Yes	-	-
HUTUBS [Brinkmann19]	96	22	FM-BEM	Yes	Yes	-	Yes	Yes
FAST Sec. 3.1	119	16	FM-BEM	Yes, registered	-	-	-	-
WiDESPREaD Sec. 3.4	1005	16	FM-BEM	Yes, registered	-	-	-	Yes

Table 2.3 – Public and private numerically simulated HRTF databases. $f_{\max}$ denotes the maximum frequency in computations.

strong18], featuring HRIR sets measured in an anechoic environment for 18 human subjects with a much higher spatial resolution than SADIE I. There are two types of measurement grids depending on the subject, in order to adapt measurement time. Both grids are of the spherical kind with an elevation resolution of 15°. They differ in their azimuth resolutions as the finest one's is of 5° for a total number of directions of 2818 while the coarser one's is of 10° for a total number of directions of 2114. 7 subjects subjects out of 18 were measured in 1.25 h using the first grid while the remaining 11 were measured in 1 h with the other. A few additional measurement points are included to the spherical grids in order to allow the perfect reproduction of 11 types of spherical harmonics configurations.

2.4.2 Numerically Simulated

There also exists synthetic datasets, built by numerically simulating HRTF sets from 3-D scans of listener morphology.

The SYMARE (Sydney York Morphological and Acoustic Recordings of Ears) database is such a dataset and was issued as part of a collaborations between the Universities of Sydney and York in 2014.

The dataset features HRTFs simulated by Fast-Multipole Boundary Element Method

(FM-BEM) for 61 subjects as well as 3-D scans of the pinnae, head and torso used for the simulations. Simulations were performed for frequencies up to 20 kHz when using the head and pinnae, and up to 16 kHz when including the torso. Spatial resolution is not an issue here. Indeed, as the reciprocity principle [Zotkin06] is easily applicable to simulations, virtually any measurement grid can be chosen with only a marginal increase in computing cost. The dataset also includes HRIRs measured in an anechoic chamber with a low spatial resolution, on a grid of radius 1.2 m and average azimuth and elevation resolutions of 10° in both cases. Only a sample of 10 subjects is freely available. In spite of this, this dataset has been the reference of databases gathering both measured and simulated HRTFs, up until very recently.

A team from the Technical University of Berlin, Huawei Technologies and Sennheiser Electronic has issued the HUTUBS database[13] [Brinkmann19] in 2019. It features both measured and simulated HRTFs for 96 subjects, as well as pinnae and head 3-D meshes and anthropometric measurements. The spatial resolution of the measurements is not particularly high, with 440 directions on a 1.47 m-radius grid with average azimuth and elevation resolutions of 10°. The choice of measurement grid accounted for compatibility with SHD up to the 17^{th} order. Acoustical simulations were performed on shoulder-less heads for frequencies up to 22 kHz.

There exists other databases of simulated HRTFs that are not accessible to the public. For example, an article published by a team from the South China University of Technology (SCUT) in 2013 [Rui13] presents the simulation by FM-BEM of the HRTF sets of 56 human subjects including near-field HRTFs, with distances ranging from 10 cm to 1.2 m.

Another example is the dataset mentioned by a research team from Yamaha in a 2016 article [Kaneko16a], which features the HRTF sets of 30 subjects, simulated by FM-BEM based on a combination of high-resolution pinnae 3-D scans and rougher head-and-torso ones.

Discussion

Over the past twenty years, a number of datasets have been built by measuring the HRIRs of various human subjects, particularly in the last half-decade, period during which eight out of the ten datasets mentioned above were issued.

While some of them (SADIE, LISTEN, SYMARE, HUTUBS) have a rather low spa-

[13]The HUTUBS database is available at `https://depositonce.tu-berlin.de/handle/11303/9429`.

tial resolution (compared to the localization blur presented in Section 1.1.5), others can be considered as having a high spatial resolution (SADIE II, ITA, BiLi). Sets of measurements of the human morphology are included in several of them (CIPIC, LISTEN, ARI, ITA and HUTUBS), a trend that was initiated by CIPIC and kick-started the active field of user-friendly HRTF individualization based on anthropometry (see Section 2.3.3). Sometimes, morphological information is included in the form of 3-D scans of the head, torso and/or pinna as in RIEC, ITA, SADIE II, SYMARE and HUTUBS. However RIEC and SADIE II do not include detailed scans of the pinnae and, while ITA does, it does not feature head or torso meshes.

Independently of their quality, due to the heavy apparatus and time that are required to make acoustic measurements, these databases are rather limited in terms of number of subjects. Indeed, the largest one, ARI, features 201 which is twice more than its two closest competitors in this area, RIEC and HUTUBS, who feature 105 and 96, respectively. Most of the other datasets mentioned above comprise data for about 50 listeners (CIPIC, LISTEN, BiLi, ITA) while both SADIE sets feature 18 and the public section of SYMARE features only 10.

When studying the inter-individual variations of HRTF sets, this may be problematic as the order of magnitude of the dimensionality of a high-resolution HRTF set [Bomhardt16b] is half a million (129 frequencies $\times$ 2300 directions $\times$ 2 ears $\simeq 6 \cdot 10^5$).

One could imagine turning to numerical simulations to create larger datasets of synthetic HRTFs. While a few such datasets exist (SYMARE, HUTUBS, Yamaha and SCUT), they are mostly private, HUTUBS being the only fully public one. Moreover, none of them features more subjects than measured HRTF databases. Indeed, the largest one, HUTUBS, includes simulated HRTFs for 96 subjects while ARI, RIEC and HUTUBS include measured HRTFs for 201, 105 and 96 subjects, respectively. The fact that synthetic datasets do not present more subjects than acoustical ones can be explained by the fact that they still rely on the acquisition and edition of 3-D morphology of the subjects, which is largely manual and time-consuming, and that simulations requires non-negligible computing resources. An additional problem is the uncertainty of the perceptual relevance of simulated HRTFs, making it possibly unworthy of the effort of building a large database.

Chapter 3

Dimensionality Reduction and Data Augmentation of Head-Related Transfer Functions

As we have seen in Section 2.3 of Chapter 2, an interesting approach to the matter of user-friendly HRTF individualization consists in tuning the parameters of a statistical model of HRTFs, either based on anthropometry or on perceptual feedback from the listener – the latter being further explored in Chapter 4. Seeing that HRTF sets are a data with hundreds of thousands of degrees of freedom (see Section 3.2 and Table 3.1), it is important in that context to reduce the dimensionality of the problem. Indeed, in the case of a perceptual feedback-based approach, for instance, a lower number of tuning parameters allows for a more efficient exploration of the inter-individual variations of HRTFs and thus a shorter and more comfortable tuning session for the listener.

However, currently available datasets are small compared to the dimensionality of the data: the largest one, ARI [Majdak10], includes data for 201 subjects (see Chapter 2, Section 2.4 for a review of HRTF databases). Furthermore, while work has been done towards combining existing databases [Andreopoulou11; Tsui18; Spagnol20], such composite databases can hardly attain the same level of homogeneity as a database made in a single campaign.

In this chapter, we investigate the matter of the dimensionality reduction of magnitude HRTF sets. To this end, we used principal component analysis (PCA). Choosing PCA over more complex machine learning techniques, was motivated by the fact that we performed the statistical modeling in a way that focuses on the inter-individual variations of HRTF sets, which has barely been addressed in the literature (see Chapter 2, Section 2.1.3 for more details).

Thus, we investigate in Section 3.2 the capacity of this inter-individual approach to PCA to reduce the dimensionality of magnitude HRTF sets for 9 different datasets. These

9 datasets include 8 public datasets and a proprietary dataset of 119 3-D ear meshes and matching simulated PRTF sets, named FAST, which we present in Section 3.1. In Section 3.3, we compare the dimensionality reduction performance of PCA on FAST magnitude PRTF sets to that of matching ear point clouds. Based on the results of this study, and in order to alleviate the aforementioned lack of large-scale datasets, we present in Section 3.4 a data augmentation method that relies on random generations of pinna meshes and on numerical simulations of corresponding PRTF sets. The resulting 1005-example dataset, named WiDESPREaD (Wide Dataset of Ear Shapes and Pinna-related transfer functions generated by Random Ear Drawings) was made public and available online[1]. Finally, in Section 3.5 we study the impact on dimensionality reduction performance of training PCA with this augmented PRTF dataset.

3.1 The FAST Dataset: 119 Ear Meshes and Matching Simulated Pinna-Related Transfer Functions

Most work presented in this chapter is based on a proprietary dataset of $n = 119$ 3-D scans of human left pinnae and matching 119 numerically simulated PRTF sets. We hereon refer to it as the FAST dataset, after our research team.

In this section, we present the constitution of this dataset. First, we introduce a basis dataset of 123 registered left ear meshes which was constituted in previous work by Ghorbal *et al.* [Ghorbal19]. Then, we go over corrections that were applied to that first dataset, including the removal of 4 problematic subjects. Finally, we describe in detail how we complemented the $n = 119$ pinna meshes with matching PRTF sets by means of boundary element method (BEM) simulations.

3.1.1 Ear Meshes

Acquisition & registration

For the major part, the dataset of ear meshes was constituted in previous work by Ghorbal *et al.* [Ghorbal19]. First, 3-D scans of the left pinna of 123 human subjects were acquired using a commercial structured-light based scanner, eFit by United Sciences. The acquisition of one pinna took about 20 min. Then, the meshes were rigidly aligned by means

[1] `https://sofacoustics.org/data/database/widespread/`

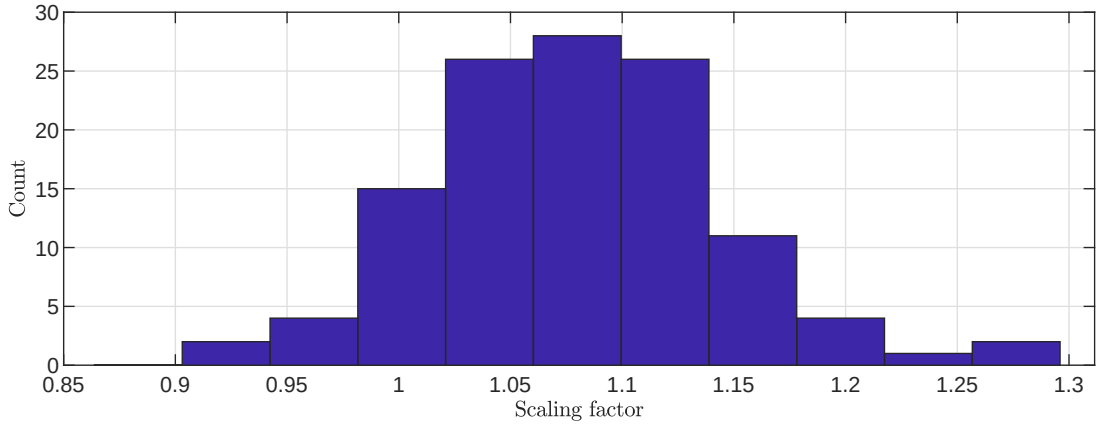


Figure 3.1 – Histogram of the scaling factors applied to the 119 ear shapes to normalize them in size.

of the Procrustes method [Gower75]. Finally, the point clouds were registered: after this step, every point cloud had the same number $n_{v_0} = 18887$ of vertices, and the vertex indexing was semantically coherent from one subject to the other. The main goal of registering the point clouds is to be able to study the variations in shape and to build a statistical shape model (SSM) (see Section 3.3.1).

The registration process was semi-automatic: a human operator identified manually a number of characteristic features on the mesh, then an algorithm derived a denser sampling of the pinna surface, designed so that the newly created point clouds were sampled in coherence with each other. A new set of triangular faces was defined from the n_{v_0} vertex indices.

Normalization in size

In addition, the pinna meshes were normalized in size and the scaling factors stored. Indeed, this practice, common when it comes to SSMs [Cootes95], forces the statistical model to learn complex variations in the shape of the pinnae at the exclusion of size.

A histogram of the scaling factors associated with the 119[2] subjects of the final FAST dataset are shown in Figure 3.1. The scaling factors are normally distributed with a significance level of 1 % according to the Anderson-Darling test, with a mean value of 1.080 and a standard deviation of 0.065.

In the following, we note $E = \{\mathbf{e}_1, \ldots \mathbf{e}_{n_0}\}$ the set of $n_0 = 123$ ear point clouds whose x, y and z coordinates are concatenated into row vectors $\mathbf{e}_1, \ldots \mathbf{e}_{n_0} \in \mathbb{R}^{3n_{v_0}}$, with

[2]Four of the initial 123 meshes were excluded during the registration fix step described just below.

$3n_{v_0} = 56661$. Thanks to the registration, the only change from one mesh to the other resides in the coordinates of the n_{v_0} vertices. Therefore, the term "ear shape" is hereon used interchangeably with "ear point cloud".

Registration fix & ear canal removal

In the initial dataset, there was a critical issue of registration in the meshes, localized in the ear canal area. The registration was sometimes so wrong that a vertex located at the tip of the ear canal for certain subjects was found in the concha for others (see Figure 3.2). As a consequence, a SSM trained on this dataset would learn unrealistic deformations of exaggerated amplitude.

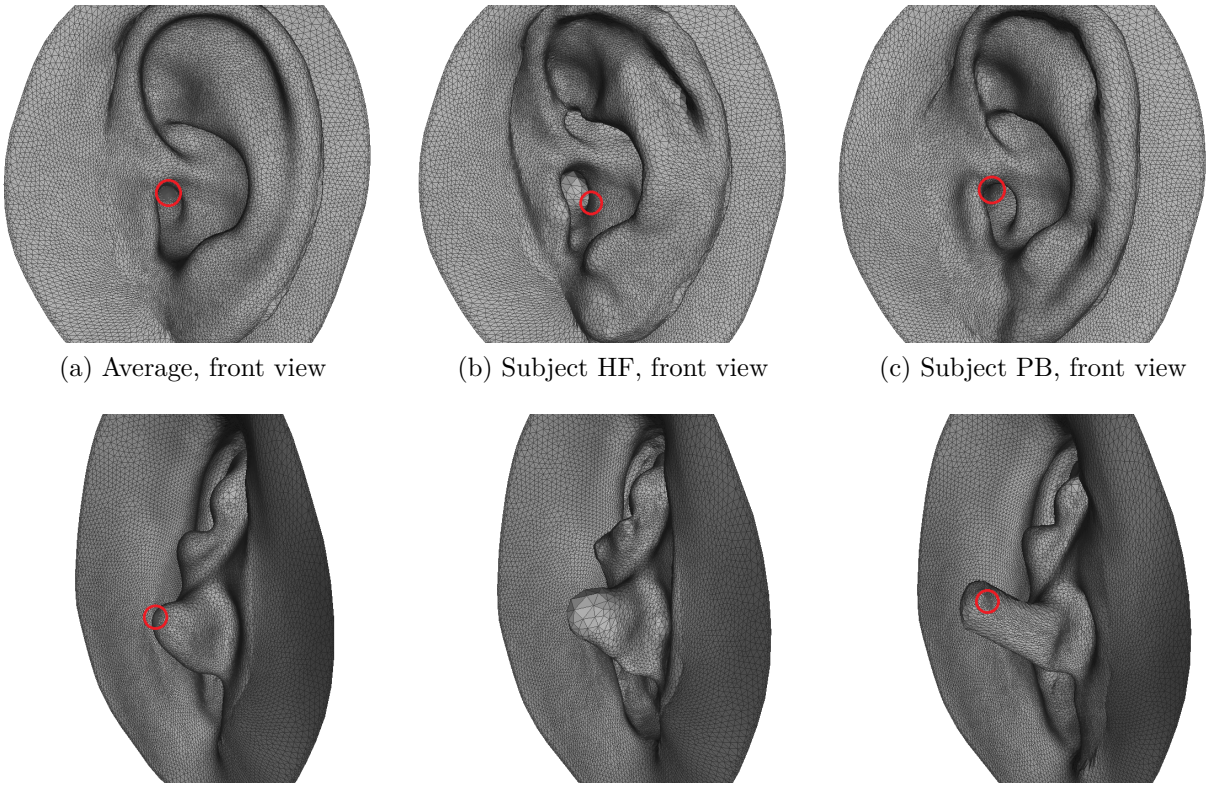
Moreover, most of this area is constituted of artificial data. Indeed, the scanning device could not acquire the ear canal down to the ear drum and closed the hole automatically. Thus, we also wanted to erase this non-realistic part of the morphology before training the SSM.

The straightforward and ideal solution to the registration issue would have been to perform the registration of the 123 ear scans all over again. However, as mentioned above, this step relies on manual annotation, which is tedious and lengthy: two to three weeks of full-time work would have been required to process the whole dataset. Furthermore, the defect is localized, moreover in an area where we would like to remove most of the vertices. Hence, we devised a automated method to correct this defect in all 123 meshes while respecting a major constraint: preserving registration.

Anchoring of the problematic vertices As a first step, we constrained the "displacement" of the vertices in the ear canal neighborhood from one subject to the other. To do so, we anchored these vertices to the average point cloud $\bar{\mathbf{e}}$, whose registration we deemed acceptable (see Figure 3.2), with

$$\bar{\mathbf{e}} = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{e}_i. \tag{3.1}$$

For each mesh in the dataset and for every one of these vertices, we applied a linear weighting that made the vertex closer to its match in the average point cloud. The weights increased progressively from the edge of the ear canal to its end, so that the ear shape progressively transitioned from the initial point cloud to the average (see Figure 3.3)
.



(a) Average, front view (b) Subject HF, front view (c) Subject PB, front view

(d) Average, rear view (e) Subject HF, rear view (f) Subject PB, rear view

Figure 3.2 – Illustration of the registration issue, and the variability of the registration in the ear canal area. Vertices expected to be at the tip of the ear canal are circled in red for three exemplary meshes. The average shape (a, d) illustrates the expected behavior, with the circled vertices well located at the end of the ear canal. Subject HF (b, e), in contrast, constitutes an extreme example of the issue: the circled vertices are not even located in the ear canal, but are in the concha. For subject PB (c, f) the registration issue is present but milder, with the circled vertices slightly on the side of the ear canal.

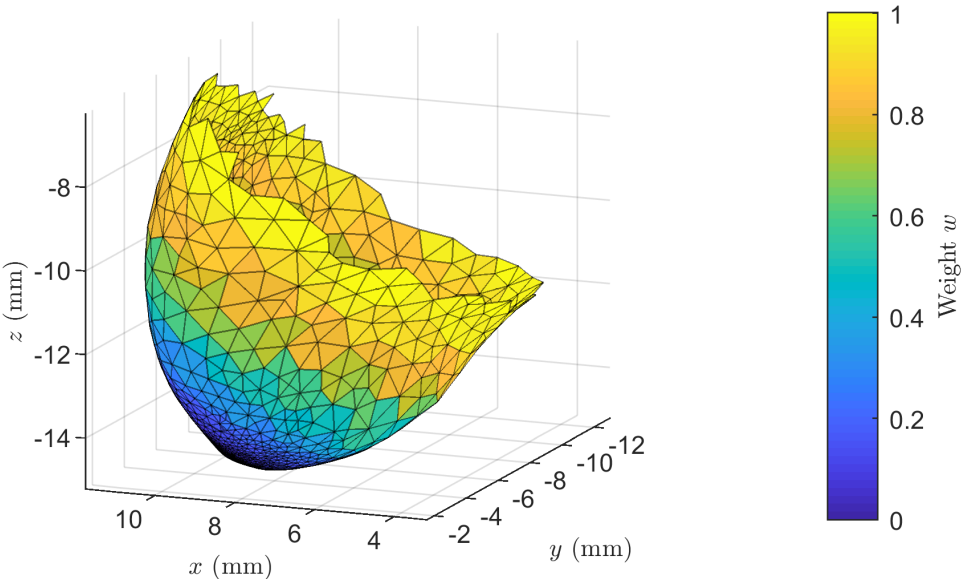


Figure 3.3 – Weights $w(i)$ for all $\bar{\mathbf{e}}(i)$ in $\mathbf{\Omega}_\mathrm{c}$, displayed on the sub-mesh that corresponds to $\mathbf{\Omega}_\mathrm{c}$

Let $\mathbf{e}$ be an ear point cloud of E. For all $k = 1, \ldots n_{v_0}$, we denote $\mathbf{e}(k) \in \mathbb{R}^3$ its k^{th} vertex. The first step was to manually elect, on the average shape, the ear canal tip vertex $\bar{\mathbf{e}}(k_{\text{tip}}) \in \mathbf{R}^3$ and a unitary vector $\mathbf{u}_\mathrm{c} \in \mathbf{R}^3$ that defined the canal axis, passing through $\bar{\mathbf{e}}(k_{\text{tip}})$. The ear canal neighborhood that we wished to constraint $\mathbf{\Omega}_\mathrm{c}$ was then defined as such:

$$\mathbf{\Omega}_\mathrm{c} = \{k \in [\![1, n_{v_0}]\!] \mid \| \bar{\mathbf{e}}(k) \cdot \mathbf{u}_\mathrm{c} \| < D\} \cap \{k \in [\![1, n_{v_0}]\!] \mid \| \bar{\mathbf{e}}(k) - \bar{\mathbf{e}}(k_{\text{tip}}) \| < D'\} \quad (3.2)$$

where D is a canal-axis distance parameter and D' an Euclidean distance parameter, manually tuned to 6.8 mm and 10.9 mm, respectively.

The weighting can then be written as:

$$\forall k \in [\![1, n_{v_0}]\!], \quad \mathbf{e}(k) := \alpha(k)\mathbf{e}(k) + (1 - \alpha(k))\, \bar{\mathbf{e}}(k) \quad (3.3)$$

where the weights are defined on the average shape:

$$\alpha(k) = \begin{cases} \sin^2\left(\dfrac{\bar{\mathbf{e}}(k) \cdot \mathbf{u}_\mathrm{c}}{D} \dfrac{\pi}{2}\right) & \text{if } k \in \mathbf{\Omega}_\mathrm{c}, \\ 1 & \text{otherwise.} \end{cases} \quad (3.4)$$

However, for four of the meshes, the registration issue was beyond correction by means of the aforementioned method. They thus were excluded from the final dataset, which includes $n = 119$ pinna meshes of $n_v = 18176$ vertices and 35750 triangular faces.

Deletion of the end of the canal Secondly, we removed the end of the canal i.e. the vertices designed by their indices:

$$\mathbf{\Omega}_{\mathrm{t}} = \mathbf{\Omega}_{\mathrm{c}} \cap \{j \in [\![1,\, n_{v_0}]\!] \mid \bar{\mathbf{e}}(j) \cdot \mathbf{u}_{\mathrm{c}} < D_e\}\,, \tag{3.5}$$

where D_e is manually tuned to 2 mm.

3.1.2 PRTFs: Numerical Simulations

For all ear shape $\mathbf{e}_i$ in E, we numerically simulated the corresponding PRTF set $\mathbf{h}_i \in \mathbb{C}^{n_f \times n_d}$, where n_f and n_d denote respectively the number of frequency bins and the number of directions of measurements. Simulations were carried out using the fast-multipole boundary element method (FM-BEM) [Gumerov05], by means of the *Mesh2HRTF*[3] software developed by the ARI team [Ziegelwanger15a; Ziegelwanger15b].

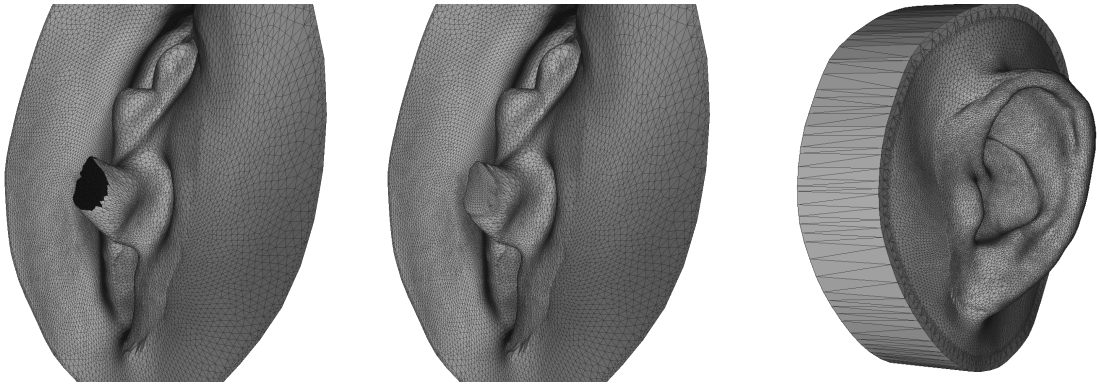
We denote $\psi : \mathbb{R}^{3n_v} \mapsto \mathbb{C}^{n_f \times n_d}$ the process of going from a registered n_v-vertex ear point cloud to the corresponding simulated PRTF set, which is described in the rest of the subsection.

Simulations were made for $n_f = 160$ frequencies from 0.1 to 16 kHz, regularly spaced with a step of 100 Hz. Let us denote $F = \{k \cdot (100\text{ Hz}) \mid k = 1,\, \ldots\, 160\}$ this set of frequency bins. The frequency resolution was chosen so that it was finer than the equivalent rectangular bandwidth (ERB)-based frequency scale in most of the frequency range. Indeed, the ERB scale is appropriate for HRTFs according to [Breebaart01] and the 100-Hz-spaced linear scale is finer than the ERB scale for frequencies above 700 Hz, which is more than sufficient in the case of PRTFs, who include little spectral variations below 4-5 kHz.

Mesh closing and grading

First, we derived the ear mesh from the ear point cloud by incorporating the 35750 triangular faces defined by the indices of the n_v vertices, as explained in Section 3.1.1.

[3] `https://sourceforge.net/projects/mesh2hrtf/`



(a) Initial mesh (b) Closed ear canal (c) Closed mesh

Figure 3.4 – Pinna mesh of subject PB throughout the closing process. (a) Initial mesh from the FAST dataset. (b) Closed ear canal. (c) Final closed mesh, after merge with the cylindric basis.

Second, we closed the ear mesh by filling the canal hole based on our prior knowledge of the boundary's vertex indices, and then by stitching the resulting mesh onto a cylindrical base mesh. Using such a small base mesh instead of one of a head and torso has consequences: spectral features that are usually found in HRTFs are altered (head shadowing effect is reduced to a smaller angular zone and shifted to higher frequencies) or absent (ripples due to the torso). However, as we did not have at our disposal a dataset of individual 3-D head and torso scans, in the latter case we would only have been able to use a generic head and torso mesh, which would have mixed non-individual spectral features with the individual pinna-related ones, at the cost of a great increase in required computing resources. These steps were scripted in Blender[4] Python and performed automatically using various Blender built-in mesh treatments.

Third, a re-sampling (also called grading) of the mesh was performed. This step is a pre-requirement to any boundary element simulation: the mesh ought to be as regular as possible and sampled finely enough with regard to the maximum simulated frequency. A widely used rule of thumb is for the mesh to present a uniform vertex distribution, equilateral triangles and at least six elements per wavelength. This rule is discussed in detail by Marburg *et al.* in [Marburg02]. In our case, we used the progressive grading approach proposed by Ziegelwanger, Kreuzer and Majdak in [Ziegelwanger16] and made available on-line as an OpenFlipper[5] [Möbius10] plug-in, which makes the mesh finer near

[4] `https://www.blender.org/`
[5] `http://www.openflipper.org/`

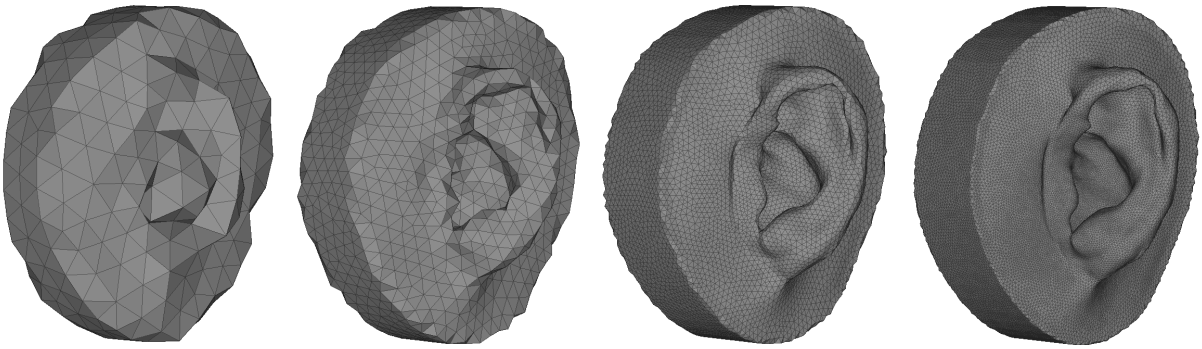


Figure 3.5 – Simulation-ready meshes derived from ear point cloud $\mathbf{e}_1$ for four mesh grading configurations, each corresponding to a frequency band. Left to right: [0.1, 0.4 kHz], [0.5, 2.0 kHz], [2.1, 3.5 kHz] and [3.6, 16 kHz].

the ear canal (where the sound source is positioned) and progressively coarser elsewhere. This considerably decreases the computing cost of the FM-BEM simulation compared to uniform re-sampling, while maintaining numerical accuracy. In this case, we used the cosine-based approach with the grading factor set to 10.

Additionally, in order to further reduce the computational cost, we adapted the mesh grading step to each of four different frequency bands. At low frequencies, a uniform re-sampling was enough due to the low number of required elements. It was performed with target edge lengths of 10 and 5 mm, in the frequency bands [0.1, 0.4 kHz] and [0.5, 2.0 kHz], respectively. At higher frequencies, the re-sampling was progressive, with target minimum and maximum edge lengths of 2 and 5 mm, and 0.7 and 5 mm, in the frequency bands [2.1, 3.5 kHz] and [3.6, 16 kHz], respectively. An example of simulation-ready meshes (each corresponding to a mesh grading configuration) is displayed in Figure 3.5.

Simulation settings

Reciprocity principle According to the reciprocity principle in acoustics, given two points in space A and B, the pressure in B due a sound source located in A, $p_{\mathrm{A}\to\mathrm{B}}$ is equal to the pressure that would be observed in A if the sound source was located in B $p_{\mathrm{B}\to\mathrm{A}}$:

$$p_{\mathrm{A}\to\mathrm{B}} = p_{\mathrm{B}\to\mathrm{A}}. \tag{3.6}$$

This is particularly interesting in the case of HRIR measurements. Instead of sequentially measuring the responses in the ear canal to sound sources located in n_d locations,

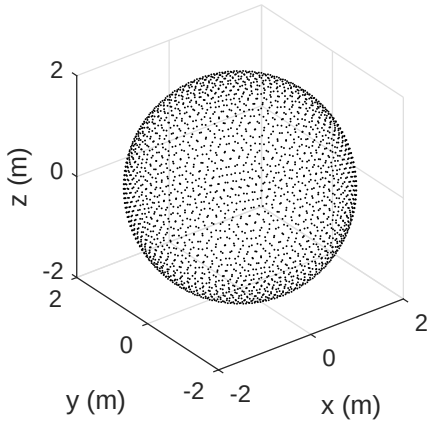


Figure 3.6 – Spherical grid used for PRTF simulations: 2-meter-radius icosahedral geodesic polyhedron of frequency 256 ($n_d = 2562$ vertices).

a single sound source can be placed inside the ear canal while the pressure is measured at the n_d points of space at once [Zotkin06]. This approach collides however with several problems related to the position of a loudspeaker near a person's ear drum: limited sound level of the impulse, small size and directivity of the loudspeaker [Matsunaga10].

In a context of numerical simulation, on the contrary, none of these problems are encountered. Hence, the reciprocity principle can be employed in order to considerably reduce the computing cost of simulating an HRTF set and to make measurements on an arbitrarily dense grid – a widespread practice [Katz01; Kreuzer09; Jin14].

Measurement grid In practice, a few triangular faces located on the ear canal plug were assigned a vibrant boundary condition (making them the sound source), while virtual microphones were disposed around the pinna mesh. The spherical measurement grid, centered on the pinna, was a 2-meter-radius icosahedral geodesic polyhedron of frequency 256 ($n_d = 2562$ directions), displayed in Figure 3.6. Let $\mathcal{D}$ be this measurement grid.

Not studied in the rest of this thesis but included in the WiDESPREaD dataset (see Section 3.4), PRTF sets were calculated on additional measurement grids: another icosahedral geodesic polyhedron of radius 1 m, and equiangular polar grids with an angular resolution of 5° ($n_d = 2522$ directions) of respective radii 2 m, 1 m, 0.5 m and 0.2 m.

Boundary conditions Except for the few vibrating triangles mentioned above (see Figure 3.7), the boundary condition was set to fully reflective (infinite impedance) everywhere on the mesh. This choice was mostly due to a technical constraint: the release of

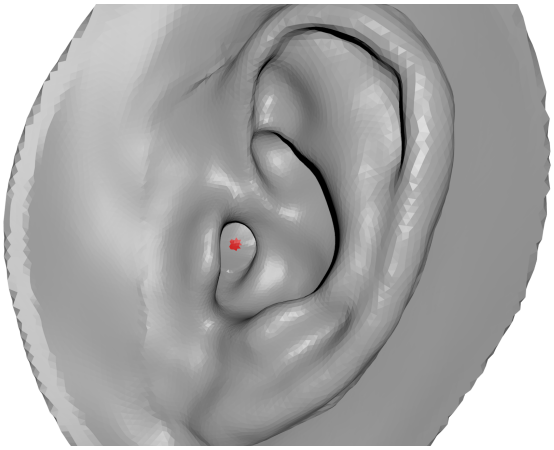


Figure 3.7 – Boundary conditions applied to subject PB's pinna mesh, graded for frequencies up to 16 kHz. Red: sound source (vibrant boundary condition). Gray: infinitely reflective.

Mesh2HRTF that we used (v0.1.3, released in June 2018) did not handle other boundary conditions properly. This is nonetheless in agreement with the literature. Indeed, Katz [Katz00], by means of impedance tube measurements, finds that the absorption coefficient of the human skin (measured at different positions of the body) is close to that of a fully rigid material for frequencies between 1 and 6 kHz. To the best of our knowledge, there is no measurement of the impedance of the human skin at higher frequencies, likely because of the limited frequency range of impedance tube measurements: up to 6 kHz for standard devices, although a recent experimental device proposed by Kimura *et al.* [Kimura14] appears to allow measurements up to 13 kHz.

Regarding the cylindric basis mesh, it might have been desirable to make it fully absorbing in order to remove its contribution to the PRTFs. In Figure 3.11 (in which the horizontal-plane PRIRs of an exemplary subject are plotted), this contribution can be observed in the form of a multiple wavefront for azimuths between $-180°$ and $-30°$, which corresponds to the propagation of sound around the basis mesh. In any case, this phenomenon is limited to the contralateral hemisphere, where PRTF data has little meaning in the absence of the head.

Post-processing

The output of the FM-BEM calculations is the Fourier transform of the pressure field $P(f, \theta, \varphi)$ at each point (θ, φ) of the measurement grid $\mathcal{D}$ and at each simulated frequency $f \in F$. As $\mathcal{D}$ is spherical, the radius is not considered here.

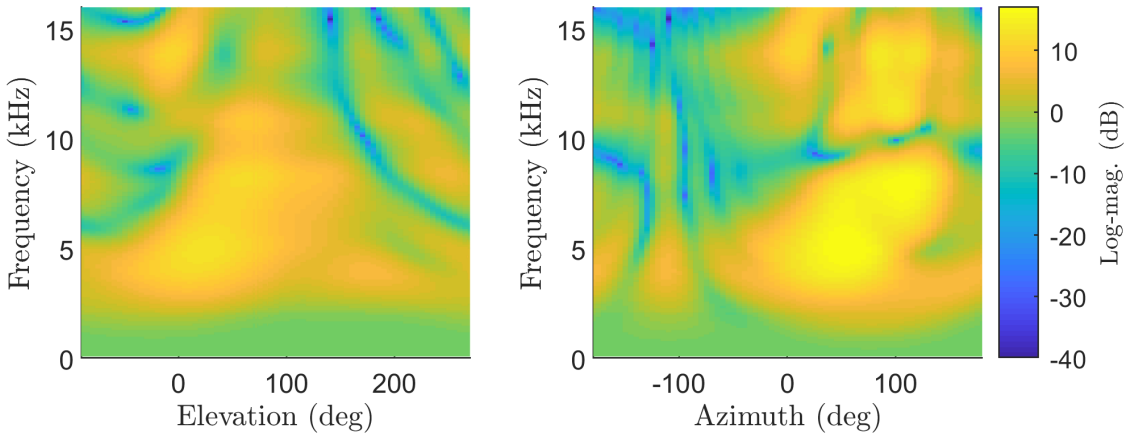


Figure 3.8 – Numerically simulated PRTF set of subject PB just after derivation from the pressure field (see Equation (3.7)). The corresponding mesh and location of the vibrating sound source are displayed in Figure 3.7. Log-magnitude PRTFs are plotted in the median (left) and the horizontal (right) planes.

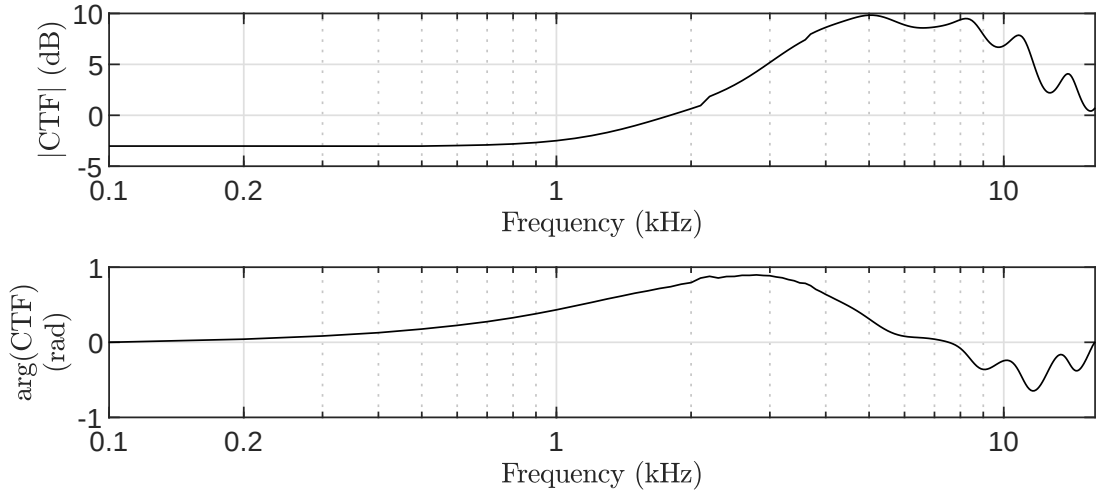


Figure 3.9 – CTF computed from the numerically simulated PRTF set of subject PB, used for its DFEQ.

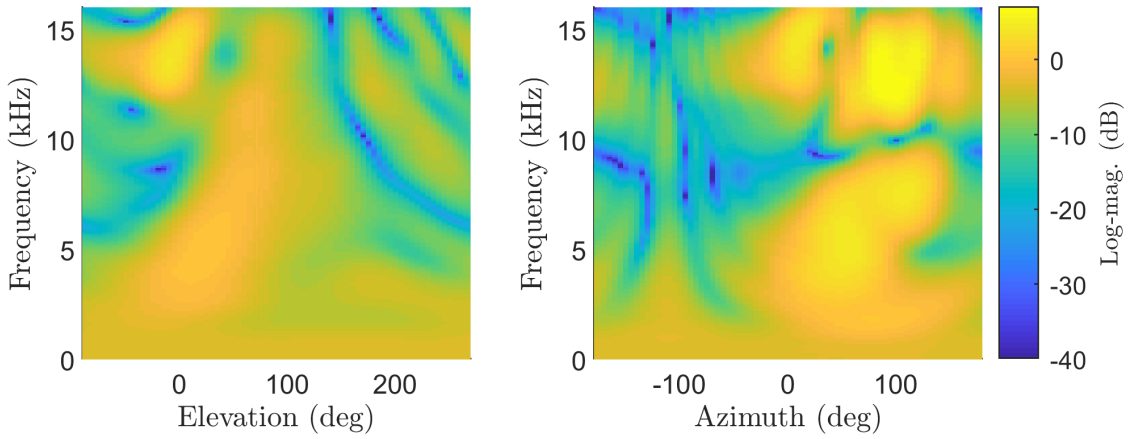


Figure 3.10 – Numerically simulated PRTF set of subject PB after post-processing (generation of a constant component and DFEQ). Log-magnitude PRTFs are plotted in the median (left) and the horizontal (right) planes.

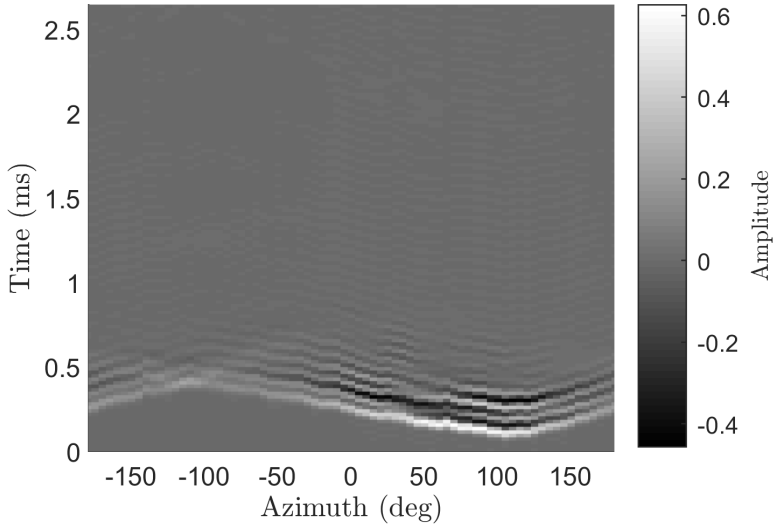


Figure 3.11 – PRIR set of subject PB after post-processing: first 2.6 ms (128 samples) of the horizontal-plane PRIRs.

According to the reciprocity principle, the pressure $P(f, \theta, \varphi)$ is identical to the pressure that would be observed if the sound source was in (θ, φ) and the microphone in the ear canal.

Derivation from the pressure field First, according to Equation (1.2), PRTFs H were directly derived from the pressure field:

$$H(f, \theta, \varphi) = \frac{P(f, \theta, \varphi)}{P_{\text{ref}}(f)}, \tag{3.7}$$

for all $(\theta, \varphi) \in \mathcal{D}$ and $f \in F$, where $P_{\text{ref}}(f)$ is the reference pressure i.e. the pressure that would be observed in the origin if the pinna was absent.

Constant component Second, a constant component was added: the PRTFs were padded in frequency zero using the 100-Hz complex values: for all $(\theta, \varphi) \in \mathcal{D}$,

$$H(f = 0, \theta, \varphi) := H(f = 100 \text{ Hz}, \theta, \varphi). \tag{3.8}$$

Diffuse-field equalization Third, a diffuse field equalization (DFEQ) of the PRTF set was performed (see Chapter 1, Section 1.2.3 for further detail on DFEQ).

For all frequency bins $f \in F$ and for all directions $(\theta, \varphi) \in \mathcal{D}$,

$$H(f, \theta, \varphi) := \frac{H(f, \theta, \varphi)}{c(f)}, \tag{3.9}$$

where $c(f) \in \mathbb{C}$ denotes the CTF. The magnitude of the CTF was obtained by computing the Voronoi-diagram-based weighted average of the log-magnitude spectra of H over all directions of $\mathcal{D}$, then by deriving the corresponding minimal phase spectrum (see Section 1.2.3).

As can be seen in Figure 3.12, the magnitude spectrum of the CTF change substantially from one pinna to the other. In particular, the central frequencies of the various peaks and notches – omni-directional resonances – are variable. This highlights the interest of performing a DFEQ, even on synthetic PRTF sets of the same dataset.

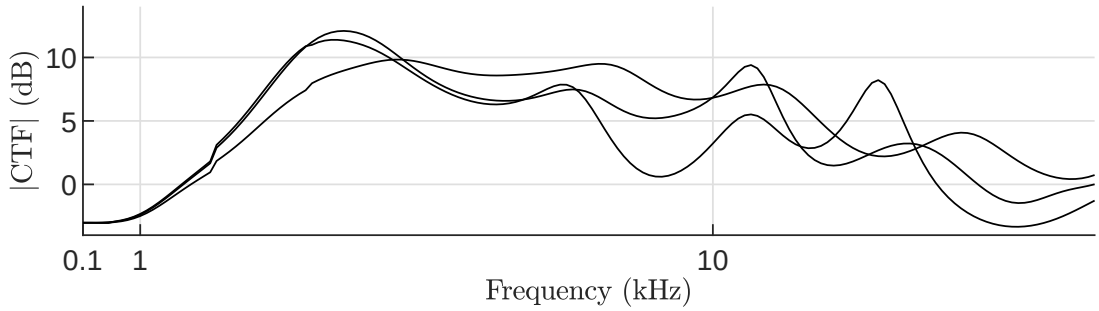


Figure 3.12 – Magnitude CTFs of three exemplary PRTF sets of the FAST dataset (prior to DFEQ), each computed from a different pinna.

Conclusion

In this section, we have presented the FAST dataset, a dataset of 119 3-D pinna meshes and matching PRTF sets computed by means of BEM simulations.

This dataset has the advantage of including both auditory and morphological data. Furthermore, the pinna meshes are registered, which makes them particularly suited for various applications such as statistical analysis, easy extraction of anthropometric measurements and/or regression between morphology and transfer functions. Finally, it includes data for more subjects than most public HRTF datasets, being surpassed only by ARI's 201 (see Chapter 2, Section 2.4).

However, the dataset does not include head and torso 3-D morphologies nor their auditory filtering contributions to HRTFs. Yet, the pinnae have a vast influence on the spectral features involved in the perceptual problems that arise from a lack of individualization [Asano90]. Furthermore, they are arguably the most complex component of HRTF-impacting morphology (i.e. pinnae, head and torso) in terms of shape, inter-individual variability and influence of physical changes on auditory filtering.

Finally, let us note that the pinnae were normalized in size and that the PRTF sets were derived from them. However, as we stored the scaling factors, the ear meshes can easily be re-scaled. Additionally, a close approximation of the corresponding PRTFs sets can be obtained by applying matching frequency scalings to the PRTF sets.

3.2 Dimensionality Reduction of HRTFs

In this section, we investigate how PCA performs at reducing the dimensionality of magnitude HRTF sets from 9 different datasets, including FAST. We start by explaining

how we performed PCA to learn inter-subject variations and reduce their dimensionality. Then, we compare the various datasets under the light of how PCA performed at reducing dimensionality. In particular, we compare some of these results to the literature.

3.2.1 Principal Component Analysis of Log-Magnitude HRTFs

We focus hereon on the magnitude spectra of HRTFs, leaving the matter of ITD individualization out of the scope of this work. Indeed, lateral perception is more robust to a lack of individualization [Wenzel93]. Furthermore, a set of ITDs is a data of lower dimensionality, as it corresponds to one value per direction, against at least a few dozens for the magnitude spectra. For a state-of-the-art on approximating and modeling ITDs, the curious reader can refer to Bomhardt's PhD thesis [Bomhardt17].

In matters of dimensionality reduction, PCA is usually an indispensable first step. Indeed, it is a statistical analysis tool that can help better understand the dataset before moving on to more complex approaches. Furthermore, it is a simple, low-complexity technique that has proved its usefulness in a wide variety of dimensionality reduction problems. Its main limitation lies in its inability to describe non-linear manifolds.

Looking into the literature, PCA is effectively the most popular machine learning approach to model HRTFs. Yet, let us mention that other techniques have been used as well, such as independent component analysis (ICA) [Larcher00; Huang09b; Liu19b], High-Order SVD [Li13] for linear techniques, and Isomap [Kapralos08; Grijalva16] and locally linear embedding (LLE) [Duraiswami05; Kapralos08] for non-linear ones. Neural networks have only come up very recently for unsupervised HRTF modeling [Yamamoto17; Chen20]. However, these approaches rarely learn inter-subject variations only, often mixing in directional variations.

In Section 2.1.3, we discussed the various ways in which HRTF data can be formatted prior to PCA. Regarding our HRTF individualization problem, the *inter-individual* one seems most adequate as, in that case, PCA only learns variations between subjects. However, it is worth noting that it has rarely been used in the literature [Hölzl14; Hold17], likely because of the limited size of currently available datasets (≤ 201) compared to the dimensionality of the data (order of magnitude between 10^4 and 10^6, see Table 3.1).

In the following, we detail how we used PCA to reduce the dimensionality of magnitude HRTF sets.

Pre-processing

Prior to PCA, all HRTF sets went through a small pre-processing step. First, we re-sampled the HRIRs to a sampling frequency of 32 kHz, that is a maximum frequency of 16 kHz. Indeed, most listeners cannot hear content at higher frequencies. Second, we performed a diffuse-field equalization of the HRTF sets. The magnitude spectrum of the CTF was computed by performing a Voronoi-diagram-based average of the log-magnitude HRTF sets (see Section 1.2.3).

The spatial grids of 3 HRTF sets from the ARI database differed slightly from that of the other 198 HRTF sets. Rather than tampering with the data by interpolating the HRTFs, we cast aside these 3 HRTF sets.

Data formatting

Let us consider a dataset of n DTF sets measured or simulated on a spherical or hemispherical grid $\mathcal{D}$ of n_d directions and on a frequency range F of n_f bins

$$\left\{H_i^{(\lambda)}(f,\theta,\varphi) \mid i = 1,\ \ldots\ n,\ \lambda \in \{\mathrm{L},\mathrm{R}\},\ f \in F,\ (\theta,\varphi) \in \mathcal{D}\right\} \tag{3.10}$$

the dataset of DTFs. In the case of PRTFs (i.e. for the FAST dataset), data in the contralateral hemisphere has little meaning due to the unrealistic contribution of the cylindric basis mesh in that area. Yet, we are dealing here with the matter of reducing the very high dimensionality of HRTF sets. Hence, in order to emulate the more general matter of HRTFs, in what follows PRTF and HRTF sets are not restricted to the ipsilateral hemisphere, unless indicated otherwise.

Following the aforementioned pre-processing step, we focused on the magnitude spectra of HRTFs. The logarithmic scale was chosen for its coherence with human perception. Furthermore, considering that HRTFs from left and right ears are largely symmetrical, and that the FAST dataset only contains left-ear data, we restricted this study to left-ear HRTFs. For all $i = 1,\ \ldots\ n$, $f \in F$ and $(\theta,\varphi) \in \mathcal{D}$, let there be such a mag-HRTF

$$G_i(f,\theta,\varphi) = 20 \cdot \log_{10}\left|H_i^{(\mathrm{L})}(f,\theta,\varphi)\right| \tag{3.11}$$

and

$$G = \{G_i(f,\theta,\varphi) \mid i = 1,\ \ldots\ n,\ f \in F,\ (\theta,\varphi) \in \mathcal{D}\} \tag{3.12}$$

the corresponding mag-HRTF dataset.

As we have reviewed in Section 2.1.3, there are several ways of performing PCA on HRTF data depending on whether the variations to be learned are along frequencies, directions and/or subjects. As mentioned above, we hereby consider the inter-individual PCA approach described in Section 2.1.3. Hence, the mag-HRTFs from the n_d directions were concatenated into a row vector $\mathbf{g}_i \in \mathbb{R}^{n_f n_d}$ for each subject $i = 1, \ldots n$

$$\mathbf{g}_i = \Big[G_i(f_1, \theta_1, \varphi_1) \ \ldots \ G_i(f_{n_f}, \theta_1, \varphi_1) \ \ldots \ G_i(f_1, \theta_{n_d}, \varphi_{n_d}) \ \ldots \ G_i(f_{n_f}, \theta_{n_d}, \varphi_{n_d}) \Big]. \quad (3.13)$$

The n row vectors were then stacked into the data matrix

$$\mathbf{X}_G = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} \in \mathbb{R}^{n \times n_f n_d}. \quad (3.14)$$

Principal component analysis

Let there be $\bar{\mathbf{g}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i$ the average mag-HRTF set and

$$\bar{\mathbf{X}}_G = \begin{bmatrix} \bar{\mathbf{g}} \\ \vdots \\ \bar{\mathbf{g}} \end{bmatrix} \in \mathbb{R}^{n \times n_f n_d} \quad (3.15)$$

the matrix constituted of the average mag-HRTF set stacked n times. Finally, let $\mathbf{\Gamma}_G \in \mathbb{R}^{n_f n_d \times n_f n_d}$ be the covariance matrix of $\mathbf{X}_G$:

$$\mathbf{\Gamma}_G = \frac{1}{n-1} \Big(\mathbf{X}_G - \bar{\mathbf{X}}_G \Big)^{\mathrm{t}} \Big(\mathbf{X}_G - \bar{\mathbf{X}}_G \Big). \quad (3.16)$$

The PCA transform is then written as

$$\mathbf{Y}_G = \Big(\mathbf{X}_G - \bar{\mathbf{X}}_G \Big) \mathbf{U}_G{}^{\mathrm{t}}, \quad (3.17)$$

where $\mathbf{U}_G$ is obtained by diagonalizing the covariance matrix $\mathbf{\Gamma}_G$

$$\mathbf{\Gamma}_G = \mathbf{U}_G^{\mathrm{t}} \mathbf{\Sigma}_G{}^2 \mathbf{U}_G. \quad (3.18)$$

In the equations above, $\mathbf{\Sigma}_G{}^2 \in \mathbb{R}^{(n-1)\times(n-1)}$ is the diagonal matrix that contains the

eigenvalues of $\mathbf{\Gamma}_G$, ${\sigma_{G_1}}^2$, ${\sigma_{G_2}}^2$, $\dots {\sigma_{G_{n-1}}}^2$, ordered so that ${\sigma_{G_1}}^2 \geq {\sigma_{G_2}}^2 \geq \cdots \geq {\sigma_{G_{n-1}}}^2$

$$\mathbf{\Sigma}_G{}^2 = \begin{bmatrix} {\sigma_{G_1}}^2 & & \\ & \ddots & \\ & & {\sigma_{G_{n-1}}}^2 \end{bmatrix}, \tag{3.19}$$

and $\mathbf{U}_G \in \mathbb{R}^{(n-1)\times n_f n_d}$ is an orthogonal matrix that contains the corresponding eigenvectors $\mathbf{u}_{G_1}, \mathbf{u}_{G_2}, \dots \mathbf{u}_{G_{n-1}} \in \mathbb{R}^{n_f n_d}$

$$\mathbf{U}_G = \begin{bmatrix} \mathbf{u}_{G_1} \\ \vdots \\ \mathbf{u}_{G_{n-1}} \end{bmatrix}. \tag{3.20}$$

The eigenvalues denote how much variance in the input data is explained by the corresponding eigenvectors.

In the equations above, we implicitly set the number of principal components (PCs) to $n-1$, because all PCs after the $(n-1)^{th}$ are trivial, i.e. of null associated eigenvalue. Indeed, the number of examples n is lower than the data dimension $n_f n_d$ and the data is centered, thus

$$r = \operatorname{rank}\left(\mathbf{X}_G - \bar{\mathbf{X}}_G\right) \leq n-1. \tag{3.21}$$

Hence, the rank of the covariance matrix does not exceed $n-1$ either:

$$\operatorname{rank}\left(\mathbf{\Gamma}_G\right) \leq \min\left(r, r\right) = r \leq n-1. \tag{3.22}$$

Dimensionality reduction

PCA can be used as a dimensionality reduction technique by retaining only the first p PCs and setting the weights of the discarded PCs to zero [Jolliffe02], where $p \in \{0, \dots n-1\}$:

$$\tilde{\mathbf{Y}}_G^{(p)} = \begin{bmatrix} y_{G_{1,1}} & \dots & y_{G_{1,p}} & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{G_{n,1}} & \dots & y_{G_{n,p}} & 0 & \dots & 0 \end{bmatrix}, \tag{3.23}$$

where $y_{G_{i,j}}$ is the value of matrix $\mathbf{Y}_G$ at the i^{th} row and j^{th} column for all $i = 1, \dots n$ and $j = 1, \dots n-1$.

The choice of the p first PCs (rather than another subset of p PCs) is motivated by

the fact that, by construction, a given PC represent more variation in the dataset than the next one.

Approximated data can then be reconstructed by inverting Equation (3.17):

$$\tilde{\mathbf{X}}_G^{(p)} = \tilde{\mathbf{Y}}_G^{(p)} \mathbf{U}_G + \bar{\mathbf{X}}_G. \tag{3.24}$$

Cumulative percentage of total variation

A simple but useful metric to evaluate the capacity of a PCA model to reduce dimensionality is the cumulative percentage of total variation (CPV) [Jolliffe02, Chap. 6, Sec. 1].

$$\mathrm{CPV}_G(p) = 100 \cdot \left(\sum_{j=1}^{p} {\sigma_{G}}_j^2 \right) / \left(\sum_{j=1}^{n-1} {\sigma_{G}}_j^2 \right), \tag{3.25}$$

and $p \in \{1, \ \ldots \ n-1\}$ is the number of retained PCs.

The CPV is closely related to the dimensionality reduction-related mean-square reconstruction error (MSE) of the training set [Jolliffe02, Chap. 6, Sec. 1]. This relation can be expressed as follows:

$$\mathrm{CPV}_G(p) = 100 \cdot \left(1 - \frac{\mathrm{MSE}(\tilde{\mathbf{X}}_G^{(p)}, \mathbf{X}_G)}{\mathrm{MSE}(\bar{\mathbf{X}}_G, \mathbf{X}_G)} \right), \tag{3.26}$$

where

$$\mathrm{MSE}(\mathbf{A}, \mathbf{B}) = \frac{1}{q} \frac{1}{r} (\mathbf{A} - \mathbf{B}) (\mathbf{A} - \mathbf{B})^{\mathrm{t}}, \tag{3.27}$$

for all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{q \times r}$ and $q, r \in \mathbb{N}^*$. Let us note that the MSE of two log-magnitude HRTF sets thus expressed is equal to their squared global SD (see Equation (2.12), Section 2.2.1).

By definition, the CPV increases from 0 to 100 % as a function of the number of retained PCs p (see Figure 3.13). A common criteria to choose how many PCs should be retained is to set an arbitrary threshold of CPV, and to select the lowest value of p that allows the CPV to overcome the threshold. As noted by Jolliffe in [Jolliffe02, Chap. 6, Sec. 1], despite the simplicity of this criteria its seems to work well in most cases, although the CPV threshold should be treated with flexibility and adapted to context. In particular, he notes that "attempts to construct rules having more sound statistical foundations seem [...] to offer little advantage over the simpler rules in most circumstances". The study by Hölzl [Hölzl14, Chap. 5] against which we compare our results, in particular, uses this CPV-based criteria to evaluate dimensionality reduction performance, with a CPV

	Subjects n_s	Data dim. $n_f n_d$	Directions n_d	Freq. bins n_f
ARI	198	133300	1550	86
FAST	119	412482	2562	161
RIEC	105	147915	865	171
$\text{HUTUBS}_{\text{meas}}$	96	41360	440	94
$\text{HUTUBS}_{\text{simu}}$	96	162620	1730	94
LISTEN	51	31977	187	171
BiLi	55	144480	1680	86
ITA	46	209902	2233	94
CIPIC	45	52718	1226	43

Table 3.1 – Number of subjects and data dimensionality for each dataset under study.

threshold of 90 %.

3.2.2 Cumulative Percentage of Total Variation of 9 Datasets

Hereon, we provide an overview of the dimensionality reduction performance of inter-individual PCA on one PRTF dataset, FAST, and 8 HRTF datasets: ARI, RIEC, measured HUTUBS, simulated HUTUBS, LISTEN, BiLi, ITA and CIPIC. The CPV curves that correspond to PCA on each dataset are plotted in Figure 3.13.

Datasets under study

While a description and review of the datasets under study is available in Chapter 2, we hereby summarize in Table 3.1 the number of subjects and the dimensionality of the data i.e. the dimensions of the data matrix $\mathbf{X}_G$.

The spatial and frequency resolutions of the HRTF sets vary greatly between datasets (see Chapter 2). Data dimensionality varies accordingly, ranging from 31977 (LISTEN) to 412482 degrees of freedom (FAST). Regarding the number of subjects, four of the datasets under study include about 50, two (HUTUBS and RIEC) about 100, FAST 119, while ARI has 200. Let us point out that we left out 3 ARI HRTF sets (subjects NH10, NH22 and NH826) because their measurement grids differed from the rest of the dataset.

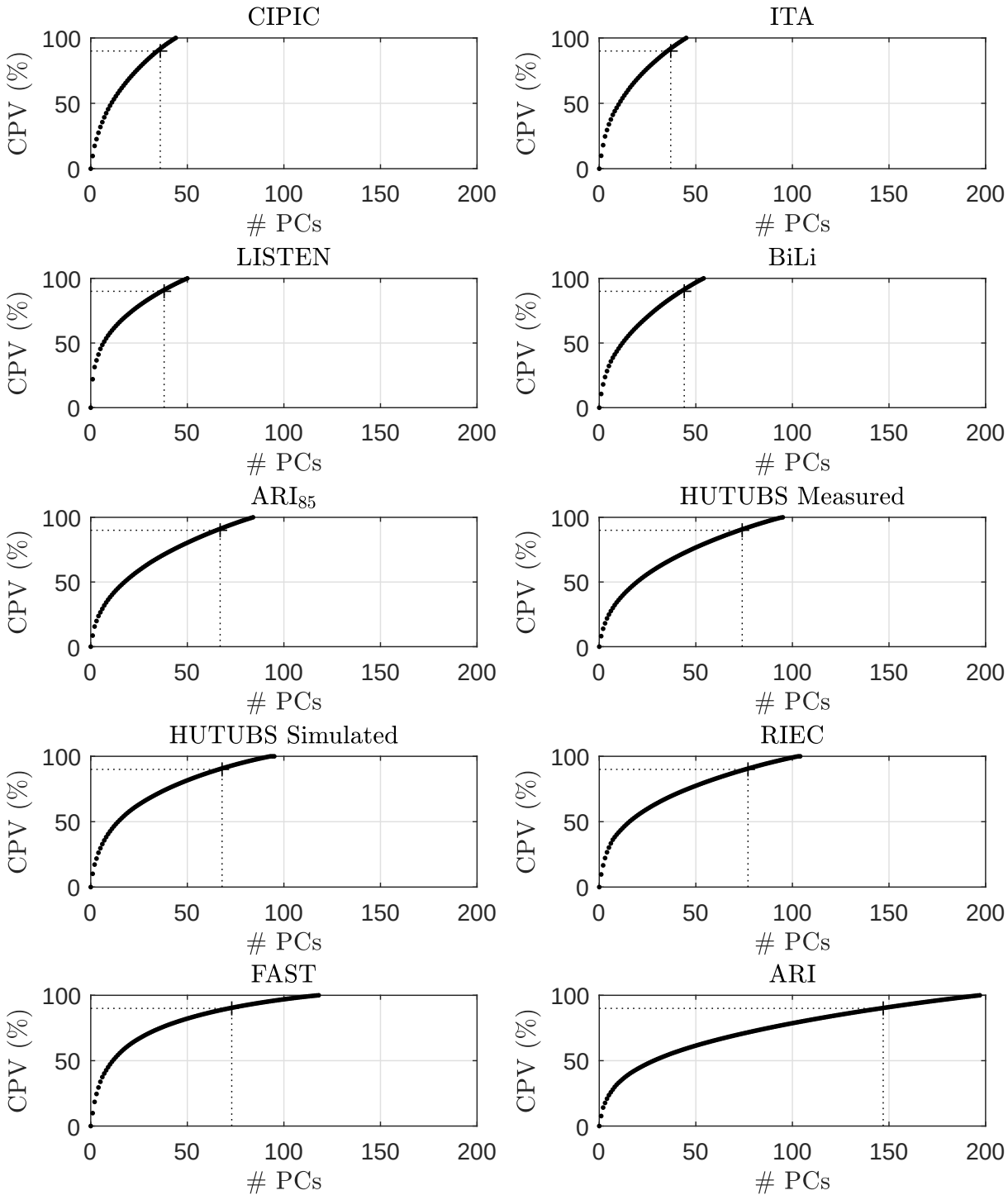


Figure 3.13 – CPV as a function of the number of the retained PCs p for each dataset under study.

	Hölzl's p_{90}	Our p_{90}	# PCs
ARI_{85}	59	66	84
CIPIC	36	35	44
LISTEN	38	37	50

Table 3.2 – Number of PCs p_{90} required to reach a CPV of 90 % according to Hölzl's study and ours. As a reference, the total number of non-trivial PCs, i.e. $n-1$, is displayed in the last column.

Comparison with the literature

In his Masters thesis [Hölzl14, Chap. 5] Hölzl compares various manners of formatting HRTF data prior to PCA, including the inter-individual approach used in the present work. The criteria used by Hölzl to evaluate the dimensionality reduction performance of PCA in the various configurations under study is the number of PCs required to reach a CPV of 90 %:

$$p_{90} = \min \{p \in \{0, \ \dots \ n-1\} \mid \text{CPV}(p) \geq 90\%\} \,. \tag{3.28}$$

Results for the configuration that corresponds to our proposed approach (inter-individual formatting, left ear only, log-magnitude HRTFs, no smoothing) can be found in the first row and last column of Tables C.1, C.2 and C.3 of [Hölzl14], for the ARI, CIPIC and LISTEN datasets, respectively. We hereby report these results in Table 3.2 for comparison with our own. Please note that in Hölzl's study, an older version of the ARI dataset was used which included only 85 subjects. As we could not find this older release of ARI, we performed PCA on a 85-subject subset of the latest version of the ARI dataset, in order for our results to be somewhat comparable. We refer to this subset as ARI_{85} in what follows.

Overall, we observe good coherence between our study and Hölzl's. In particular, we observe very close results for the CIPIC and LISTEN datasets (a difference of only one PC from one study to the other). With the ARI dataset, we can note a difference of p_{90} of about 10 % (66 in our case versus 59 in Hölzl's). However, it seems reasonable to attribute this difference to our approximation of the ARI dataset used by Hölzl, in view of the very good coherence of our results with the two other datasets.

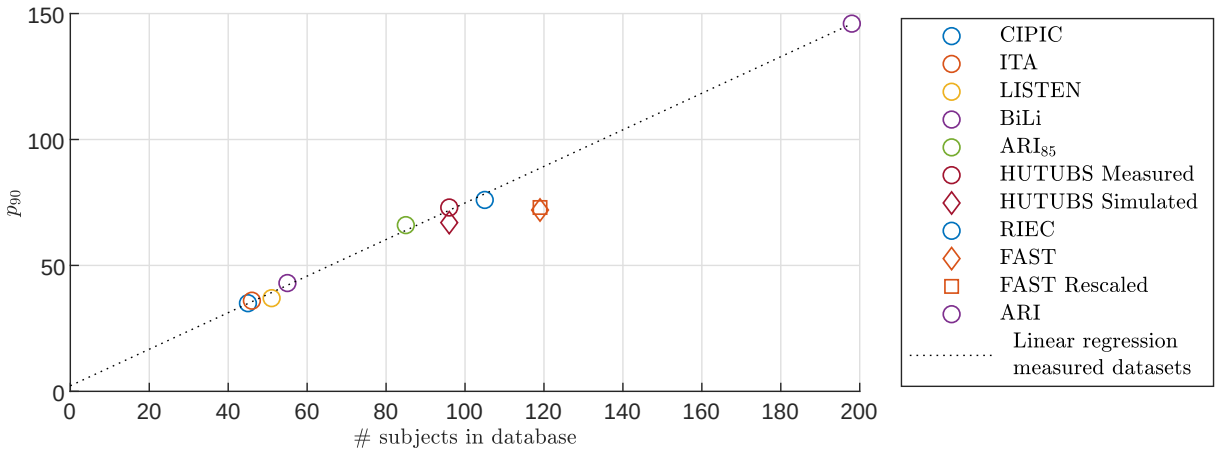


Figure 3.14 – Number of PCs p_{90} required to reach a CPV of 90 % for each dataset under study, displayed as a function of the number of subjects. Circles: acoustically measured datasets. Diamonds: numerically simulated datasets. Square: re-scaled FAST dataset.

Overview across datasets

The p_{90} calculated for each datasets is plotted in Figure 3.14, as a function of the number of subjects. The circles stand for datasets obtained though "classical" acoustical measurements, while the diamonds correspond to datasets constituted by means of numerical simulations.

Remarkably, all points are mostly aligned. In particular, performing a linear regression on acoustically measured datasets gave an excellent fit, with a coefficient of determination $R^2 = 0.998$ (for slope and offset parameters of 0.73 and 2.21, respectively).

Comparing the two HUTUBs datasets, one can note that, in spite of a higher data dimensionality (due to a denser spatial grid), the simulated one needs less PCs to retain 90 % of the total variance. A likely explanation is that acoustic measurement is more prone to variations from one session to the other, even for a motionless manikin [Andreopoulou15], which could induce increased variability in the dataset. While the 3-D morphology acquisition procedure may be variable from one acquisition to the other, the simulation itself is perfectly consistent across subjects. The fact that HUTUBS HRTFs were simulated from heads without torso may account for some decrease in variability as well.

The only other simulated dataset, FAST, also falls below the straight line. It however seems to be more of an outlier than HUTUBS (19 % below the line, against 7 % for HUTUBS). While the simulation approach probably explains part of it, two other

hypotheses are plausible. The first one is the fact that the ear shapes used for simulation were normalized in size, which corresponds to a normalization in frequency scaling of the PRTF sets. However, by re-introducing the frequency scaling factors in the PRTF sets, and performing PCA on this rescaled version of the FAST dataset, we found that the p_{90} is higher only by 1 PC. A second one is that contributions from the head and torso are absent from the PRTFs, thus reducing variability compared to full-morphology measurements or simulations. This reduced variability should be more prominent in the contralateral hemisphere and at low frequencies. Thus, we look below into the impact of leaving out these particular spatial and frequency ranges from PCA on the p_{90} of all datasets.

Despite the considerable variety in data dimensionality (spatial sampling, notably) and HRTF acquisition conditions, the number of PCs required to retain 90 % of total variation, p_{90}, increases in an approximately linear fashion with the number of subjects contained in the datasets. The measured datasets fit very well this linear trend, while the FAST and, to a lesser extent, the simulated HUTUBS datasets, lie slightly out of it. Both of them were simulated instead of measured, and were generated from 3-D geometries that excluded the torso (HUTUBS) or both the torso and head (FAST).

Effect of restricting the spatial and frequency ranges

As we have seen above, the magnitude PRTF sets from the FAST dataset seem to present a lesser inter-individual variability than other HRTF datasets. This difference is likely due to the absence of contribution from the head and torso in the PRTFs. We herein study the effect on the p_{90} of leaving out of PCA spatial or frequency ranges where contribution from the pinna is less prominent: the contralateral hemisphere and frequencies below 4 kHz.

Frequency range Restricting the frequency range to frequencies above 4 kHz had little to no effect on the p_{90}. The p_{90} remained identical for most datasets and decreased only by 1 for the three datasets which exhibited change, that is ARI, ARI_{85} and LISTEN.

It appears that data in the lower frequency range (where only the torso and head contribute to directional filtering) correspond to a very small proportion of the variability between magnitude HRTF sets of a given dataset. This is coherent with the fact that spectral features of HRTFs at these low frequencies are rather "smooth", unlike the sharp peaks and notches caused by the pinna, whose central frequencies and gains are very

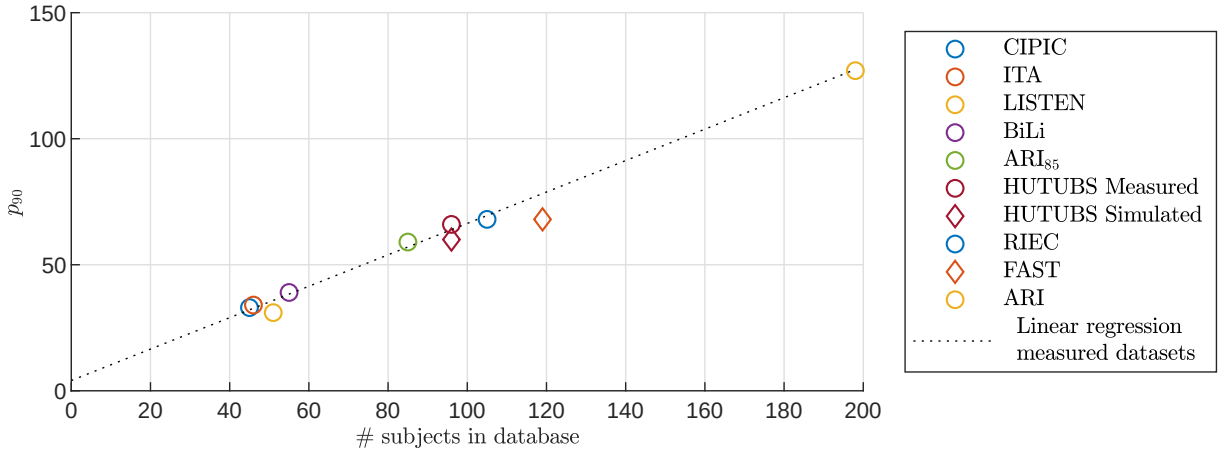


Figure 3.15 – Number of PCs p_{90} required to reach a CPV of 90 % for each dataset under study when HRTF sets are restricted to the ipsilateral hemisphere, displayed as a function of the number of subjects. Circles: acoustically measured datasets. Diamonds: numerically simulated datasets.

variable between individuals.

Ipsilateral hemisphere As can be seen in Figure 3.15, restricting the PRTF and HRTF sets to the ipsilateral hemisphere had a notable impact on the p_{90} of all datasets, decreasing it by 6 to 13 %. For all datasets, leaving the contralateral filters out of the statistical analysis reduced the total variance in the training set, thus decreasing the number of PCs required to represent a given percentage – 90 % – of that variance.

The previously observed linear trends was preserved by the reduction of the spatial grid: a linear regression on acoustically measured datasets yielded an excellent fit, with a coefficient of determination $R^2 = 0.995$ (for slope and offset parameters of 0.62 and 4.1, respectively).

The FAST PRTF dataset remains an outlier to that linear trend, but is somewhat closer to it. This can be explained by the fact that the contribution of the cylindric basis mesh (more prominent in the contralateral hemisphere) is not subject to variation between subjects. In contrast, other datasets include the contribution of a head and/or torso whose shape varies between individuals.

Overall, restriction to higher frequencies had almost no effect, and restriction to the ipsilateral hemisphere reduced the variability of all datasets while mostly preserving the previously observed trends. Although contralateral data has little meaning in the case of

PRTFs, removing the contralateral hemisphere reduces for all datasets the dimensionality of the data by half and removes a substantial part of the variability (p_{90} decreased by 7 % for the FAST PRTF dataset). Thus, in order to stay close to the more general matter of reducing the dimensionality of HRTF sets, we hereon consider PCA models built on unrestricted spatial and frequency grids.

3.2.3 Reconstruction Error Distribution

In the previous section, we looked into the number of PCs required to retain 90 % of the variance of 9 datasets of HRTF or PRTF sets. However, the CPV does not inform us on the type of information that is lost when reducing the dimensionality of magnitude HRTF sets. Thus, we herein look into the distribution of that loss of information (i.e. the reconstruction error) over the frequency and spatial domains for two models: the FAST PRTF model and the ARI HRTF model.

Frequency dependency The dimensionality-reduced magnitude mag-HRTF set of exemplary subjects from the FAST and ARI datasets are plotted in Figure 3.16, along with the original mag-HRTF set and the difference between them in the dB domain. As can be seen in that figure, the magnitude HRTF sets are somewhat "smoothed" by the dimensionality reduction process: progressive changes in gain across frequency bins and directions are better reconstructed than sharper ones.

For both HRTF sets, reconstruction errors are low below 1 kHz, and at their largest beyond 4-5 kHz, which is coherent with the average behavior observed in Figure 3.18. Indeed, the root-mean-squared reconstruction error (across all subjects and directions) increases with frequency. In contrast with the ARI HRTF model, this error is almost zero for the FAST PRTF model for frequencies up to 4 kHz. This is coherent with the fact that the pinna has little effect on HRTFs and PRTFs in that frequency range, and that in the meshes used to compute the FAST PRTFs, only the pinna varies from one "subject" to the other. There is a large increase of this error around 16 kHz, which is a side effect of our pre-processing: re-sampling the HRTFs independently at each direction caused them to have little coherence between directions and subjects at these high frequencies. The data beyond 15 kHz thus has little meaning but, due to the aforementioned lack of directional coherence, is "seen" as noise by the PCA i.e. associated to the very last PCs.

Finally, the reconstruction error for the FAST PRTF model is lower than that of the ARI HRTF model, regardless of the frequency. This was expected, seeing that the

reconstruction MSE with $p = p_{90}$ corresponds to 90 % of the total variance and that the total variance of the FAST PRTF dataset is lower than that of the ARI HRTF dataset: 13 dB^2 (standard deviation 3.5 dB) and 18 dB^2 (standard deviation 4.3 dB), respectively.

As can be seen in Figure 3.17, the distribution over the frequency range of the reconstruction error with p_{90} PCs is similar to that of the dataset's variance, i.e. the reconstruction error with $p = 0$ PCs.

Directional dependency The root-mean-squared reconstruction error with p_{90} (averaged over all subjects and frequencies) is plotted as a function of direction in Figure 3.20 for both datasets. As can be seen in that figure and in the exemplary reconstructions (see Figure 3.16), the error is more important in the contralateral hemisphere than in the ipsilateral one, in particular for the ARI model.

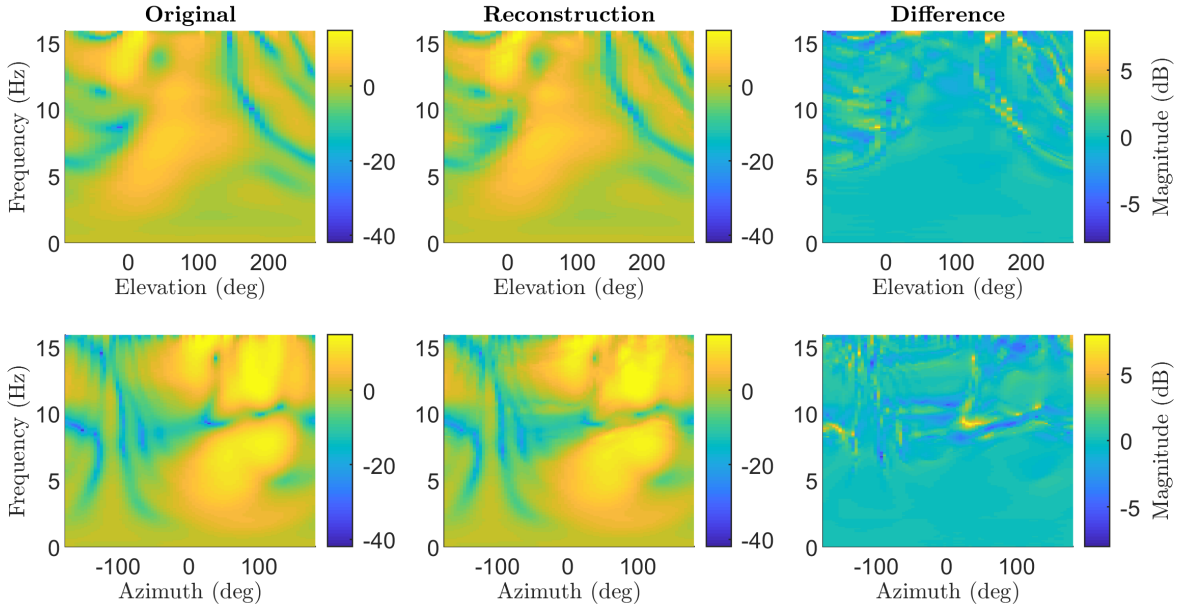
By comparison, the spatial distribution of the root-mean-squared reconstruction error for $p = 0$ – i.e. the variability – of each dataset is plotted in Figure 3.19. The variance of the ARI mag-HRTF dataset is substantially larger in the ipsilateral hemisphere than in the contralateral one. Indeed, head shadowing causes the magnitude of HRTFs to be generally lower in the contralateral region than in the ipsilateral one. In contrast, this shadowing effect is almost absent in PRTFs, and the variance of the FAST mag-PRTF dataset is more uniformly distributed between both hemispheres.

Conclusion

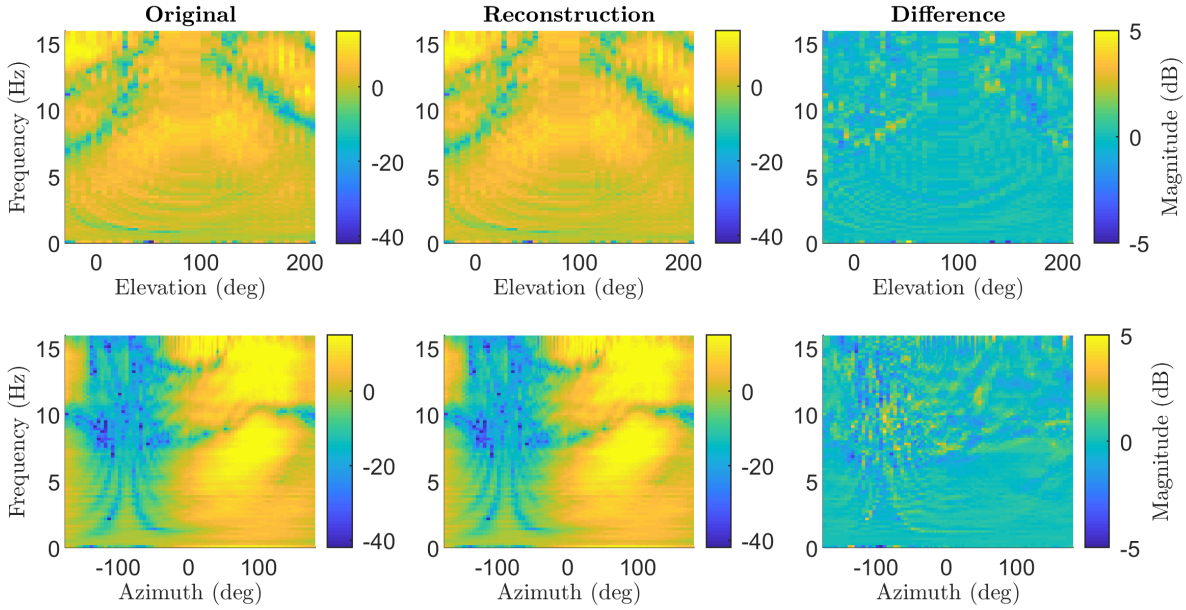
In this section, we have investigated the dimensionality reduction performance of 9 PCA models of log-magnitude HRTFs, trained on 8 public HRTF datasets and FAST.

Having checked that our results on the ARI, CIPIC and LISTEN datasets were coherent with the literature, we observed an interesting trend. Indeed, the number of PCs required to retain 90 % of the information, p_{90}, increases linearly with the size of the dataset. This suggests that these datasets are too small to be representative of the space of log-magnitude HRTF sets in general – by means of linear combinations. Otherwise, a slowdown in p_{90}'s increase would be observed.

In other words, if there exists a linear manifold representative of the inter-individual variations of log-magnitude HRTF sets, currently available datasets are too small for PCA to identify it. Although a non-linear manifold could exist, there are few examples compared to the dimensionality of the data, possibly too few for a more complex, non-linear machine learning technique. Under both hypotheses, a larger-scale dataset would

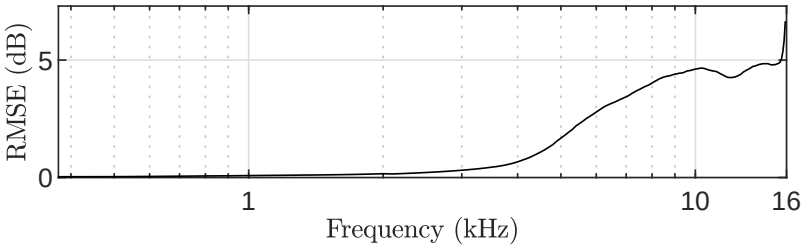


(a) FAST model, subject PB, 72/118 PCs

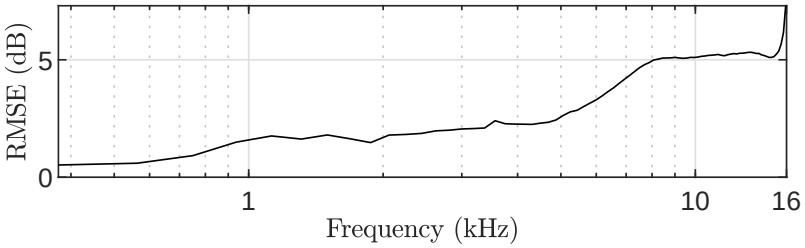


(b) ARI model, subject NH08, 146/197 PCs

Figure 3.16 – Reconstruction of the mag-HRTF set of two exemplary subjects: (a) subject PB from the FAST dataset with $p_{90} = 72$ out of 118 PCs, and (b) subject NH08 from the ARI dataset with $p_{90} = 146$ out of 197 PCs. Left to right: original, reconstructed and difference (in the dB domain) between original and reconstructed magnitude HRTF sets. All magnitude HRTF sets are plotted for directions of the median (first row) and horizontal planes (second row).

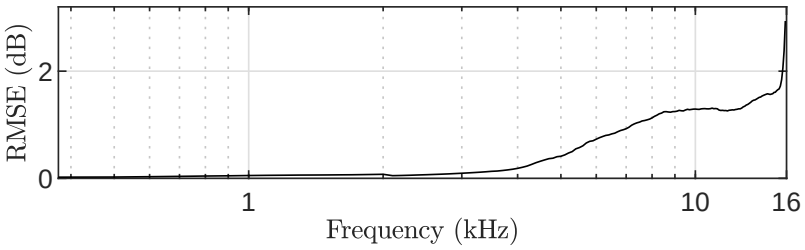


(a) FAST model, 0/118 PCs

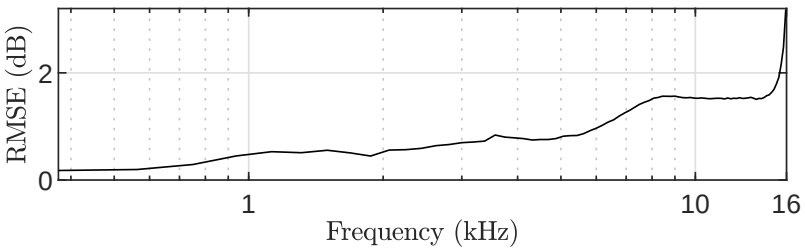


(b) ARI model, 0/197 PCs

Figure 3.17 – Reconstruction RMSE as a function of frequency for (a) the FAST PCA model with 0/118 PCs, and (b) the ARI PCA model with 0/197 PCs.

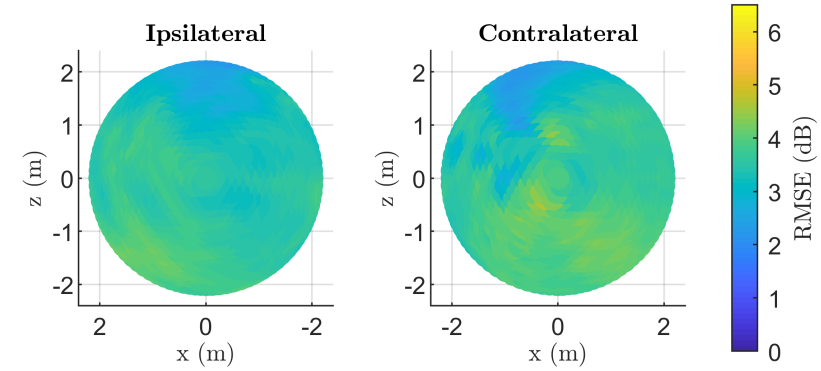


(a) FAST model, 72/118 PCs

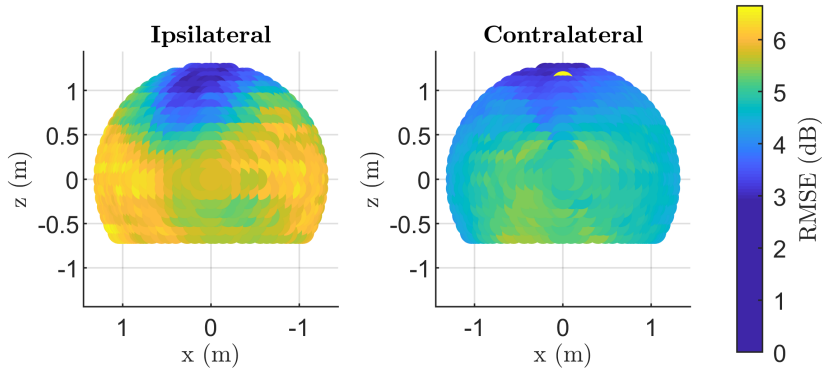


(b) ARI model, 146/197 PCs

Figure 3.18 – Reconstruction RMSE as a function of frequency for (a) the FAST PCA model with $p_{90} = 72/118$ PCs, and (b) the ARI PCA model with $p_{90} = 146/197$ PCs.

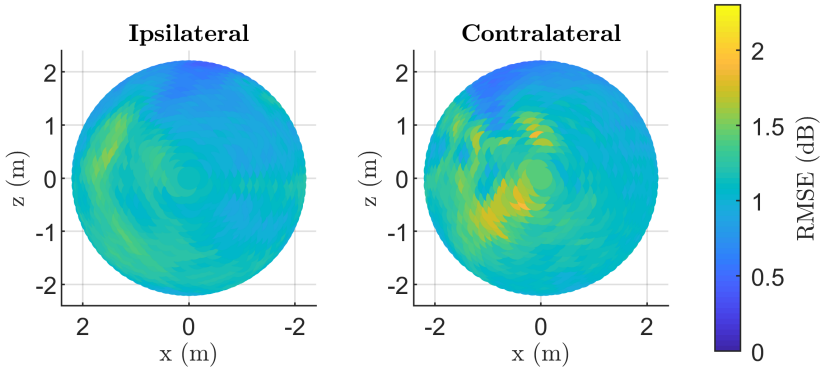


(a) FAST model, 0/118 PCs

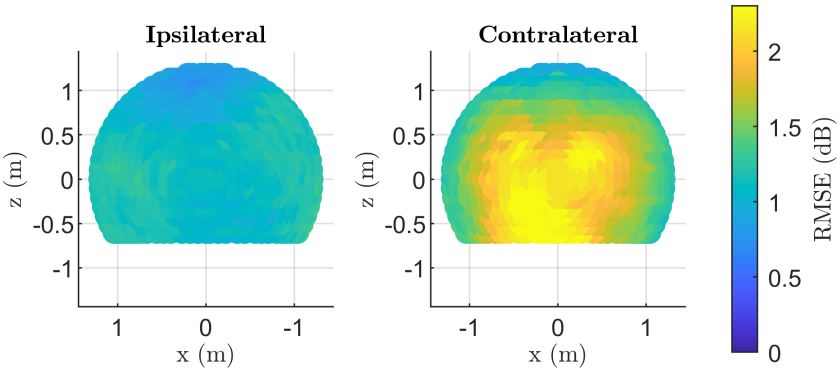


(b) ARI model, 0/197 PCs

Figure 3.19 – Reconstruction RMSE as a function of direction for (a) the FAST PCA model with 0/118 PCs, and (b) the ARI PCA model with 0/197 PCs.



(a) FAST model, 72/118 PCs



(b) ARI model, 146/197 PCs

Figure 3.20 – Reconstruction RMSE as a function of direction for (a) the FAST PCA model with $p_{90} = 72/118$ PCs, and (b) the ARI PCA model with $p_{90} = 146/197$ PCs.

be desirable.

3.3 Compared Dimensionality Reductions of Ear Shapes and Matching PRTF Sets

In the previous section, we studied the dimensionality reduction performance of PCA on log-magnitude HRTF sets from various datasets, including the FAST one. Results suggested that current datasets include too few examples for PCA to be able to find a linear subspace representative of log-magnitude HRTF sets in general.

In this section, we deal with the preliminary study of the FAST dataset that led to designing a data augmentation method. Taking advantage of the fact that the FAST dataset includes registered pinna meshes, we investigated whether PCA performs better at reducing the dimensionality of 3-D ear morphology than of matching computed PRTF sets.

First, we study the ability of PCA to reduce the dimensionality of the 119 log-magnitude PRTF sets. Second, we present how we performed PCA on the corresponding 119 ear point clouds. Third, we compare the dimensionality reduction performances of both PCA models, as well as the statistical distribution of their respective PCs. Finally, we draw the conclusions that led us to propose the data augmentation scheme.

3.3.1 Principal Component Analysis of Ear Shapes

Let $E = \{\mathbf{e}_1, \ldots \mathbf{e}_n\}$ be the set of $n = 119$ ear point clouds from the FAST dataset whose x, y and z coordinates are concatenated into row vectors $\mathbf{e}_1, \ldots \mathbf{e}_n \in \mathbb{R}^{3n_v}$, with $3n_v = 54528$.

In order to build a statistical shape model of the pinna, the ear point clouds were gathered into a data matrix as follows. Let there be $\mathbf{X}_E = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{bmatrix} \in \mathbb{R}^{n \times 3n_v}$ the data

matrix, $\bar{\mathbf{e}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_i$ the average ear shape and

$$\bar{\mathbf{X}}_E = \begin{bmatrix} \bar{\mathbf{e}} \\ \vdots \\ \bar{\mathbf{e}} \end{bmatrix} \in \mathbb{R}^{n\times 3n_v} \tag{3.29}$$

the matrix constituted of the average shape stacked n times. Finally, let $\mathbf{\Gamma}_E \in \mathbb{R}^{3n_v\times 3n_v}$ be the covariance matrix of $\mathbf{X}_E$:

$$\mathbf{\Gamma}_E = \frac{1}{n-1}\left(\mathbf{X}_E - \bar{\mathbf{X}}_E\right)^{\mathrm{t}}\left(\mathbf{X}_E - \bar{\mathbf{X}}_E\right). \tag{3.30}$$

Similarly to the case of magnitude PRTF sets (see Section 3.2), we performed PCA on the data matrix $\mathbf{X}_E$ according to Equations (3.17), (3.18) and (3.20). The number of non-trivial PCs is $(n-1)$ in this case as well, due to the fact that $n < 3n_v$. From the set of ear point clouds E described in Section 3.1.1, we classically constructed a statistical 3-D shape model of the pinna using PCA [Rajamani07].

Behavior of the first principal components

The behavior of the first principal components can be observed as follows.

For each PC of index $j \in \{1, 2, 3\}$, we set the j^{th} PC weight to $\lambda\sigma_{E_j}$ and all other PC weights to zero, with $\lambda \in \{-5, -3, -1, +1, +3, +5\}$ and reconstructed the corresponding ear point cloud $\mathbf{e}_{v_j}(\lambda)$ by inverting Equation (3.17)

$$\mathbf{e}_{v_j}(\lambda) = \left(0 \ \ldots \ 0 \ \lambda\sigma_{E_j} \ 0 \ \ldots \ 0\right)\mathbf{U}_E + \bar{\mathbf{e}}. \tag{3.31}$$

Meshes derived from these ear point clouds are displayed in Figure 3.21, colored with the vertex-to-vertex euclidean distance to the average shape.

The first one seems to control vertical pinna elongation including concha height and lobe length up to disappearance, as well as some pinna vertical axis rotation. The second one seems to encode the intensity of some topography features such as triangular fossa depth or helix prominence. It also has an impact on concha shape and vertical axis rotation. The third PC seems to have a strong influence on concha depth, triangular fossa depth as well as upper helix shape.

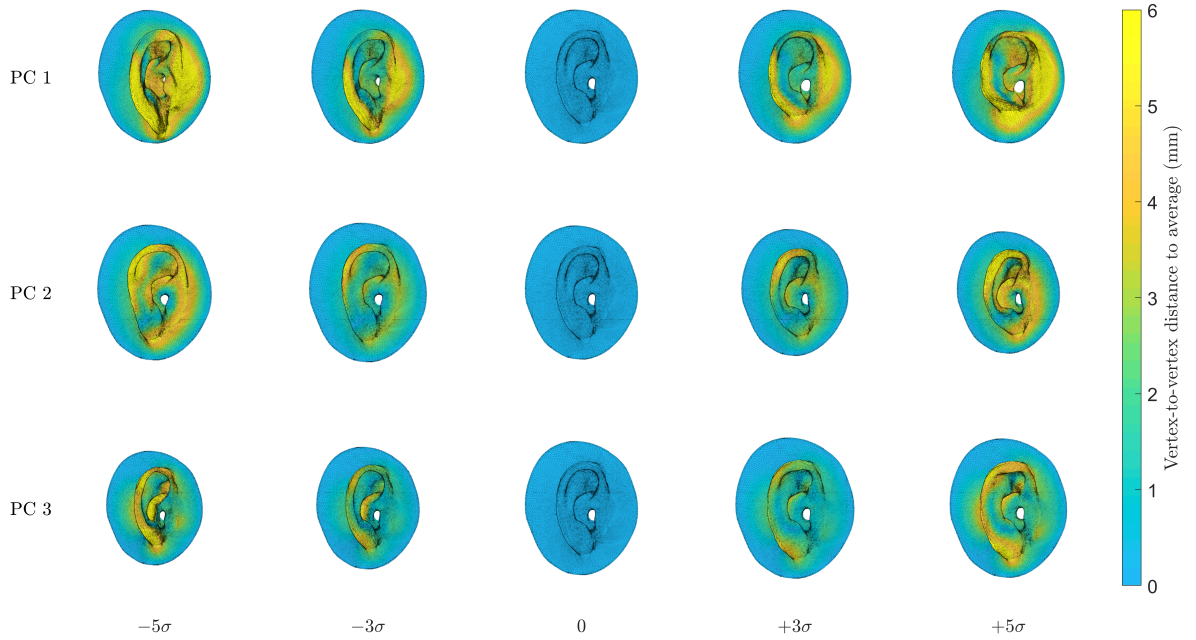


Figure 3.21 – First three principal components (PCs) of the PCA ear shape model. Rows: PC of index $j \in \{1, 2, 3\}$. Columns: Weight assigned to given PC, indicated in proportion of its standard deviation σ_{E_j}.

Behavior of the first principal components

Various log-magnitude PRTF sets that illustrate the behavior of the three first PCs were reconstructed according to Equation (3.31).

They are plotted in Figure 3.22 for directions that belong to the median sagittal plane. As it was expected, no variations are visible below 5 kHz: at these wavelengths the pinna have little impact on sound propagation. Each PC appears to represent a different pattern of change in anterior and posterior directions, although only the first one seems to have a strong influence on directions above the head. However, it does not seem possible to distinguish patterns that are limited to a certain range of directions and/or frequencies. Interestingly, it seems that changes in the first PC weight results in a frequency shift in the PRTFs. As the pinnae used to construct the model are normalized in size, this effect likely corresponds to variations in the volume of the pinna's interior cavities, such as the concha or triangular fossa.

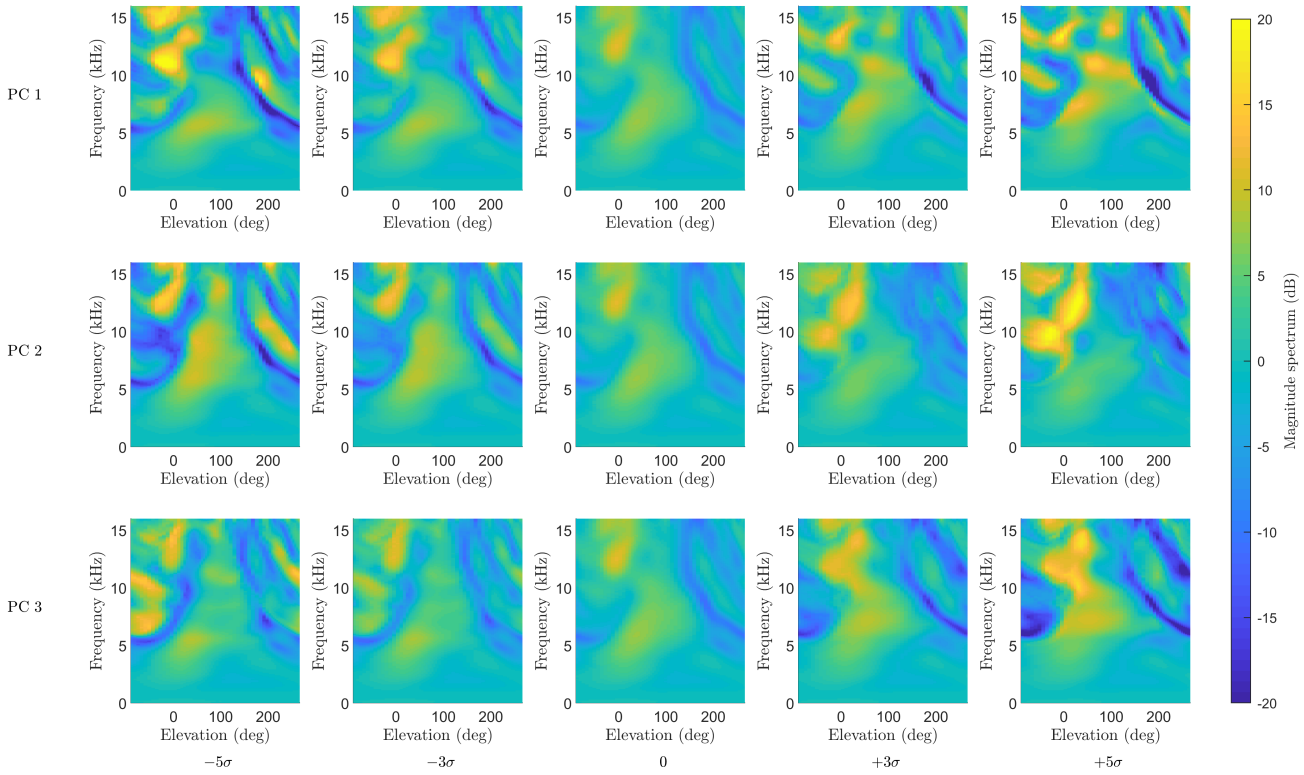


Figure 3.22 – First Principal Components (PCs) of the PCA model of log-magnitude PRTFs. Reconstructed PRTF sets are plotted in the median sagittal plane. Rows: PC. Columns: Weight assigned to a given PC, indicated in proportion of its standard deviation σ.

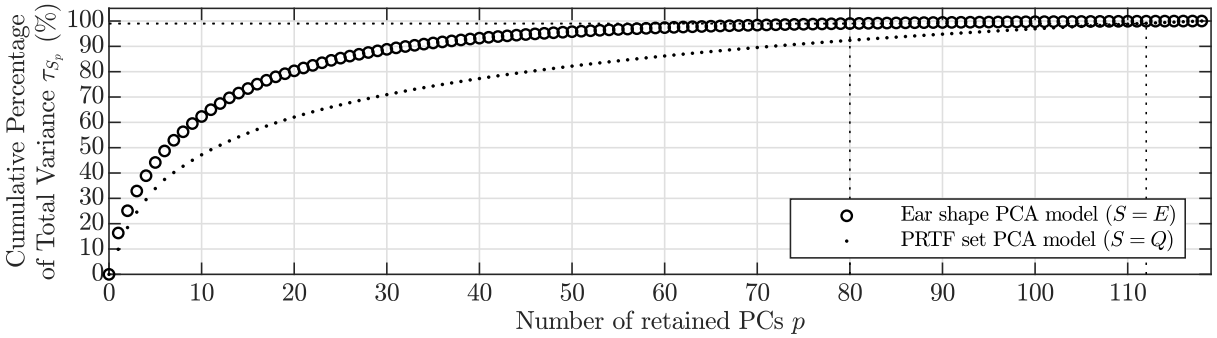


Figure 3.23 – $\mathrm{CPV}_S(p)$ as a function of the number of retained PCs $p \in \{0, \ldots n-1\}$ for either PCA model. Circles: ear shape model ($S = E$). Dots: PRTF set model ($S = Q$).

3.3.2 Comparison of Both PCA Models

For all ear shapes $\mathbf{e}_i \in E$, let us denote $\mathbf{h}_i = \psi(\mathbf{e}_i) \in \mathbb{C}^{n_f \times n_d}$ the corresponding PRTF set, computed according to the process described in Section 3.1.

Additionally, let $\mathbf{g}_i \in \mathbb{R}^{n_f n_d}$ be the log-magnitude PRTF set derived from $\mathbf{h}_i$ according to the pre-processing step described in Section 3.2. Accordingly, let $\Psi : \mathbb{R}^{3n_v} \mapsto \mathbb{R}^{n_f \times n_d}$, defined by $\mathbf{e} \longmapsto \mathbf{g} = 20 \cdot \log_{10}(|\psi(\mathbf{e})|)$, be the process of deriving a log-magnitude PRTF set from an ear point cloud, and let $G = \{\mathbf{g}_1, \ldots \mathbf{g}_n\} = \{\Psi(\mathbf{e}_1) \ldots \Psi(\mathbf{e}_n)\}$.

PCA was performed on the 119 log-magnitude PRTF sets from the FAST dataset in the inter-individual fashion described in Section 3.2. The number of non-trivial PCs is $(n-1)$ in this case as well, due to the fact that $n < n_f n_d$.

Dimensionality reduction performance

As in Section 3.2, we use CPV to compare the dimensionality reduction performances of both PCA models. CPVs for both models are plotted in Figure 3.23. While we previously used a CPV threshold of 90 % as a basis for comparison with the literature, we hereon prefer a more selective threshold of 99 %.

A first notable result is that, for the ear shape model, the 99 %-of-total-variance threshold is reached for $p = 80$ retained PCs, i.e. only $\frac{p}{n-1} = \frac{80}{118} = 67.8$ % of the maximum number of PCs.

In other words, the 118-dimensional linear subspace of $\mathbb{R}^{3n_v} = \mathbb{R}^{56661}$ defined by the $n = 119$ pinnae of our database can be described using only 80 parameters while maintaining a 'reasonable' reconstruction accuracy, in the sense of a vertex-to-vertex MSE. In the present example of a CPV of 99 %, this accuracy corresponds to a MSE of 1 % of its

maximum value. Indeed, for a CPV of 99 %, Equation (3.26) gives us:

$$\begin{aligned} \mathrm{MSE}(\tilde{\mathbf{X}}_E^{(p_{99})}, \mathbf{X}_E) &= \left(1 - \frac{\mathrm{CPV}_E(p_{99})}{100}\right) \cdot \mathrm{MSE}(\bar{\mathbf{X}}_E, \mathbf{X}_E) \\ &= 0.01 \cdot \mathrm{MSE}(\bar{\mathbf{X}}_E, \mathbf{X}_E). \end{aligned} \tag{3.32}$$

More importantly, PCA appears to be largely more successful at reducing the dimension of ear shapes $\mathbf{e}_i$ than that of magnitude PRTF sets calculated from the same ear shapes $\mathbf{g}_i = \Psi(\mathbf{e}_i)$. Indeed, the PRTF CPV is substantially lower than the ear shape CPV for any number of retained PCs. For instance, the 99 %-of-total-variance threshold is reached for 112 PCs out of 118 for the PRTF model against 80 out of 118 for the ear shape one.

Statistical distribution

Going further in our comparison of both PCA models, we looked into the statistical distribution of the data in both 118-dimensional PCA subspaces.

To do so, we tested the PCs of each model for multivariate normal distribution using Royston's test [Royston83], performed on the columns of the PC weights matrix $\mathbf{Y}_S$, where $S \in \{E, G\}$ denotes the dataset.

The outcome of the test was an associated p-value of 0.037 in the case of ear point clouds, and 0.000 in the case of mag-PRTF sets, where the p-value refers to the null hypothesis that the distribution is not multivariate normal. In other words, the ear model's PC weights can be considered to be multivariate-normally distributed with a significance level of 3.7 %, while its PRTF counterpart's fail the test for any significance level.

Conclusion

We found that PCA performs largely better at reducing the dimensionality of the 119 3-D ear shapes than of the log-magnitude PRTF sets derived from them. In particular, in contrast with the case of log-magnitude PRTF sets, PCA allowed us to identify an 80-dimensional linear subspace in which the 119 training examples can be represented while retaining 99 % of the information. Moreover, the ear point cloud PC weights follow a multivariate normal distribution.

Overall, the ear shape PCA model seems more suited than its PRTF counterpart for the random generation of new data.

3.4 Dataset Augmentation

Based on the conclusions drawn in Section 3.3, we devised and implemented a method to augment the FAST dataset that uses the space of 3-D ear morphology as a back door to randomly generate new examples (see Figure 3.24 for an overview). This method allows the generation of new data (pinna meshes and matching PRTF sets) from existing data, based on the statistical distribution of the latter. Such a process is commonly referred to as "data augmentation" in the field of machine learning, and is used to overcome the recurring problem of limited dataset size in applications that require a lot of data – generally neural-network-based.

In the present section, we introduce this process and the resulting dataset, named WiDESPREaD (a wide dataset of ear shapes and pinna-related transfer functions). First, we explain how we used the PCA model of ear shapes presented in Section 3.3 to randomly generate over a thousand ear meshes. Then, we go over how PRTF sets were derived from those meshes by means of FM-BEM calculations. Finally, we take a look at a few examples from the augmented dataset.

3.4.1 Random Generation of Ear Meshes

The statistical ear shape model learned from dataset E and presented in Section 3.3.1 can be used as a generative model. By construction, the model's PCs (i.e. the columns of $\mathbf{Y}_E$) are of zero mean and are mutually uncorrelated, i.e. statistically independent up to the second order. Furthermore, as we have shown in Section 3.3, the columns of $\mathbf{Y}_E$ follow a multivariate normal distribution. They are thus mutually statistically independent (up to any order) and follow respective normal probability laws of zero mean and σ_{E_j} standard deviation $\mathcal{N}(0, \sigma_{E_j})$, where $j \in \{1, \ldots n-1\}$ represents the PC index. As a consequence, a new statistically realistic ear point cloud can be conveniently generated by randomly drawing a vector of PC weights according to the distribution of probability observed in the FAST dataset.

To constitute the WiDESPREaD dataset, an arbitrarily large number N of ear shapes $\mathbf{e}'_1, \ldots \mathbf{e}'_N \in \mathbb{R}^{3n_v}$ were thus randomly generated as follows. First, for all $i = 1, \ldots N$,

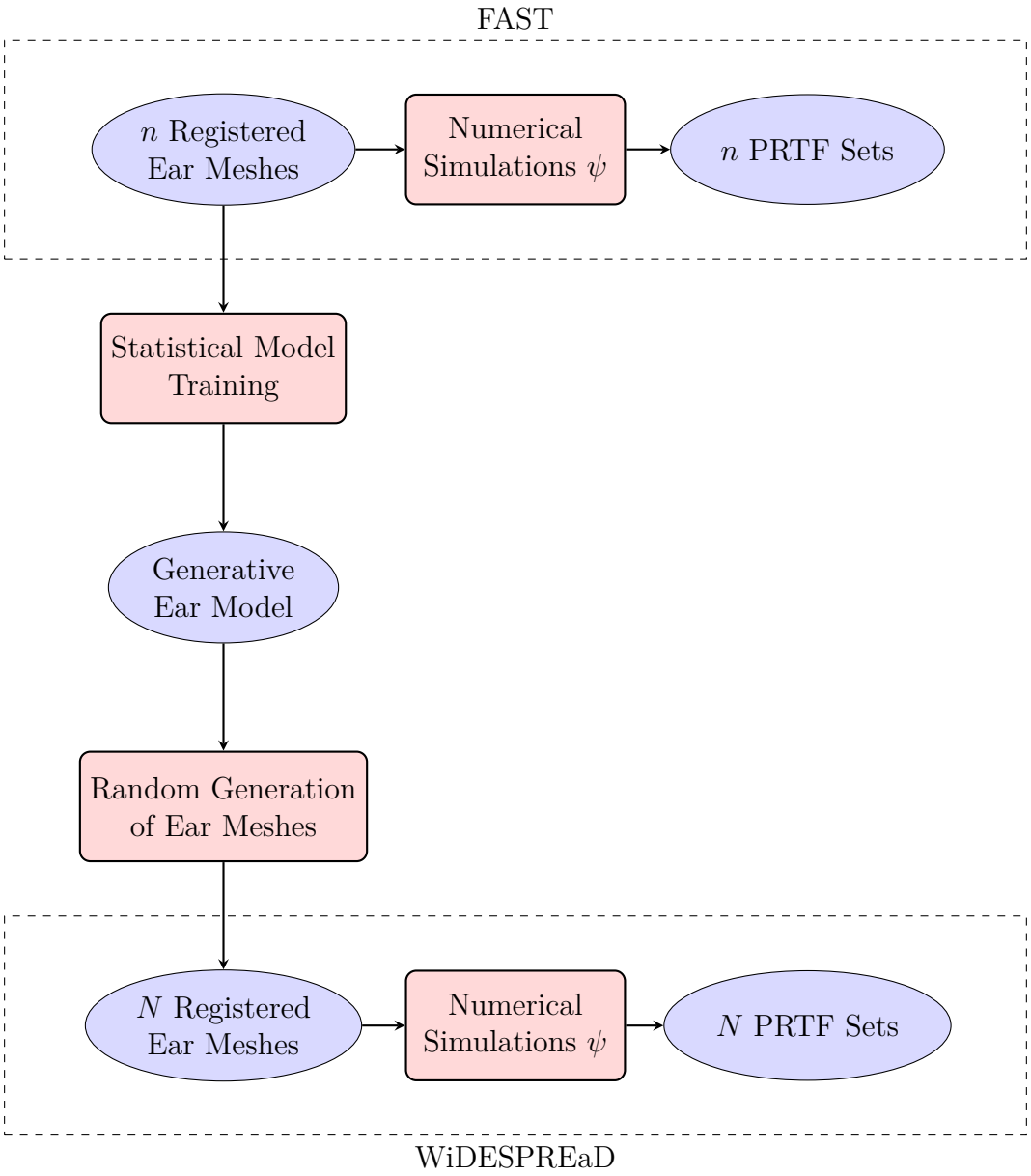


Figure 3.24 – Overview of the data augmentation process.

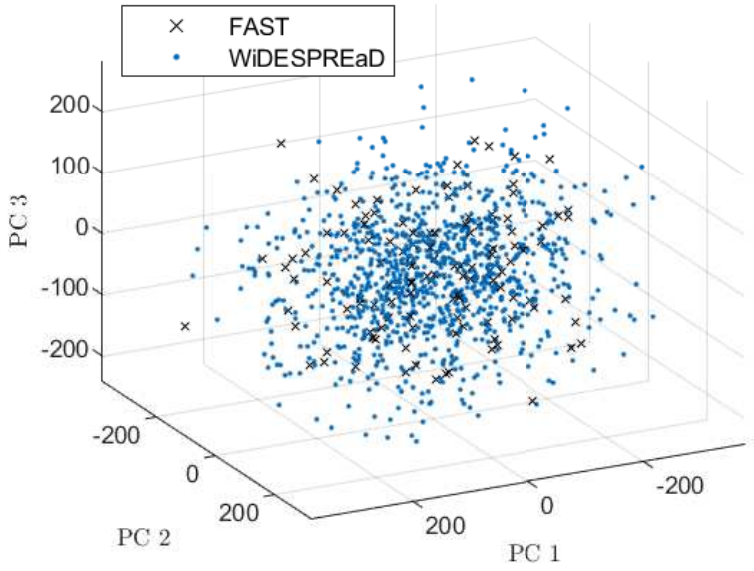


Figure 3.25 – The respective 119 and 1005 ear point clouds of the FAST (black crosses) and WiDESPREaD (blue dots) datasets, displayed in the space of the three first PCs.

a PC weights vector $\mathbf{y}_{E_i} = (y_{E_{i,1}}, \ \dots \ y_{E_{i,n-1}}) \in \mathbb{R}^{n-1}$ was obtained by drawing the $(n-1)$ PC weights $y_{E_{i,1}}, \ \dots \ y_{E_{i,n-1}}$ independently according to their respective probability laws $\mathcal{N}(0, \sigma_{E_1}), \dots \mathcal{N}(0, \sigma_{E_{n-1}})$. All $(n-1)$ PC weights were retained so as to follow the distribution observed in the initial dataset without introducing any bias related to dimensionality reduction. By construction, the N generated PC weights vectors populate the space of $\mathbb{R}^{n-1}$ in a manner that is statistically realistic with regard to what we have observed on real data, that is our dataset of ear point clouds from 119 human subjects. This is illustrated in three dimensions in Figure 3.25, where the pinna PC weight vectors of both the FAST and WiDESPREaD datasets are plotted in the space of the three first PCs.

Second, the corresponding ear shapes were reconstructed by inverting Equation (3.17)

$$\mathbf{X}'_E = \mathbf{U}_E \mathbf{Y}'_E + \bar{\mathbf{X}}_E, \tag{3.33}$$

where $\mathbf{Y}'_E \in \mathbb{R}^{N\times(n-1)}$ is the matrix whose rows are the N PC weights vectors

$$\mathbf{Y}'_E = \begin{bmatrix} \mathbf{y}'_{E_1} \\ \vdots \\ \mathbf{y}'_{E_N} \end{bmatrix} = \begin{bmatrix} y'_{E_{1,1}} & \cdots & y'_{E_{1,n-1}} \\ \vdots & \ddots & \vdots \\ y'_{E_{N,1}} & \cdots & y'_{E_{N,n-1}} \end{bmatrix}, \tag{3.34}$$

and $\mathbf{X}'_E \in \mathbb{R}^{N\times 3n_v}$ is the data matrix whose rows are the N ear shapes $\mathbf{e}'_1,\ \ldots\ \mathbf{e}'_N \in \mathbb{R}^{3n_v}$

$$\mathbf{X}'_E = \begin{bmatrix} \mathbf{e}'_1 \\ \vdots \\ \mathbf{e}'_N \end{bmatrix}. \tag{3.35}$$

Quality check

At the end of the ear shape generation process, meshes were derived from the point clouds as in the case of the FAST dataset (see Section 3.1.1). We then verified that the meshes were not aberrant and that they were fit for numerical simulation: any mesh that presented at least one self-intersecting face was left out.

In total, 24 % (320 out of 1325) of the meshes were discarded. Performing the Royston's multivariate normality test on the 1325 randomly drawn ear PC weights then on the 1005 remaining ones, we observed a decrease in the significance level of the test from 4.8 % to 0.8 %: it appears that the statistical distribution of the ear PC weights was somewhat degraded by the selection process. However, when looking into the distribution of each PC of the selected ear shapes separately (using the Shapiro-Wilk univariate normality test with a significance level of 5 %), we observe that the 9 rejected PCs account only for 3.7 % of the total variance.

For simplicity, we consider further on that N is the number of retained meshes i.e. $N = 1005$.

3.4.2 Numerical Simulations

Finally, PRTF sets were numerically simulated from the ear shapes of the new set E' according to the process described in Section 3.1.2

$$\mathbf{h}'_i = \psi\left(\mathbf{e}'_i\right), \quad \forall\ i = 1,\ \ldots\ N. \tag{3.36}$$

While virtually any number of pinna meshes could have been generated, the size of WiDESPREaD was limited by the computing resources required to calculate the PRTF sets from the pinnae. Indeed, calculating the $N = 1005$ PRTF sets from meshes of about 55000 triangular elements required a total of 40 days of 24 hours, on a workstation that features 12 CPU and 32 GB of RAM.

3.4.3 Visualization of the Augmented Dataset

By means of a visual review, we verified that the synthesized ear shapes and PRTF sets looked realistic. The first 10 pairs of ear shapes and PRTF sets of the WiDESPREaD dataset are displayed in Figure 3.26. We can see that the ear shapes are very diverse and that the PRTF sets vary accordingly.

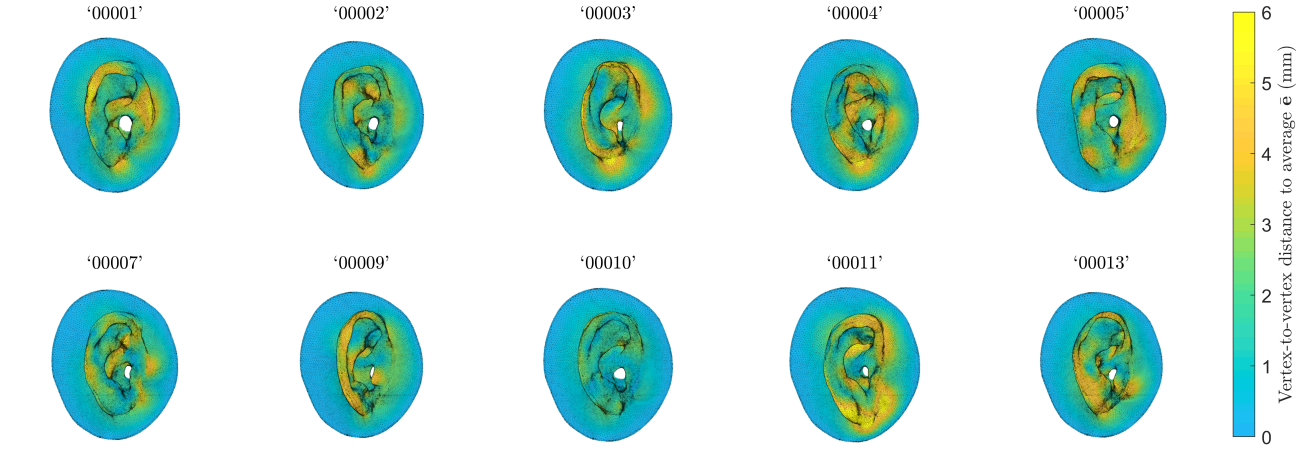
Conclusion

Our study of a joint dataset of 119 pinna ear meshes and matching simulated PRTF sets, FAST, resulted in our designing of what is, to the best of our knowledge, the first approach to HRTF data augmentation in a context of individualization.

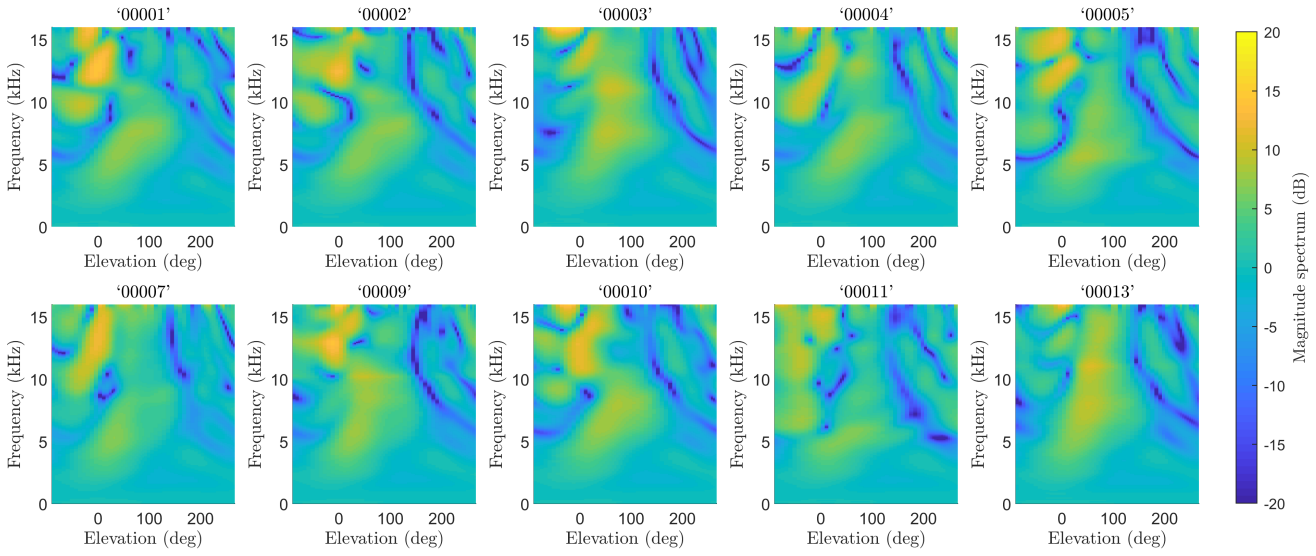
The resulting dataset, WiDESPREaD, is public and available online – kindly hosted by the ARI team on `sofacoustics.org`[6]. With its 1005 pairs of registered pinna meshes and corresponding computed PRTF sets, its is larger than any other currently available HRTF datasets by an order of magnitude. Its vastness opens up new possibilities regarding HRTF statistical modeling, user-friendly individualization and spatial interpolation from sparse measurements.

On another note, it is uniquely interesting for applications that rely on morphological data to provide individualized HRTFs. Indeed, it is the only HRTF dataset, to the best of our knowledge, that includes 3-D meshes that are registered. This fact makes it very easy to automatically extract various measurements from the meshes, a particularly interesting feature for the active field of user-friendly HRTF individualization based on anthropometry (see Chapter 2 for a review). It also facilitates linear and non-linear regressions between 3-D ear point clouds and PRTF sets.

[6] `https://sofacoustics.org/data/database/widespread/`



(a)



(b)

Figure 3.26 – Visualization of the first 10 artificial subjects of WiDESPREaD, designated by their ID. (a) Meshes derived from the synthetic ear shapes $\mathbf{e}'_1$, … $\mathbf{e}'_{10}$. Color represents the vertex-to-vertex euclidean distance to the generative model's average $\bar{\mathbf{e}}$. (b) Log-magnitude PRTF sets $20 \cdot \log_{10}(\mathbf{h}'_1)$, … $20 \cdot \log_{10}(\mathbf{h}'_{10})$ displayed in the median sagittal plane.

3.5 Dimensionality Reduction of the Augmented PRTF Dataset

In this section, we investigate how using the augmented dataset, WiDESPREaD, to train a PCA model of log-magnitude PRTF sets impacts dimensionality reduction performance. We start by comparing its CPV with that of 10 PCA models from Section 3.2, trained on various HRTF datasets including FAST. Going further, we then compare the results of 20-fold cross-validations performed respectively on the FAST and WiDESPREaD PCA models.

3.5.1 Cumulative Percentage of Total Variation

Pre-processing and PCA of the WiDESPREaD log-magnitude PRTF sets was performed as for the other HRTF datasets (see Section 3.2). Let us look into its CPV, as we have done for other PCA models throughout this chapter, and compare it with that of other HRTF datasets.

Comparison with FAST

In particular, let us compare the CPV of the WiDESPREaD model with that of FAST, dataset from which it derives. CPVs of both log-magnitude PRTF PCA models are plotted in Figure 3.27.

A first observation that can be made is that, for equal numbers of retained PCs, the FAST CPV is lower than the WiDESPREaD one. In other words, to achieve a given CPV, the WiDESPREaD model requires more PCs than the FAST one. For instance, to retain 90 % of the variability, $p_{90} = 321$ PCs are required for the former, against $p_{90} = 72$ for the latter (see Figure 3.27). In that sense, this could be seen as a regression: more PCs are needed to achieve a CPV of 90 %.

Yet, it actually corroborates our choice of augmenting the FAST dataset. Indeed, the aim of our dataset augmentation method was to produce a large yet statistically realistic population of PRTF sets, by using the more PCA-compatible ear shape space as a back door. According to the aforementioned observation, the WiDESPREaD PCA model has captured variations in magnitude PRTF sets that were not present in the initial dataset. If that was not the case, using the space of ear shapes to generate new data by means of PCA would have had little interest.

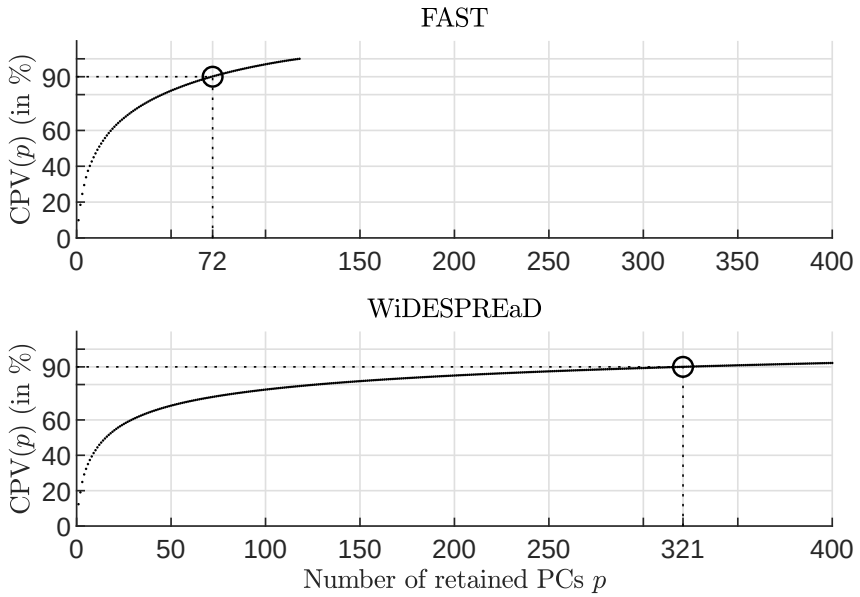


Figure 3.27 – Cumulative percentage of total variation (CPV) of log-magnitude PRTF PCA models as a function of number of retained principal components. Please note that the WiDESPREaD curve is plotted for a limited range of $p \in \{0, \dots 400\}$, due to the large difference in number of PCs between models. Top: FAST. Bottom: WiDESPREaD.

Furthermore, although the p_{90} of the WiDESPREaD model is larger than that of the FAST one in absolute terms, it is much smaller relatively to the maximum number of PCs. Indeed, for WiDESPREaD the $\frac{p_{90}}{N-1}$ ratio is $\frac{321}{1004} = 32$ %, whereas for FAST it is $\frac{p_{90}}{n-1} = \frac{72}{118} = 61$ %. This can be interpreted as the larger dataset having more redundancy, thus enabling PCA to store a same ratio of total information into less components relatively to the number of training examples.

Comparison with other datasets

In order to replace these observations in a more general context, we herein extend our Section 3.2 study to WiDESPREaD PRTFs. We thus compare the p_{90} – the number of PCs required to reach a CPV of 90 % – of the WiDESPREaD log-magnitude PRTF PCA model with those of 10 other models trained respectively on FAST and 9 public HRTF datasets. To this end, similarly to Figure 3.14, we present in Figure 3.28 a scatter plot of p_{90} as a function of the size of the dataset. Seeing that WiDESPREaD is much larger than any other dataset, we also include models trained on randomly drawn subsets of WiDESPREaD of sizes ranging from 100 to 800, denoted WiDESPREaD$_{100}$, ... WiDESPREaD$_{800}$.

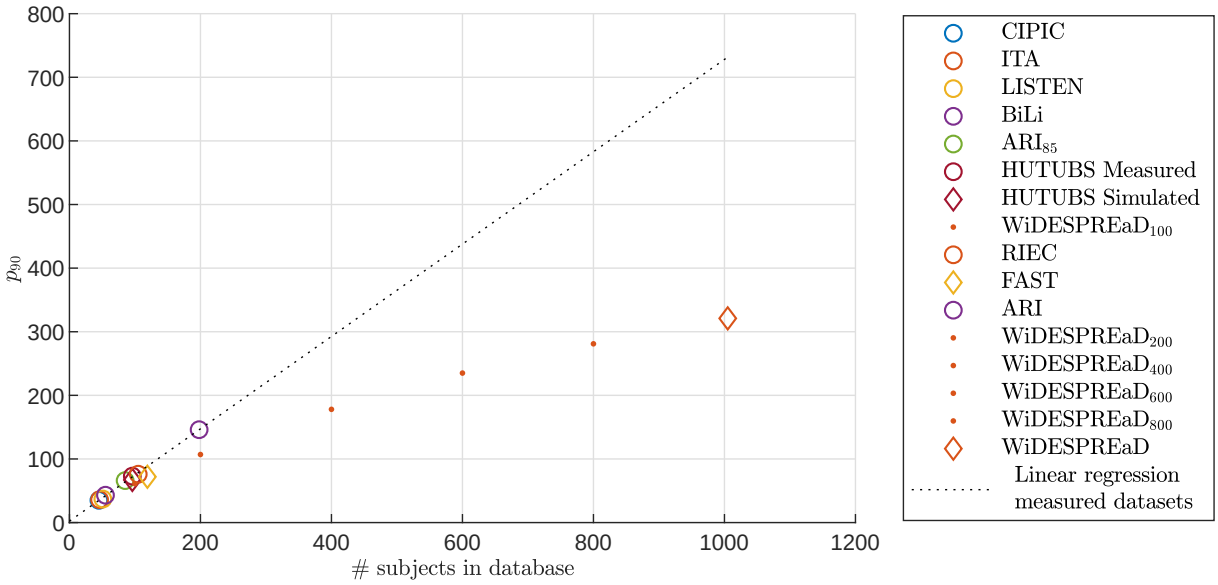


Figure 3.28 – Number of PCs p_{90} required to reach a CPV of 90 % for WiDESPREaD, 5 WiDESPREaD subsets, FAST and 9 public HRTF datasets, displayed as a function of the number of subjects. Circles: acoustically measured datasets. Diamonds: numerically simulated datasets. Dots: WiDESPREaD subsets.

We can see in Figure 3.28 that WiDESPREaD's p_{90} falls largely below the linear trend followed by the smaller datasets: the p_{90} (301) is worth less than half the linear prediction (731). As hoped, augmenting the FAST dataset has allowed us to reach a number of subjects high enough to observe a slowdown in p_{90}'s increase.

Overall, studying the CPV of WiDESPREaD and comparing it to that of FAST and other HRTF datasets has given us indications that the WiDESPREaD model may perform better at representing log-magnitude PRTF sets in general.

3.5.2 Cross-Validation

In order to assess and compare the capacity of the WiDESPREaD and FAST PCA models to generalize to new examples, we performed a 20-fold cross-validation on each one of them.

Method

Let us denote $\mathbf{g}_1^{(\mathrm{F})}, \ \ldots \ \mathbf{g}_n^{(\mathrm{F})} \in \mathbb{R}^{n_f n_d}$, and $\mathbf{g}_1^{(\mathrm{W})}, \ \ldots \ \mathbf{g}_N^{(\mathrm{W})} \in \mathbb{R}^{n_f n_d}$ the log-magnitude PRTF sets from the FAST and WiDESPREaD datasets, respectively. Additionally, let

there be $N_{\mathrm{F}} = n$ and $N_{\mathrm{W}} = N$.

Each dataset was equally divided into $K = 20$ sub-groups, each containing about 5 % of the subjects. Each sub-group of index $k = 1, \ldots K$ was then used in turn as a validation set for a PCA model trained the subjects of the remaining $K - 1$ folds.

For all $k = 1, \ldots K$ and for all dataset $S \in \{\mathrm{F}, \mathrm{W}\}$, let there be $I_{S_{\mathrm{train}}(k)} \subset \{1, \ldots N_S\}$ and $I_{S_{\mathrm{val}}(k)} \subset \{1, \ldots N_S\}$ the sets of subject indices that constitute the k^{th} fold's training and validation sets, respectively. For every fold, the number of training subjects is thus $N'_S = (K-1)\left\lfloor \frac{N_S}{K} \right\rfloor$, which is worth $N'_{\mathrm{W}} = 950$ for WiDESPREaD and $N'_{\mathrm{F}} = 95$ for FAST.

Let there be a fold $k = 1, \ldots K$ and a dataset $S \in \{\mathrm{F}, \mathrm{W}\}$. PCA was performed on the data matrix $\mathbf{X}_{S_{\mathrm{train}}(k)} = \left(\mathbf{g}_i^{(S)}\right)_{i \in I_{S_{\mathrm{train}}(k)}}$. Re-writing Equation (3.17) using this notation, the PCA transform can be written:

$$\mathbf{Y}_{S_{\mathrm{train}}(k)} = \left(\mathbf{X}_{S_{\mathrm{train}}(k)} - \bar{\mathbf{X}}_{S_{\mathrm{train}}(k)}\right) \mathbf{U}_{S_{\mathrm{train}}(k)}{}^{\mathrm{t}}. \tag{3.37}$$

Examples from the validation set $\mathbf{X}_{S_{\mathrm{val}}(k)} = \left(\mathbf{g}_{S_i}\right)_{i \in I_{S_{\mathrm{val}}(k)}}$ were then projected in the training space as follows:

$$\mathbf{Y}_{S_{\mathrm{val}}(k)} = \left(\mathbf{X}_{S_{\mathrm{val}}(k)} - \bar{\mathbf{X}}_{S_{\mathrm{train}}(k)}\right) \mathbf{U}_{S_{\mathrm{train}}(k)}{}^{\mathrm{t}}. \tag{3.38}$$

Finally, the training and validation data matrices were reconstructed from the PC weights. The number of PCs retained for reconstruction, m, varied in $\{0, \ldots N'_S\}$. Thus, using the same notation as in Equation (3.23) and according to (3.24), training and validation sets were reconstructed according to the following equations:

$$\tilde{\mathbf{X}}^{(p)}_{S_{\mathrm{train}}(k)} = \tilde{\mathbf{Y}}^{(p)}_{S_{\mathrm{train}}(k)} \mathbf{U}_{S_{\mathrm{train}}(k)} + \bar{\mathbf{X}}_{S_{\mathrm{train}}(k)}, \tag{3.39}$$

and

$$\tilde{\mathbf{X}}^{(p)}_{S_{\mathrm{val}}(k)} = \tilde{\mathbf{Y}}^{(p)}_{S_{\mathrm{val}}(k)} \mathbf{U}_{S_{\mathrm{train}}(k)} + \bar{\mathbf{X}}_{S_{\mathrm{train}}(k)}. \tag{3.40}$$

The MSE reconstruction error was then averaged across all folds for both training sets

$$\varepsilon_{S_{\mathrm{train}}}(p) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{MSE}\left(\tilde{\mathbf{X}}^{(p)}_{S_{\mathrm{train}}(k)}, \mathbf{X}_{S_{\mathrm{train}}(k)}\right), \tag{3.41}$$

and validation sets

$$\varepsilon_{S_{\text{val}}}(p) = \frac{1}{K}\sum_{k=1}^{K} \text{MSE}\left(\tilde{\mathbf{X}}^{(p)}_{S_{\text{val}}(k)}, \mathbf{X}_{S_{\text{train}}(k)}\right). \tag{3.42}$$

Results

The training and validation reconstruction errors for both FAST and WiDESPREaD PCA models are shown in Figure 3.29.

In either case, we observe a decreasing mean-square training error $\varepsilon_{S_{\text{train}}}(p)$ which becomes null when all PCs are retained, which ensues from the definition of PCA.

When looking at the cross-validation errors, a first observation that can be made is that, when all principal components are retained, the WiDESPREaD error ($\varepsilon_{\text{W}_{\text{val}}}(N'_{\text{W}} - 1) = 2.3$ dB^2) is much lower than the FAST one ($\varepsilon_{\text{F}_{\text{val}}}(N'_{\text{F}} - 1) = 6.0$ dB^2) [7] . This could be expected, seeing that WiDESPREaD includes about 8 times more examples of the same type of data than FAST. Indeed, approximating new data thanks to a PCA model with all PCs retained is equivalent to a projection into the $(N'_S - 1)$-dimensional space generated by linear combinations of the N'_S training examples.

More importantly, we can see that for any number of retained components $p = 0, \ldots N'_{\text{F}} - 1$, the WiDESPREaD cross-validation error is lower than that of the FAST model: $\varepsilon_{\text{W}_{\text{val}}}(p) \leq \varepsilon_{\text{F}_{\text{val}}}(p)$.

Let us imagine that we choose to retain p_{90} PCs – a typical way of choosing how many PCs to retain (see Section 3.2 and [Jolliffe02, Chap. 6, Sec. 1]). Doing so for each model, we would obtain for WiDESPREaD and FAST, respectively, average generalization errors of 2.84 dB^2 and 6.3 dB^2, for values of p_{90} of 312 and 60. In that context, the reduced WiDESPREaD model generalizes much better than the FAST one to new examples.

However, the WiDESPREaD model with p_{90} PCs thus holds 312 coefficients, which may still be a lot for certain applications – the tuning of an HRTF model's parameters by the listener for instance (see Chapter 4). As a consequence, we may want to choose an arbitrarily low number of PCs. For $p = 10$, for instance, the average generalization errors for the WiDESPREaD and FAST models would be 7.7 dB^2 and 8.6 dB^2, respectively. Hence, using the WiDESPREaD model would be an improvement over the FAST one in this context as well.

Finally, it is worth noting that only 35 components (out of 949) are needed for the

[7] This is not visible in Figure 3.29 as we limited the x-axis range for both models to be on a comparable range.

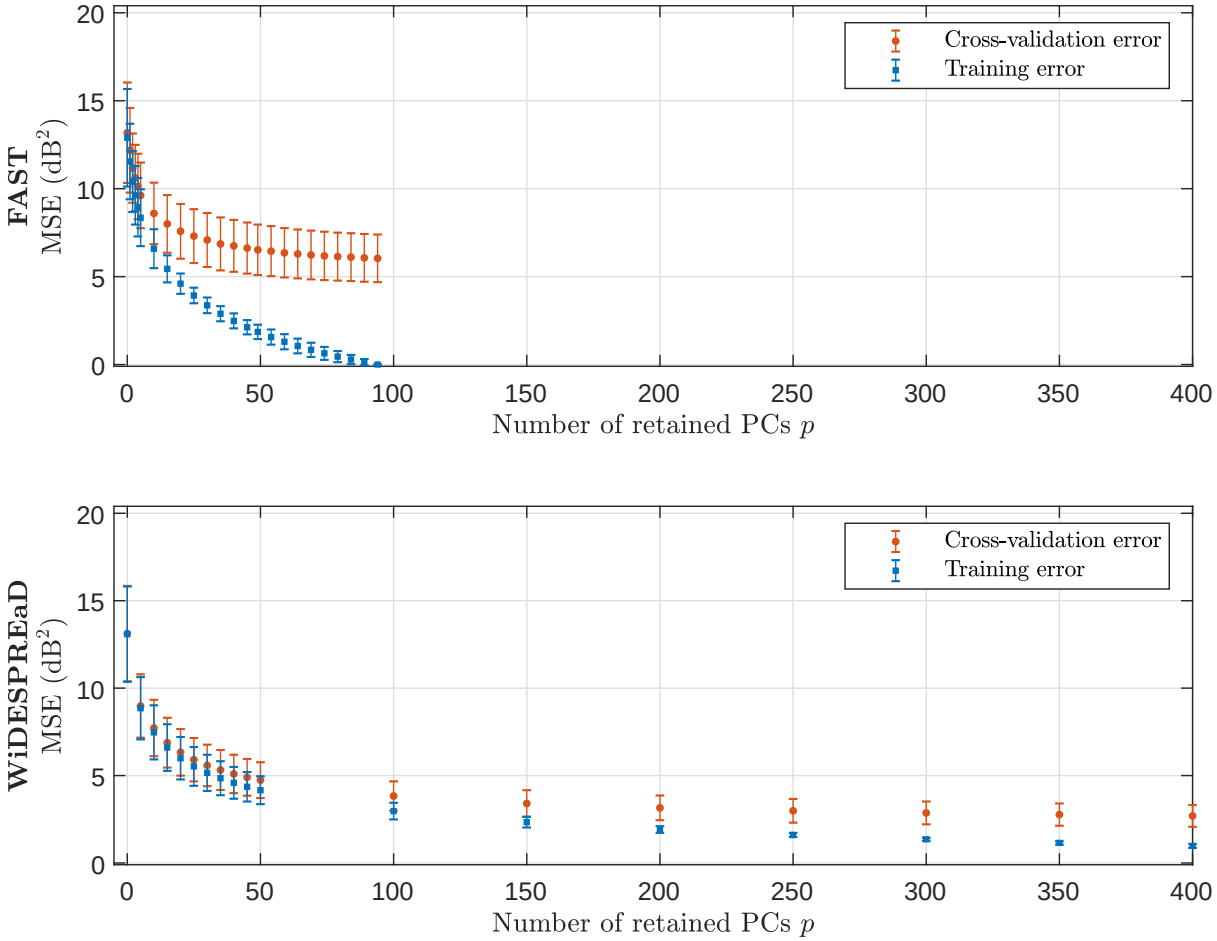


Figure 3.29 – Training MSE (blue) and cross-validation (red) MSE for various numbers of retained PCs p, for both FAST (top) and WiDESPREaD (bottom) datasets. While the curves represent the average errors across subjects and folds, the error bars stand for the standard deviation.
Please note that the WiDESPREaD curves are plotted for a limited range of $p \in \{0, \dots 400\}$, due to the large difference in number of subjects between models.

WiDESPREaD cross-validation reconstruction error to subceed the lowest cross-validation error ever attained in the case of the FAST dataset ($\varepsilon_{\mathrm{F}_{\mathrm{val}}}(N'_{\mathrm{F}}-1) = 6.0\ \mathrm{dB}^2$), that is with 94 retained PCs (out of 94).

Conclusion

In this section, in order to investigate WiDESPREaD's potential for PRTF dimensionality reduction, we have performed PCA on its log-magnitude PRTFs and compared its dimensionality reduction performance with that of other PCA models of log-magnitude HRTFs.

By comparing the CPV of the WiDESPREaD model with that of 10 log-magnitude HRTF PCA models (previously studied in Section 3.2), we corroborated our choice of augmenting the FAST dataset. Indeed, we seem to have sufficiently increased the number of subject for PCA to be able to compress more observed variability into fewer PCs. WiDESPREaD is thus the only dataset of the 10 models under study that is able to clearly escape the linear trend observed with the smaller datasets.

These results suggest that the WiDESPREaD PCA model is more representative of log-magnitude PRTF sets in general. Thus, in order to confirm it, we performed 20-cross-validations of the FAST and WiDESPREaD models. We find that, indeed, much better generalization is obtained with the WiDESPREaD model, regardless of the number of retained PCs.

3.6 Conclusion & Perspectives

The contributions in this chapter are five-fold. First, we presented the constitution of a joint dataset of 119 3-D registered meshes of human pinnae and matching simulated PRTF sets. Second, choosing an *inter-individual* approach to the PCA of HRTFs – one that has barely been covered in the literature, we studied and compared the dimensionality reduction performance of PCA on log-magnitude HRTF sets from 9 datasets including FAST. This led us to the conclusion that current datasets are too small to be representative of log-magnitude HRTF sets in general. Third, focusing on the FAST dataset, we compared the dimensionality reduction performance of PCA on the ear point clouds and that of the corresponding log-magnitude PRTF sets. We found that PCA-based dimensionality reduction performed considerably better in the space of 3-D ear morphology. Fourth, based

on this result, we presented a data augmentation process that allows the generation of an arbitrarily large synthetic PRTF database by means of random ear shape generations and FM-BEM calculations. The resulting dataset of 1005 ear meshes and matching PRTF sets, named WiDESPREaD, is public and freely available online[8]. Fifth and finally, we compared the dimensionality reduction performance of PCA on log-magnitude PRTFs from WiDESPREaD with that of other HRTF datasets, both on training and test data. We found that the WiDESPREaD seems to generalize better to new data than any other HRTF PCA model under study. In particular, much better generalization is obtained with the WiDESPREaD model that with the FAST one, regardless of the number of retained PCs.

Increasing the number of PRTF sets by generating new data in the ear shape space, where linear modeling seems adequate, may allow us to better understand the complexity of the link between morphology and HRTFs, as well as improve supervised and unsupervised HRTF statistical modeling. In particular, non-linear machine-learning techniques such as neural networks can benefit from the scalability of this synthetic dataset generation, as they generally require a large amount of data. As it is, WiDESPREaD is the first database, to our knowledge, with over a thousand PRTF sets and matching registered ear meshes. Although PRTFs are not complete HRTFs, they include an important part of the information relevant to HRTF individualization and, as the dataset includes about 5 times more subjects than any available HRTF dataset, it has great potential to help develop and improve methods for HRTF modeling, dimensionality reduction and manifold learning, as well as spatial interpolation of sparsely measured HRTFs.

Going further, it would be interesting to look for a potential non-linear manifold among WiDESPREaD magnitude PRTF sets. For that purpose, non-linear machine learning techniques such as locally linear embedding or neural networks could be used. Indeed, thanks to its size, WiDESPREaD is more suitable such techniques than any other dataset.

The dataset augmentation process itself could be improved on several aspects. In particular, including the contributions of a head and torso is an indispensable next step, as it would allow us to produce complete HRTFs instead of PRTFs. This could be done by randomly generating head and torso meshes in parallel of the pinnae, combining them then numerically simulating the corresponding HRTF set. This would however considerably increase the computing cost. Another option is to approximate complete HRTF sets by including the acoustic filtering effect of the head and torso *a posteriori* into the PRTFs

[8] `https://www.sofacoustics.org/data/database/widespread`

by means of structural composition [Algazi01b].

On another note, there is the pending question of the validity of numerically simulated HRTFs (see Chapter 2). However, the simulation process being completely deterministic, any upgrade could be easily included in the dataset augmentation method.

Finally, our generative ear model is quite rudimentary and may be further improved either using a simple trick such as probabilistic PCA [Tipping99] or a more complex machine learning technique altogether, although our work suggests that PCA fares rather well on ear point clouds.

Chapter 4

Individualization of Head-Related Transfer Functions based on Perceptual Feedback

4.1 Introduction

In Chapter 2, Section 2.3, we established a state of the art of HRTF individualization techniques. In particular, we underlined how direct methods such as acoustic measurements and numerical simulations are ill-suited for an end-user application. On the contrary, we reported that indirect methods – either based on sparse morphological data or on perceptual feedback from the listener – are designed to be user-friendly. In this thesis, we focus on the second – and less-explored – kind of indirect methods: the ones based on perceptual feedback. Indeed, they have the advantage of relying on a perceptual assessment of the quality of the produced HRTF set throughout the individualization process. Furthermore, they require no specific equipment and can allow a trade-off between tuning time and perceptual quality.

As detailed in our state of the art, a popular approach among such methods is to select a best-fit non-individual HRTF set among a database [Seeber03; Iwaya06; Katz12] and/or to adapt a non-individual HRTF set so as to improve localization performance [Tan98; Middlebrooks00; Runkle00]. These methods are however rudimentary and cannot claim to embrace the full complexity of the inter-individual variations of HRTF sets. In contrast with these, a more ambitious alternative has been to synthesize an HRTF set by means of a statistical model whose parameters are tuned based on perceptual feedback [Shin08; Hwang08a; Hölzl14; Fink15; Yamamoto17].

In this chapter, we present and evaluate such a method, which consists in tuning the parameters of a PCA model of magnitude HRTF set based on the outcome of listening experiments. The parameters are optimized by means of a Nelder-Mead simplex algorithm,

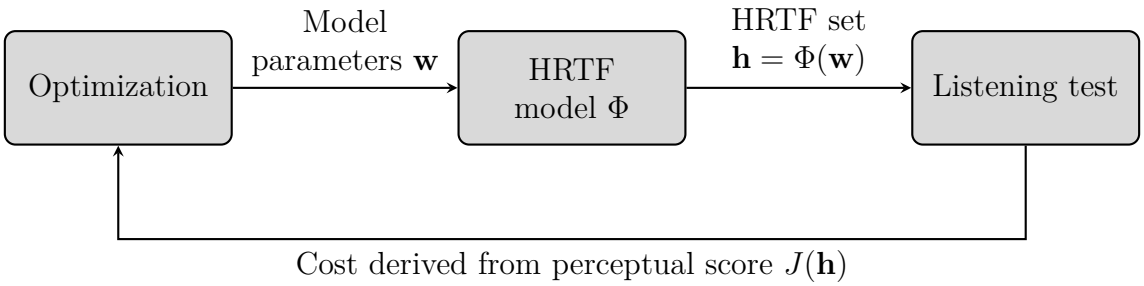


Figure 4.1 – General architecture of the HRTF tuning method.

based on a cost function directly derived from localization error.

Optimization Rather than letting the listener tune the model parameters himself as in [Shin08; Hwang08a; Hölzl14; Fink15], an optimization of the model parameters is performed by an algorithm – as in [Yamamoto17]. The listener is only prompted for subjective evaluation. While the former has the advantage of letting the listener decide what the best tuning duration/HRTF quality trade-off is, the latter gives us more control on the optimization scheme, seeing that it is performed by the algorithm instead of being entrusted to a human subject whose behavior is hardly predictable.

HRTF model In most similar work, the underlying model of HRTFs [Shin08; Hwang08a; Hölzl14; Fink15] is PCA-based. Interestingly, Yamamoto *et al.* [Yamamoto17] differed and used a variational autoencoder neural network to model the magnitude HRTFs. In the present work, we model the magnitude HRTFs by means of the *inter-individual* approach to PCA introduced in Chapter 2, Section 2.1.3 – which focuses on the inter-individual variations of magnitude HRTFs. Indeed, PCA was not performed in this fashion in any of the aforementioned PCA-based studies. Thus, for the same reasons as the ones invoked in Chapter 3, we use PCA to model magnitude HRTF sets, before potentially moving on to more complex unsupervised learning techniques.

With this particular way of performing PCA on HRTF data, a set of parameters (the PC weights) corresponds to a collection of magnitude HRTFs over the whole sphere – a mag-HRTF set as per the definition proposed in Chapter 1, Section 1.2.1. What is more, the PCs encode the inter-individual variations of HRTF sets. In our proposed method, the mag-HRTF set is thus tuned globally, as in [Hölzl14] or [Yamamoto17]. However, unlike Hölzl [Hölzl14], who used a SHD of the PC weights of a *spectral* PCA model of magnitude HRTFs, here the spatial patterns that underly our model's PCs were statistically inferred

from the training data.

Listening tests In the present work, the listening tests are localization tasks and the cost is derived from a localization error metric. Indeed, localization tasks allow for an absolute and quantitative rating of the perceptual quality of an HRTF set, as opposed to judgment tasks were an HRTF set is rated relatively to other HRTF sets according to a certain set of criteria (see Chapter 2, Section 2.2.2). This absolute character is particularly convenient in the present context: it allows to carry out independently the perceptual evaluation of every new HRTF set presented to the listener throughout the optimization process. Furthermore, according to Zagala *et al.* [Zagala20], localization performance appear to be also a good predictor of overall preference based on virtual sound trajectories.

The perceptual evaluations were restricted to directions of the median-plane, where ILD and ITD are almost null, allowing us to focus on monaural spectral auditory cues, which are the core problem in HRTF individualization (see Chapter 1).

The present chapter is laid out as follows. First, we detail the HRTF individualization method. Second, we present a preliminary experiment, in which the localization tasks were simulated by means of an auditory model. Third, the HRTF tuning method is evaluated in an actual listening experiment with 12 participants.

4.2 Method

The general architecture of the tuning algorithm is laid out in Figure 4.1. At each iteration, an HRTF set is generated by the HRTF model. Then, the HRTF set is presented to the listener for a listening test which yields a perceptual score. Based on that score, the optimization algorithm then updates the model's parameters.

4.2.1 HRTF Model

As mentioned above, the HRTF model used in our implementation was a PCA model trained in the *inter-individual* fashion, as in the work presented in Chapter 3. In this sub-section, we go over the process Φ of reconstructing a complex two-ear HRTF set $\mathbf{h} \in \mathbb{R}^{2n_f n_d}$ from a set of PC weights $\mathbf{w} \in \mathbb{R}^p$.

Let N be the number of training subjects and

$$\mathbf{X} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_N \end{bmatrix} \in \mathbb{R}^{N \times n_f n_d}$$

the training data matrix. The PCA transform is expressed according to Equation (3.37) (see Section 3.3 for more detail):

$$\mathbf{Y} = \left(\mathbf{X} - \bar{\mathbf{X}}\right) \mathbf{U}^{\mathrm{t}}, \tag{4.1}$$

where $\mathbf{U} \in \mathbb{R}^{(N-1) \times n_f n_d}$ is the transform matrix. The rows of $\mathbf{U}$ are the $(N-1)$ eigen vectors $\mathbf{u}_1, \ \ldots \ \mathbf{u}_{N-1} \in \mathbb{R}^{n_f n_d}$ that correspond to the PCs:

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{N-1}. \end{bmatrix} \tag{4.2}$$

Log-magnitude HRTF set

Conversely, let there be a row vector $\mathbf{w} \in \mathbb{R}^p$ of weights for the first $p \in \{0, \ \ldots \ N-1\}$ PCs. The corresponding log-mag-HRTF set $\mathbf{g} \in \mathbb{R}^{n_f n_d}$ is reconstructed as follows

$$\mathbf{g} = \mathbf{w}\tilde{\mathbf{U}}^{(p)} + \bar{\mathbf{g}}, \tag{4.3}$$

where

$$\tilde{\mathbf{U}}^{(p)} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \end{bmatrix} \tag{4.4}$$

is the sub-set of the transform matrix $\mathbf{U}$ that corresponds to the p first eigen vectors.

Complex two-ear PRTF set

By construction, $\mathbf{g}$ contains left-ear log-magnitude HRTFs $G_{\mathrm{dB}}^{(\mathrm{L})}(f_i, \theta_j, \varphi_j)$ for all frequencies $f_i = f_1, \ \ldots \ f_{n_f}$ and all directions $(\theta_j, \varphi_j) = (\theta_1, \varphi_1), \ \ldots \ (\theta_{n_d}, \varphi_{n_d})$:

$$\mathbf{g} = \left[G_{\mathrm{dB}}^{(\mathrm{L})}(f_1, \theta_1, \varphi_1) \ \ldots \ G_{\mathrm{dB}}^{(\mathrm{L})}(f_1, \theta_{n_d}, \varphi_{n_d}) \ \ldots \ G_{\mathrm{dB}}^{(\mathrm{L})}(f_{n_f}, \theta_1, \varphi_1) \ \ldots \ G_{\mathrm{dB}}^{(\mathrm{L})}(f_{n_f}, \theta_{n_d}, \varphi_{n_d})\right]. \tag{4.5}$$

Corresponding left-ear minimal phase HRTFs $H^{(\mathrm{L})}$ were obtained by deriving minimal phase filters from the magnitude spectra

$$H^{(\mathrm{L})}(f,\theta,\varphi) = G^{(\mathrm{L})}(f,\theta,\varphi) \cdot \exp\left[j\mathcal{H}\left(-\ln\left(G^{(\mathrm{L})}(f,\theta,\varphi)\right)\right)\right], \tag{4.6}$$

where $G^{(\mathrm{L})}(f,\theta,\varphi) = 10^{G^{(\mathrm{L})}_{\mathrm{dB}}(f,\theta,\varphi)}$ is the linear magnitude.

The left-ear HRTFs were then mirrored with regard to the median plane to constitute right-ear HRTFs

$$H^{(\mathrm{R})}(f,\theta,\varphi) = H^{(\mathrm{L})}(f,-\theta,\varphi). \tag{4.7}$$

Although in a more general context ITD would need to be tuned along with the magnitude HRTF model and the corresponding TOAs combined with the minimum-phase filters, it is irrelevant here, in the case of median-plane localization tests – where the ITD is close to zero.

Overall, $\Phi(\mathbf{w}) = \mathbf{h} \in \mathbb{C}^{2n_f n_d}$, with

$$\begin{aligned}\mathbf{h} = [&H^{(\mathrm{L})}(f_1,\theta_1,\varphi_1)\ \ldots\ H^{(\mathrm{L})}(f_1,\theta_{n_d},\varphi_{n_d})\ \ldots\ H^{(\mathrm{L})}(f_{n_f},\theta_1,\varphi_1)\ \ldots\ H^{(\mathrm{L})}(f_{n_f},\theta_{n_d},\varphi_{n_d})\ \ldots\\ &H^{(\mathrm{R})}(f_1,\theta_1,\varphi_1)\ \ldots\ H^{(\mathrm{R})}(f_1,\theta_{n_d},\varphi_{n_d})\ \ldots\ H^{(\mathrm{R})}(f_{n_f},\theta_1,\varphi_1)\ \ldots\ H^{(\mathrm{R})}(f_{n_f},\theta_{n_d},\varphi_{n_d})].\end{aligned} \tag{4.8}$$

4.2.2 Cost Function

The tuning process can be formulated as an optimization problem, where we seek to minimize a localization-error-based cost function J:

$$\tilde{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmin}}\left[J(\Phi(\mathbf{w}))\right], \tag{4.9}$$

where $p \in \mathbb{N}^*$ is the number of model parameters.

The cost function was composed of two components J_{loc} and J_{reg}:

$$J = J_{\mathrm{loc}} + J_{\mathrm{reg}}. \tag{4.10}$$

Localization error cost

The former, J_{loc}, is directly related to the localization error.

For the present application, we use the *absolute polar error* (APE) which is the ex-

pectation of the absolute error in elevation

$$\varepsilon(\mathbf{h}) = \frac{1}{N_\varphi} \sum_{k=1}^{N_\varphi} \sum_{l=1}^{N_\varphi} P_{\mathbf{h}}(\varphi_k \mid \varphi_l) \cdot |\varphi_k - \varphi_l|. \tag{4.11}$$

The cost was then computed by normalizing the APE, by dividing it by the APE that would be observed for random answers $\varepsilon_{\text{chance}}$

$$J_{\text{loc}}(\mathbf{w}) = \frac{\varepsilon\left(\mathbf{h}\right)}{\varepsilon_{\text{chance}}} = \frac{\varepsilon\left(\Phi(\mathbf{w})\right)}{\varepsilon_{\text{chance}}}. \tag{4.12}$$

Regularization cost

The second term, J_{reg}, is a regularization cost that encourages the PCWs to be in a "plausible" range, i.e. that discourages extreme values.

We based the cost on a multivariate normal probability density function whose mean in the null vector and whose covariance matrix is the diagonal matrix composed of the variances associated with each PC $\mathbf{\Sigma}^2_{G'}$, multiplied by a factor $\alpha \in \mathbb{R}^+$, used to control the harshness of the constraint.

The probability density is then normalized by its maximum value, i.e. its value in the null vector.

$$J_{\text{reg}}(\mathbf{w}) = 1 - \frac{\rho_{\mathbf{0},(\alpha\mathbf{\Sigma}_{G'})^2}(\mathbf{w})}{\rho_{\mathbf{0},\mathbf{\Sigma}^2_{G'}}(\mathbf{0})}, \tag{4.13}$$

where $\rho_{\boldsymbol{\mu},\mathbf{\Sigma}^2} : \mathbb{R}^p \mapsto [0,1]$ designates the multivariate probability density function of mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{\Sigma}^2 \in \mathbb{R}^{p\times p}$, defined by

$$\rho_{\boldsymbol{\mu},\mathbf{\Sigma}^2}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{t}}\mathbf{\Sigma}^{-2}(\mathbf{x}-\boldsymbol{\mu})\right]. \tag{4.14}$$

In the following experiments, α was tuned manually to 6.

4.2.3 Optimization Algorithm

To solve the optimization problem, we used the Nelder-Mead simplex method [Nelder65].

This general-purpose approach is appropriate to the present case, where the cost function is provided by a black box system, that is a human subject participating in

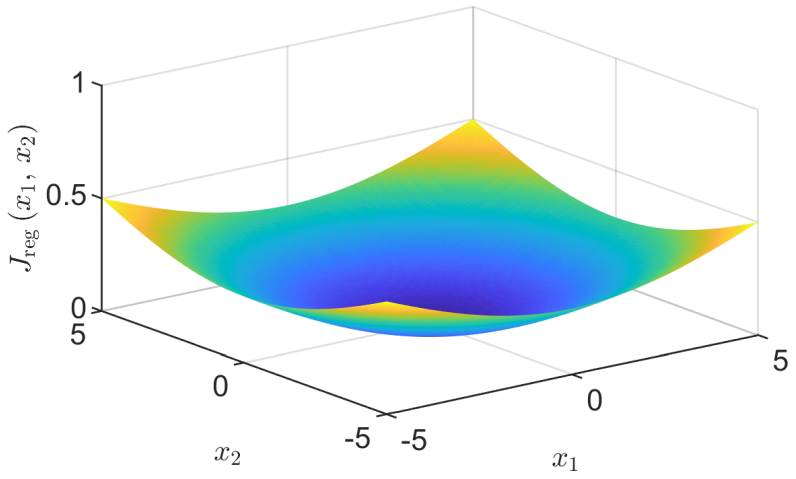


Figure 4.2 – Regularization cost J_{reg} in two dimensions ($p = 2$) for $\alpha = 6$ and $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

a localization experiment. Indeed, the Nelder-Mead algorithm is aimed at minimizing a scalar-valued non-linear cost function of $\mathbb{R}^p$ without any derivative information, explicit or implicit.

Furthermore, according to Lagarias *et al.* [Lagarias98], the method is parsimonious in cost function evaluations, a desirable trait in our case where limiting the number of subjective evaluations is desirable in order to limit the duration of the tuning procedure.

Initialization The optimization process was initiated with PC weights set to zero, which corresponds to the average log-magnitude HRTF set $\bar{\mathbf{g}}$.

Convergence The optimization process was considered to have converged when the absolute difference of two subsequent evaluations of the cost function subceeded a lower bound of 10^{-3}:

$$|J(\mathbf{w}[n+1]) - J(\mathbf{w}[n])| < 10^{-3}, \tag{4.15}$$

where $n \in \mathbb{N}^+$ denotes the iteration. If that criterion was not reached before, the process stopped at 500 iterations. These parameters were tuned manually after a number of trials and errors.

4.3 Simulated Listening Tests

4.3.1 Auditory Model

To simulate localization tasks, we used the Baumgartner auditory model for median-plane localization [Baumgartner14], described in more details in Section 2.2.3 of Chapter 2. Given two sets of median-plane HRTFs, e.g. the listener's own $\mathbf{h}_0$ and the one listened to $\mathbf{h}$, the model outputs a map of response probabilities. The result is a probability mass vector (PMV). This PMV contains, for all elevations φ and φ_{req}, the probability that the listener's answer is φ given that the requested elevation is φ_{req}. We denote this probability $P_{\mathbf{h}}(\varphi|\varphi_{\text{req}})$.

The code, included in the freely available Auditory Modeling Toolbox[1], also includes tools to compute common localization error metrics from the probabilities, such as the quadrant error (QE) and polar error (PE) presented in Section 2.2.2 of Chapter 2.

4.3.2 Configurations

Several configurations of the tuning method were explored.

Datasets

Three of the HRTF datasets studied in Chapter 3, WiDESPREaD, FAST and ARI, were used in turn to build the model of HRTF magnitudes. Each time, approximately 95% of the log-magnitude HRTF sets were used to train the PCA model. The remaining 5% were then used as targets for the tuning process.

The WiDESPREaD dataset was chosen because of its large number of examples. As we have seen in Chapter 3, it allows the PCA model to generalize well compared to other datasets. As WiDESPREaD was generated by augmenting the FAST dataset, the latter is a good comparison point. Finally, the FAST and WiDESPREaD datasets are composed of synthetic PRTF sets, simulated from pinnae normalized in size. It thus seemed desirable to also perform the tuning procedure on a more conventional dataset, made of acoustically measured HRTFs. We chose ARI in particular for its size, the good spatial accuracy of its HRTF sets, and its popularity among the community.

[1] `http://amtoolbox.sourceforge.net/`

Number of principal components

Different numbers of tuning PCs were studied. The higher the number of parameters to be tuned by the Nelder-Mead optimization algorithm, the higher the number of evaluations of the cost function and thus of virtual localization experiments. With in mind the goal of simulating real localization experiments and thus to keep the tuning time as low as possible, the number of PCs was kept arbitrarily low. Hence, we tested the tuning procedure for 3, 5, 10, 20 and 40 retained PCs.

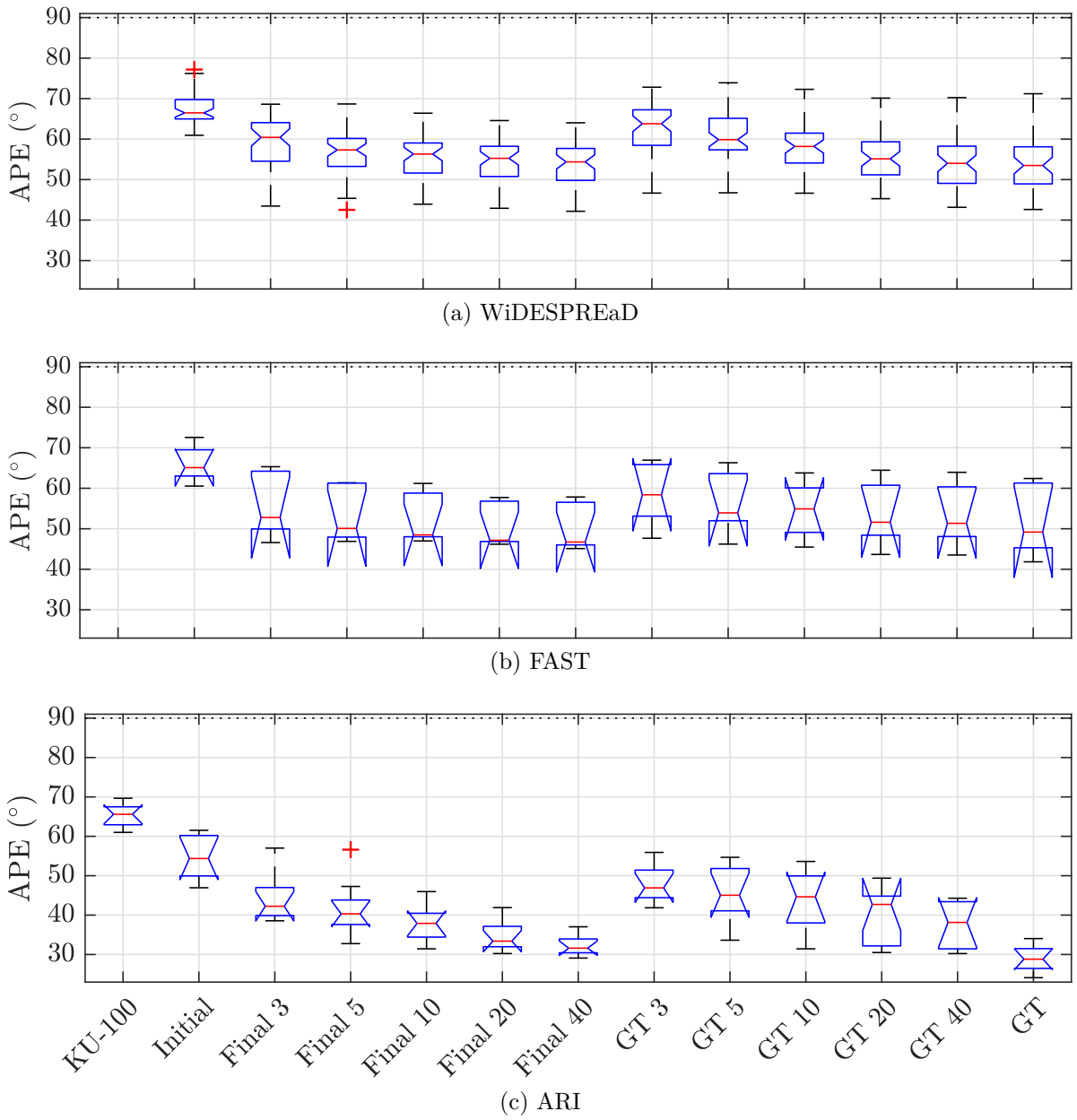
4.3.3 Results

In Figures 4.3, 4.4 and 4.5, we report the localization errors (the APEs, QEs and PEs, respectively) obtained at the beginning and at the end of the tuning process for each dataset and for the various numbers of PCs under test. In addition, we include a *ground truth* (GT) localization error which corresponds to the case where the "virtual listener" (VL) is presented with his own HRTF set. Finally, for each number of PCs under study, we also provide a *reduced ground truth* localization error which corresponds to the VL being presented with the approximation by the reduced PCA model of his own HRTF set. This log-magnitude HRTF set is also the best fit of the reduced PCA model to the target in terms of MSE. A baseline condition is included as well for the ARI dataset case: the HRTF set of a Neumann KU-100 manikin, commonly used in the literature to generate a generic non-individual VAS. For coherence with the ARI dataset, the KU-100 HRTF set measurement used in this work is the one made at the ARI as part of the Club Fritz project [Andreopoulou15]. Seeing that the baseline, initial and ground truth localization errors do not depend on the number of PCs, they are plotted only once. The results are reported as box plots in order to represent statistical variation across test subjects.

Additionally, an exemplary outcome of the optimization process is displayed in Figure 4.6. For ARI subject NH825 and $p = 20$ tuning PCs, the initial, final, reduced-GT and GT mag-HRTF sets are plotted for directions of the median plane, as well as the corresponding localization PMVs output by the Baumgartner model.

Ground truth – comparison with the literature

Before going on, let us compare the GT localization errors that we obtained in the ARI case to those reported in [Baumgartner14]. In that work, Baumgartner *et al.* used the auditory model to predict the localization performance of 23 listeners from the ARI dataset

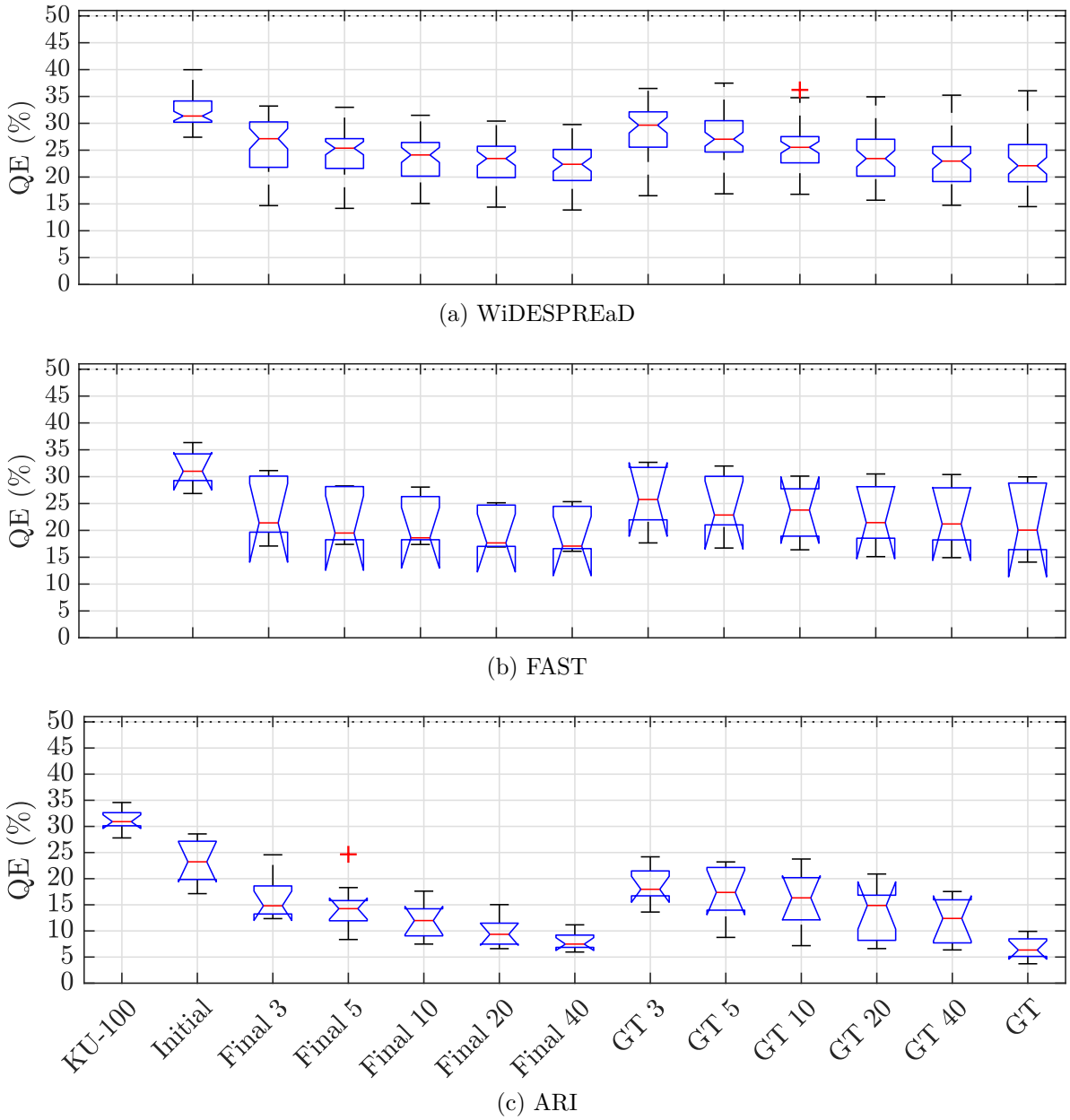


(a) WiDESPREaD

(b) FAST

(c) ARI

Figure 4.3 – Localization error outcome of the simulated tuning experiments: notched box plots of the APEs of the baseline (KU-100), initial, final (Final p), reduced ground truth (GT p) and ground truth (GT) for all numbers $p = 3, 5, 10, 20, 40$ of retained PCs. Each subplot corresponds to a dataset condition: WiDESPREaD (a), FAST (b) and ARI (c). The horizontal dotted line shows the localization error associated with random answers.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].

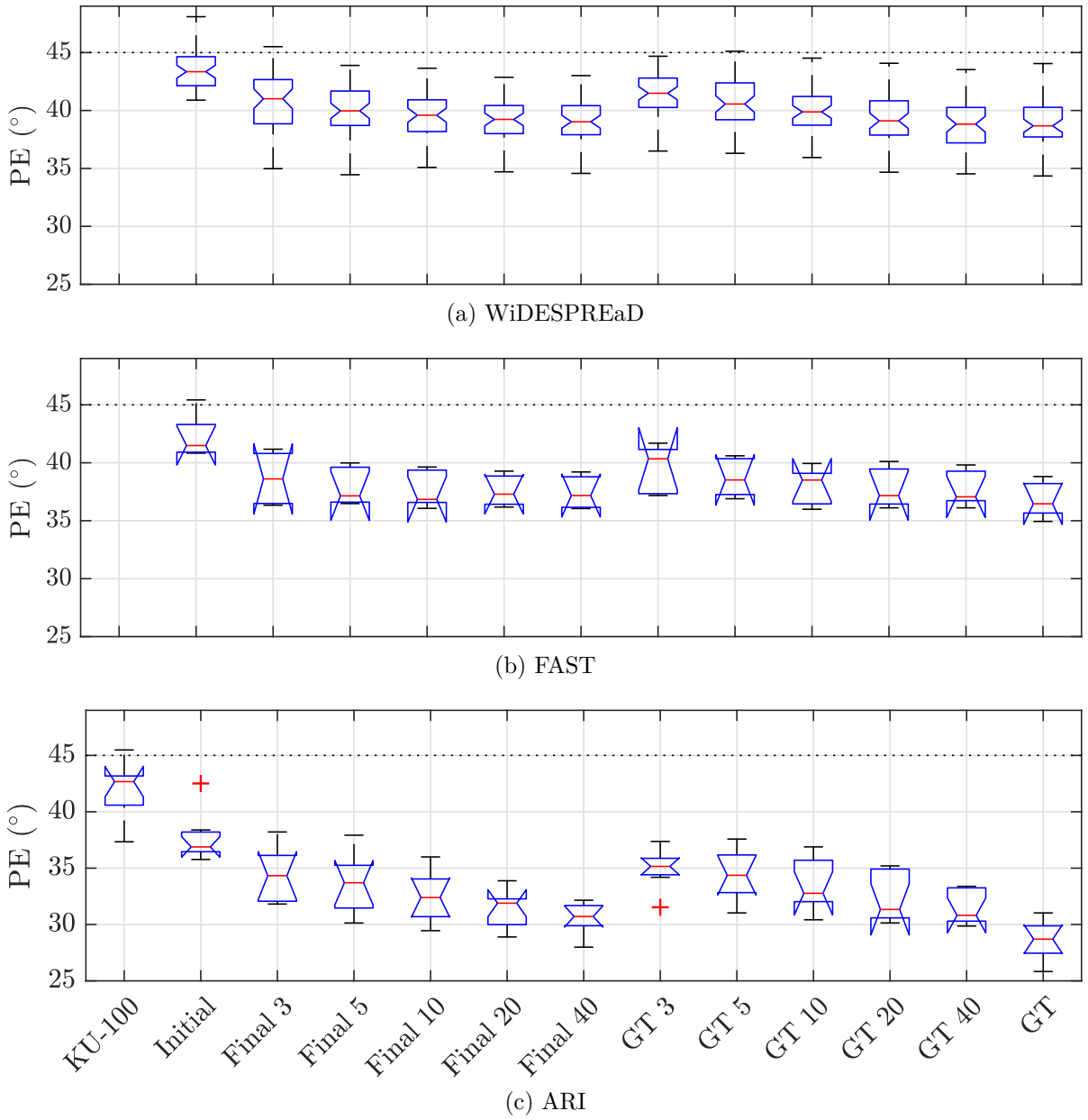


(a) WiDESPREaD

(b) FAST

(c) ARI

Figure 4.4 – Localization error outcome of the simulated tuning experiment: notched box plots of the QEs of the baseline (KU-100), initial, final (Final p), reduced ground truth (GT p) and ground truth (GT) for all numbers $p = 3, 5, 10, 20, 40$ of retained PCs. Each subplot corresponds to a dataset condition: WiDESPREaD (a), FAST (b) and ARI (c). The horizontal dotted line shows the localization error associated with random answers.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].



(a) WiDESPREaD

(b) FAST

(c) ARI

Figure 4.5 – Localization error outcome of the simulated tuning experiment: notched box plots of the PEs of the baseline (KU-100), initial, final (Final p), reduced ground truth (GT p) and ground truth (GT) for all numbers $p = 3, 5, 10, 20, 40$ of retained PCs. Each subplot corresponds to a dataset condition: WiDESPREaD (a), FAST (b) and ARI (c). The horizontal dotted line shows the localization error associated with random answers.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].

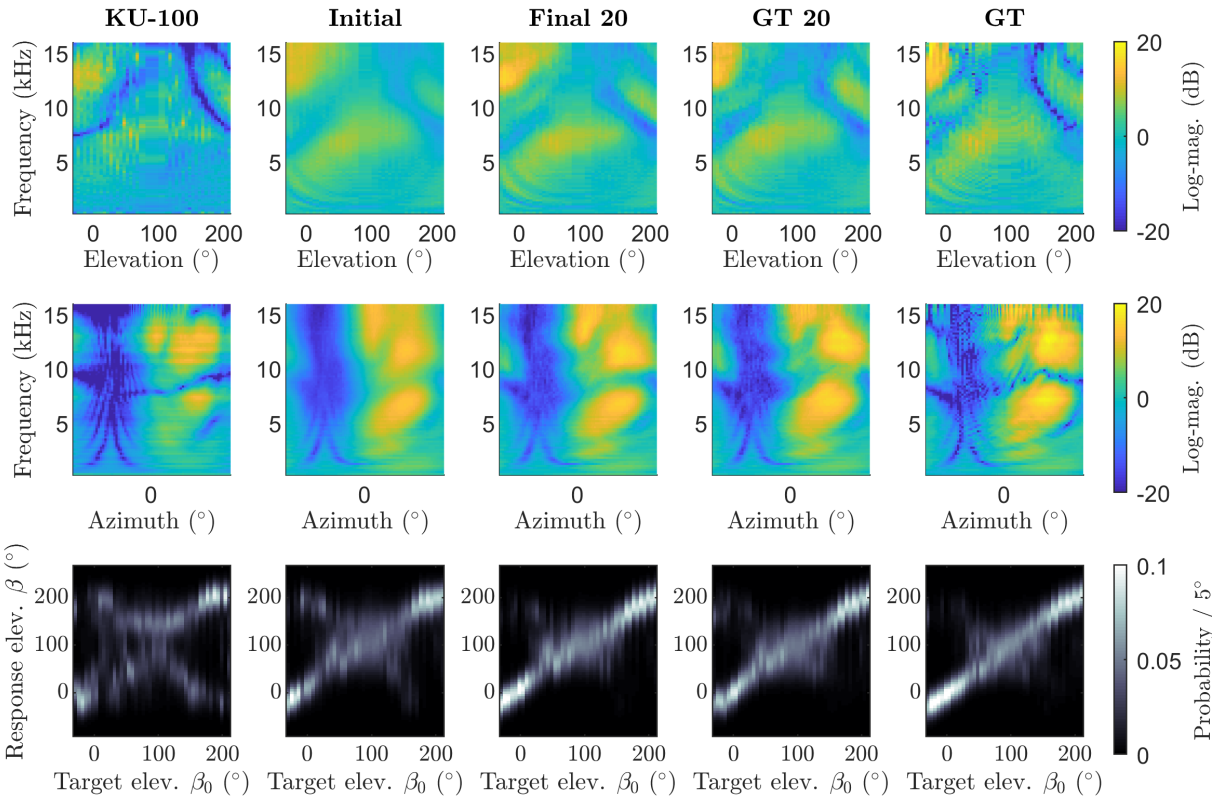


Figure 4.6 – Exemplary outcome of the optimization process, for subject NH825 of the ARI dataset and 20 tuning PCs. The baseline (KU-100), initial, final, reduced ground truth (GT 20) and ground truth (GT) magnitude HRTF sets are shown on the first and second rows, for directions in the median and horizontal planes, respectively. The corresponding PMVs are plotted on the second row. Matching APEs are 66.1°, 61.5°, 39.2°, 43.3° and 34.0°, respectively.

who were presented with their own HRTF sets, which they compare to the outcome of actual localization experiments.

We report in Figure 4.7, in the form of notched boxplots, the simulated and actual QEs and PEs from Table I of [Baumgartner14], alongside our own ARI GT QEs and PEs, simulated for a random ARI subset of 9 virtual listeners presented with their own HRTF set – also reported in subplot (c) of Figure 4.4 and Figure 4.5. Our own simulated median QE and PE (6.3 % and 29°, respectively) are somewhat lower than the simulated QE and PE from [Baumgartner14] (9.7 % and 32°). The difference in medians appears to be significant for local angular errors (PEs), but it is not the case for quadrant errors. Possible explanations for this modest mismatch include the fact that we used a fixed sensitivity parameter in the auditory model, while they tuned it for each individual. Also, we considered a different and smaller subset of the ARI dataset.

Compared to actual localization errors with individual HRTF sets found in the literature, our simulated GT for ARI virtual listeners is in rather good agreement. In a study by Middlebrooks [Middlebrooks99b, Figure 13], in which 11 listeners participated in actual localization experiments, the author reports a median QE of about 4 % and a median PE of about 27° with individual HRTF sets. In a similar study by Middlebrooks *et al.* [Middlebrooks00], the QEs for 5 listeners having listened once or twice to their own HRTF set (for a total of 9 cases) are reported in Figure 3 and correspond to a median QE of about 8 %. In [Baumgartner14], Baumgartner *et al.* report median QE and PE for the actual localization experiments of 9.6 % and 34°, respectively (see Figure 4.7). It is worth noting that the outcome of these experiments are, by construction, in excellent agreement with the aforementioned simulated localization errors from the same study: the sensitivity parameter of the auditory model had been tuned individually for each listener in order to fit the results of the actual experiments. Our median QE in the simulated ARI GT condition is comprised between the median QEs reported in [Middlebrooks99b] on the one hand, and [Middlebrooks99b] and [Baumgartner14] on the other hand, and our median PE is comparable to that of [Middlebrooks99b] and slightly lower than that of [Baumgartner14].

Differences between datasets

When comparing datasets, we can see that all localization errors are higher in the FAST and WiDESPREaD cases than in the ARI case. In particular, the median ground truth APEs are largely and significantly higher for FAST and WiDESPREaD (49° and 54°)

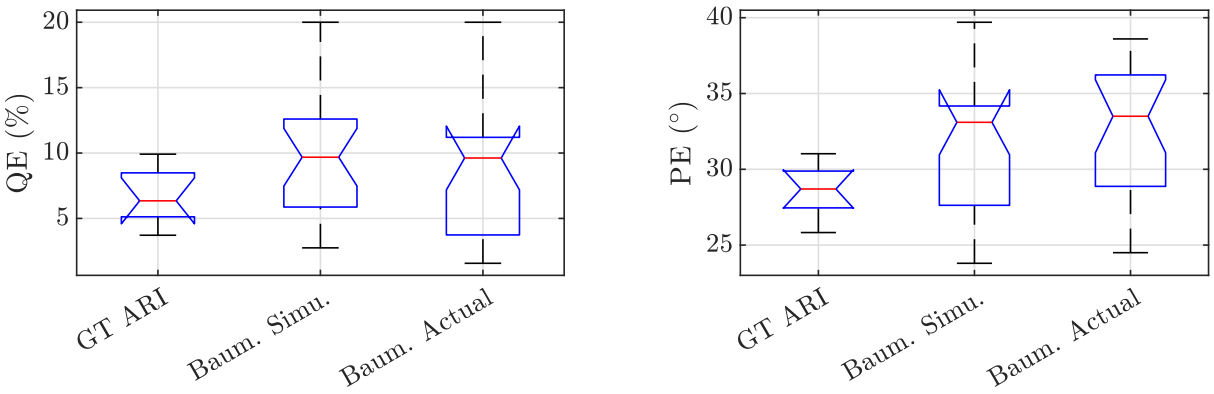


Figure 4.7 – Boxplots comparing the QEs (left) and PEs (right) that we simulated for 9 ARI virtual listeners with their own HRTF set (GT ARI) with the ones reported by Baumgartner *et al.* in [Baumgartner14] for both simulated (Baum. Simu.) and actual (Baum. Actual) localization experiments of 23 listeners.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].

than for ARI (29°). In terms of QEs, the median GT errors are of 20 % and 22 % for FAST and WiDESPREaD against 6.3 % for ARI, a significant difference of more than a factor 3. As discussed above, the latter is of the same order of magnitude (although somewhat lower) than QEs reported in [Baumgartner14] for both simulated and actual localization tasks with individual HRTF sets. In contrast, the GT simulated localization performances for FAST and WiDESPREaD are much poorer than the usually expected localization performance with individual HRTFs.

It would seem that the absence of head- and torso-related spectral features in PRTFs cause the Baumgartner model to yield considerably higher localization errors than what would be obtained in similar conditions with HRTFs, even when a PRTF set is designated as the internal template – i.e. the individual HRTF set – of the virtual listener.

Initial and baseline conditions

The initial median APE (54°) is notably lower than the baseline KU-100 one (66°), although not significantly so. It is somewhat surprising, seeing that the initial condition corresponds to the average log-magnitude HRTF set of the ARI dataset. Indeed, in such

an HRTF set the spectral features are smoothed and the peaks and notches useful to elevation perception and front-back disambiguation are shallower and less sharp than the ones found in a measured HRTF set, such as the KU-100 one (see Figure 4.14).

As we have seen above, simulated ARI GT localization tasks seem to yield localization errors that are in good agreement with the outcome of actual localization tasks with individual HRTFs found in the literature. In Figure 13 of [Middlebrooks99b], in addition to localization errors with individual HRTF sets, Middlebrooks reports the outcome of localization tasks with non-individual HRTF sets (those of other participants in the experiment): the median QE in the latter condition is about 19 % and the median PE about 41°. Similarly, in [Middlebrooks00], Middlebrooks *et al.* report in Figure 3 the QEs of 5 subjects having listened to the HRTF sets of one or two other participants, for a total of 9 non-individual conditions, and a median QE of about 33 %. Our initial condition, a non-individualized VAS based on an average HRTF set, yields simulated localization performance comparable to the first study with a median QE of 23 % and a median PE of 37°. In contrast, our baseline KU-100 condition, a generic non-individualized VAS based on the HRTF set of a manikin, results in significantly poorer simulated localization performance, with a median QE and a median PE of 31 % and 43°, respectively. This median QE is nevertheless comparable to that of the second study.

Regarding the FAST and WiDESPREaD datasets, the initial localization performances are much poorer than with the ARI dataset, with median QEs and PEs of 31 % and 41° for the former and 31 % and 43° for the latter, all significantly lower than the ARI initial median QE and PE.

Optimization outcome

General trends For all datasets, we observe that the tuning procedure significantly decreased the median APE and QE compared to initialization (training set's average log-magnitude HRTF set). The only exception occurred with FAST and 3 tuning PCs, in which case the standard deviation is very high, although the median is indeed lower than the initial APE by 19 %. In the case of the ARI HRTF model, for instance, the QE decreased in median from 23 % to between 7.5 % (for $p = 40$) and 15 % (for $p = 3$), depending on the number of PCs p – against a ground truth median QE of 6.3 %.

For all datasets, the localization errors – APE, QE and PE – tend to decrease with the number of PCs. The decrease is the most important between the *Initial* and *Final 3* conditions, and is significant in terms of APE and QE for the WiDESPREaD and ARI

datasets. The decrease gets however more modest when more PCs are retained. In particular, there seems to be a plateau for FAST and WiDESPREaD when the number of PCs exceeds 5.

Nevertheless, for all datasets, when at least $p = 20$ PCs are retained ($p \geq 10$ for WiDESPREaD, $p \geq 3$ for FAST), the difference between the median final APE (respectively QE) and the median ground truth APE (respectively QE) is not statistically significant. In the particular case of the FAST dataset, due the high variability of the localization error results in all conditions, the difference between the median *Final* p and ground truth APE, QE and PE is not significant for any $p \in \{3, 5, 10, 20, 40\}$.

Interestingly, the median final APE for a given p is generally lower than the APE of the corresponding projected GT – excepted for the WiDESPREaD dataset when $p \geq 20$, where the difference in median APEs is lower than 0.5°. This difference is not significant for any p or dataset. Nevertheless, it seems to exhibit some capacity of the optimization process to overcome – in terms of localization performance – the projection of the listener's own HRTF set in the space of the p first PCs.

Number of iterations and cost function evaluations

The number of iterations required to converge for all three datasets and all 5 numbers of PCs are reported in box plots in Figure 4.8. The corresponding number of evaluations of the cost function – i.e. the number of virtual localization tasks – are reported in the same fashion in Figure 4.9.

A first observation that we can make is that the number of iterations needed to converge is very consistent from one dataset to the other, for all numbers of PCs. Moreover, the number of iterations increases with the number of tuning parameters, which could be expected seeing that more tuning parameters means more dimensions to explore for the optimization algorithm.

Before going on, let us establish a rough estimate of the time that one cost function evaluation could take in real life, i.e. with a human subject participating in a localization experiment. Let us say that reporting the perceived direction for one stimulus (binauralized at a given direction) would take 2 seconds. Then, for 27 positions in the median plane (elevations between $-45°$ and $225°$ with a $10°$ step) and 2 repetitions at each position, one localization experiment would take $27 \times 2 \times 2$ s $= 108$ s $= 1.8$ min.

As mentioned above, in order for the difference between the final and ground truth median APE (or QE) to be non-significant, at least 20 PCs are needed in the ARI case.

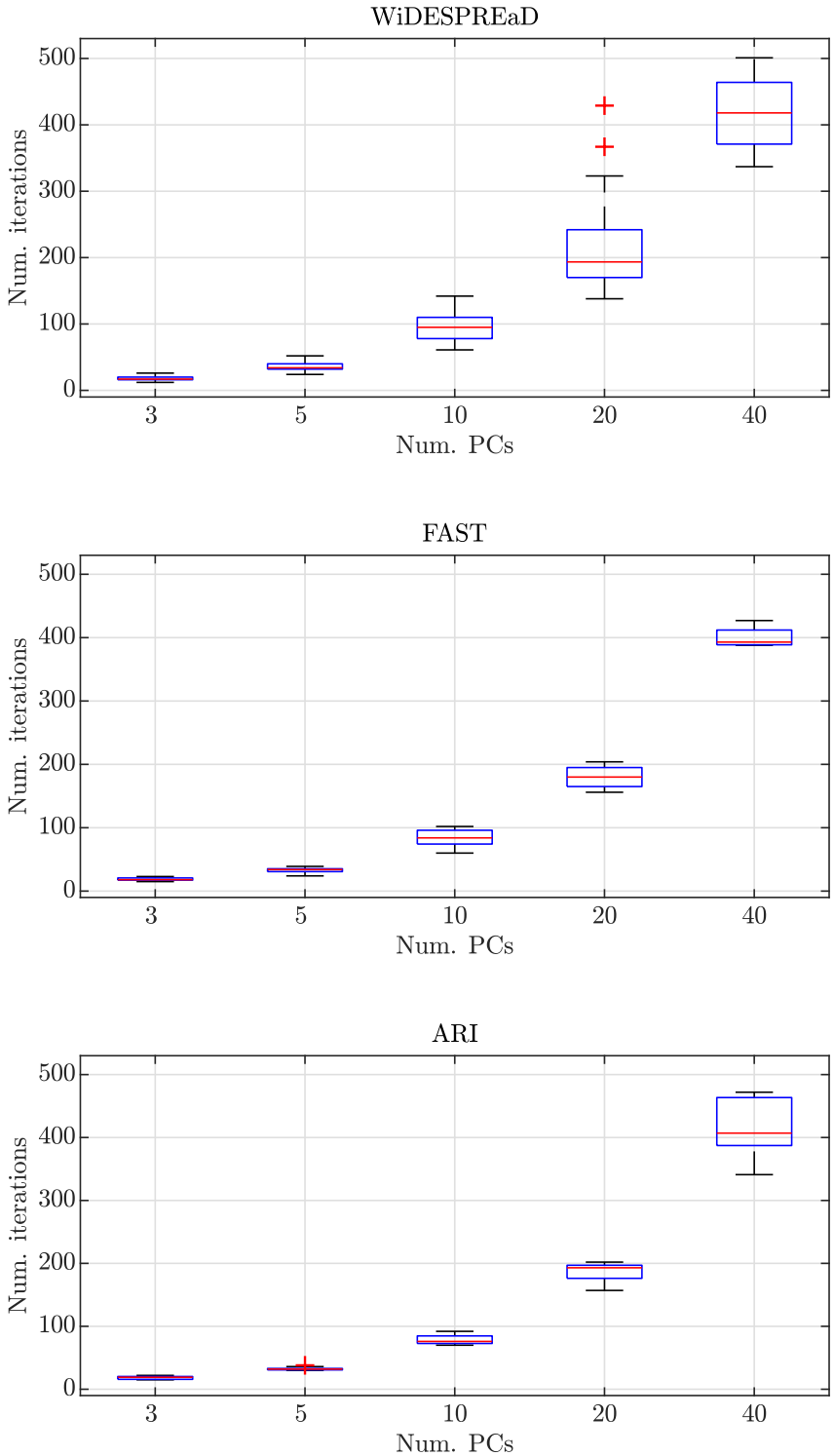


Figure 4.8 – Boxplots of the number of iterations needed to converge, as a function of the number of tuning PCs for the WiDESPREaD (top), FAST (middle) and ARI (bottom) datasets.

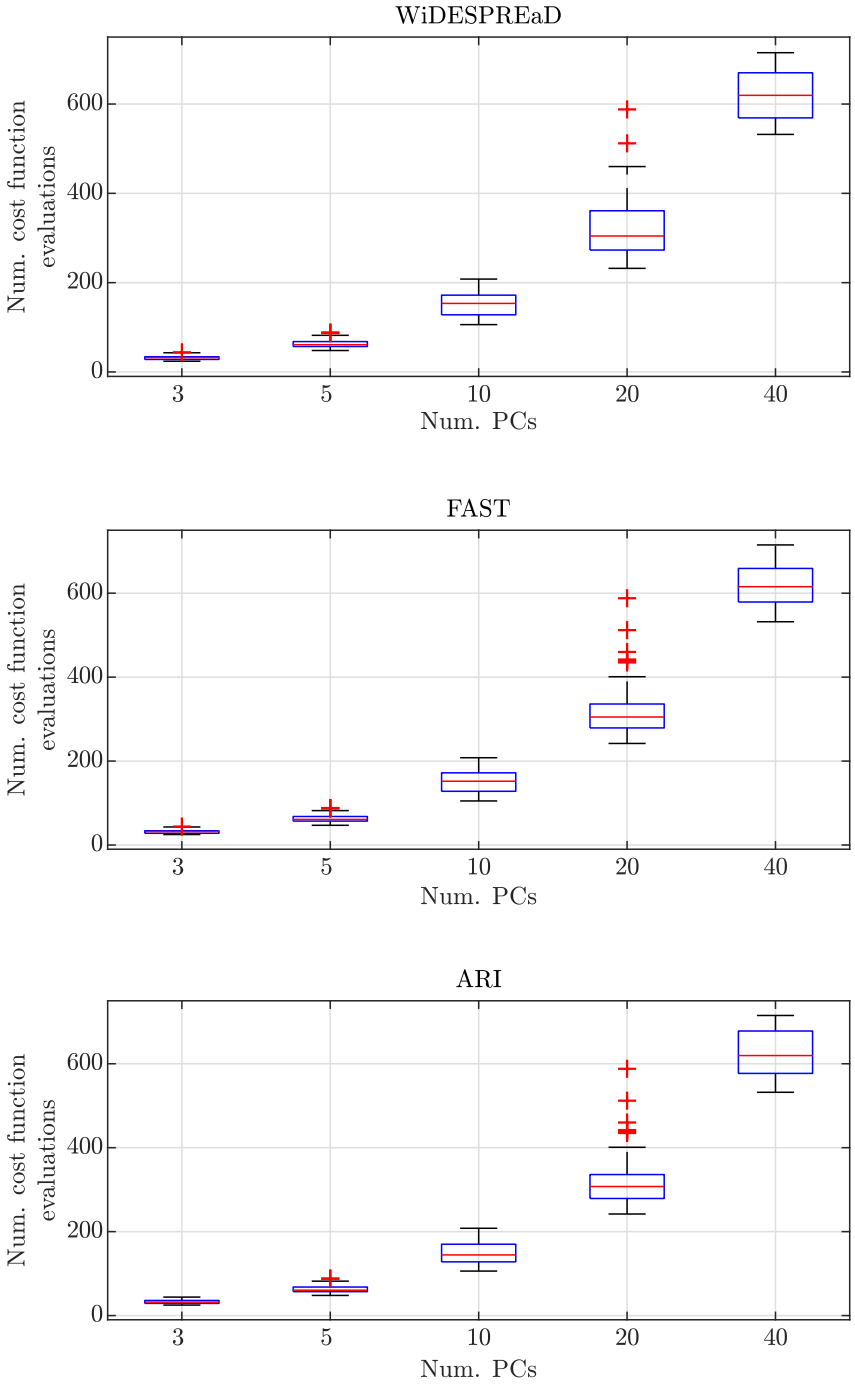


Figure 4.9 – Boxplots of the number of cost function evaluations needed to converge, as a function of the number of tuning PCs for the WiDESPREaD (top), FAST (middle) and ARI (bottom) datasets.

However, the conditions with 20 and 40 PCs require many iterations to converge: the median numbers of iterations are about 200 (193, 180 and 194) and about 400 (407, 427 and 418), respectively. The matching median numbers of cost function evaluations are about 300 (308, 305 and 305) and about 600 (620, 616 and 620), which would roughly correspond to tuning times of 300×1.8 min $\simeq 9$ h and 600×1.8 min $\simeq 18$ h, which are highly impractical.

In contrast, for the conditions with 3 and 5 PCs, convergence is reached in about 20 (medians of 20, 22 and 20) and 40 (medians of 38, 38 and 41) iterations, respectively. This corresponds to about 30 (medians of 34, 34.5 and 34) and 70 (medians of 68.5, 68 and 37.5) cost function evaluations, i.e. respective total tuning time estimates of 30×1.8 min $\simeq 1$ h and 70×1.8 min $\simeq 2$ h. Despite being long, such sessions of localization experiments may be feasible for a real listener, in particular if less than 27 positions are tested in the localization task.

As discussed above, the final median APE, QE and PE in those conditions are significantly higher that the ground truth ones. However, the final median APE, QE and PE are also significantly lower than the initial and baseline conditions. It thus appears that such tuning sessions would offer partial but substantial individualization in terms of localization performance. In the ARI case with $p = 5$ PCs, for instance, the distance to the median ground truth APE (29°) is reduced by more than half ($56\ \% = \frac{40°-54°}{29°-54°}$) between initialization (median of 54°) and convergence (median of 40°). When looking at the baseline KU-100 condition (median APE of 66°) which corresponds to a standard non-individualized VAS, the distance to the median ground truth APE is even more largely reduced, by $70\ \% = \frac{40°-66°}{29°-66°}$. Regarding quadrant errors, the improvement rates are very similar: $54\ \% = \frac{14\%-23\%}{6.3\%-23\%}$ between the Initial and Final 5 conditions, and $69\ \% = \frac{14\%-31\%}{6.3\%-31\%}$ between the KU-100 and Final 5 conditions.

Evolution throughout optimization

The evolution of the APE (QE, respectively) throughout the optimization process is shown for all test virtual listeners in Figure 4.10 (Figure 4.11, respectively). In general, the median APE and QE decrease with the number of iterations. However, sometimes the APE and QE can slightly increase, due to the regularization scheme having found a solution less extreme in terms of PCWs at the cost of a small increase in APE. This has a particularly strong impact on the median behavior of the APE and QE in the FAST case, due to the small number (5) of virtual listeners and large inter-individual difference

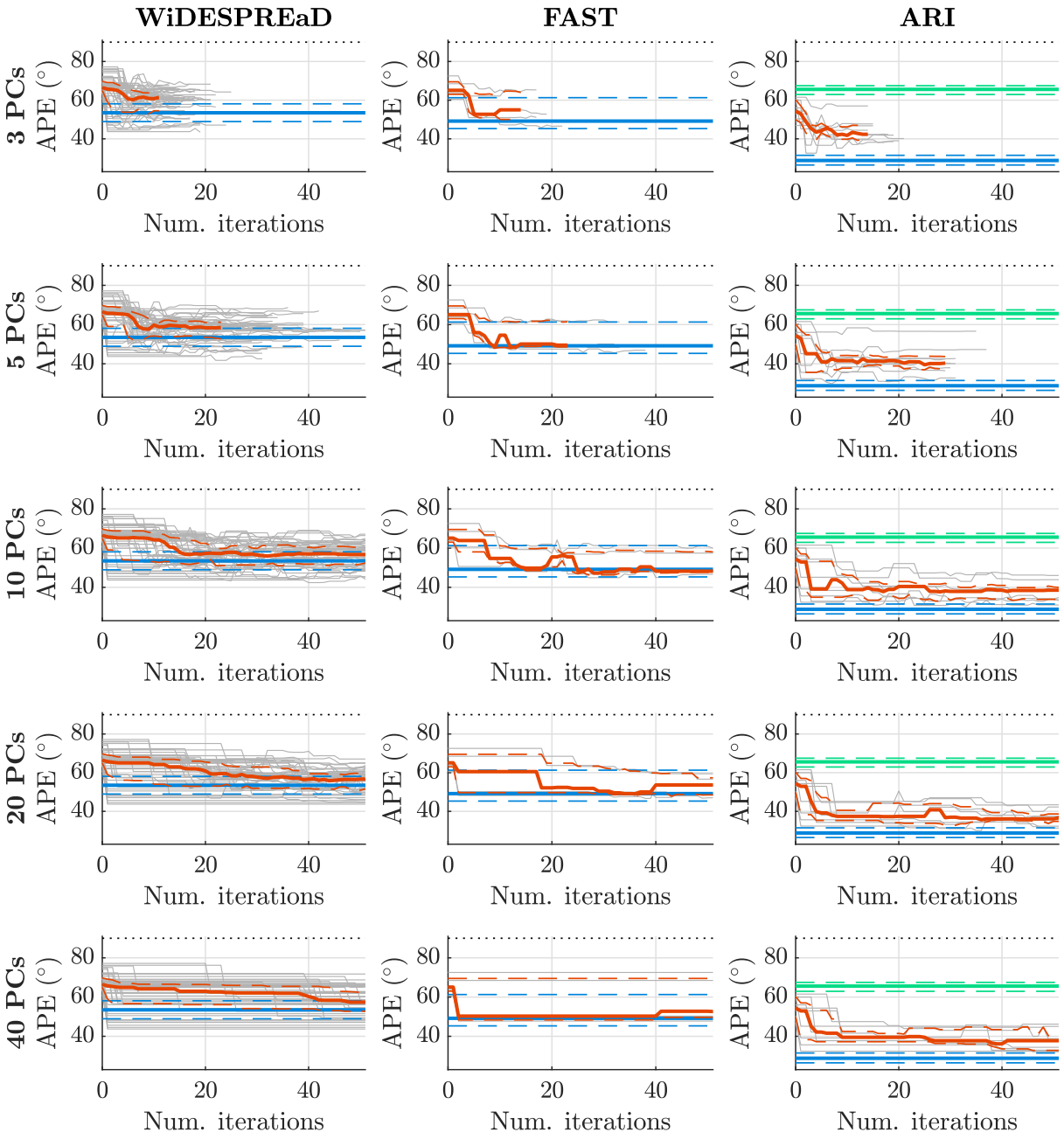


Figure 4.10 – APE throughout the first 50 iterations of the optimization process in all conditions for all test subjects (light gray). Median and quartiles of the APE across subjects are plotted as continuous and dashed red lines, respectively. The median and quartiles of the ground truth are plotted as horizontal blue lines, continuous and dashed, respectively. Finally, the median and quartiles of the baseline condition are plotted as horizontal green lines, continuous and dashed, respectively. The horizontal dotted line shows the localization error associated with random answers.
Top to bottom row: $p = 3$, 5, 10, 20 and 40 PCs. Left to right column: WiDESPREaD, FAST and ARI datasets.

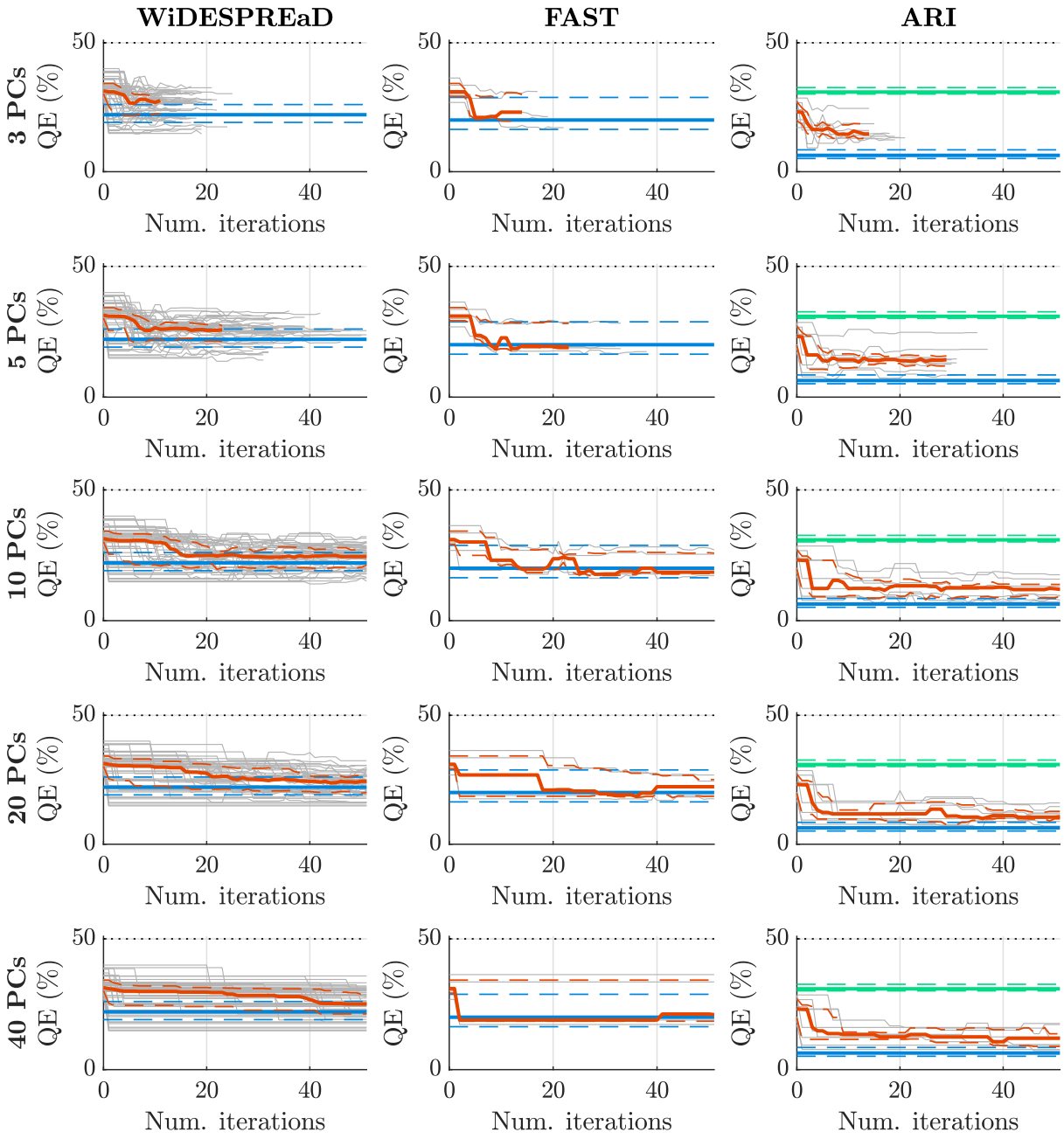


Figure 4.11 – QE throughout the first 50 iterations of the optimization process in all conditions for all test subjects (light gray). Median and quartiles of the QE across subjects are plotted as continuous and dashed red lines, respectively. The median and quartiles of the ground truth are plotted as horizontal blue lines, continuous and dashed, respectively. Finally, the median and quartiles of the baseline condition are plotted as horizontal green lines, continuous and dashed, respectively. The horizontal dotted line shows the localization error associated with random answers.
Top to bottom row: $p = 3$, 5, 10, 20 and 40 PCs. Left to right column: WiDESPREaD, FAST and ARI datasets.

in localization error from initialization to convergence. The decrease in median APE and QE appears to be slower for WiDESPREaD than for the other datasets, due to an early stagnation phase whose duration varies from subject to subject. For WiDESPREaD, the decrease gets slower when the number of PCs increases, but this behavior is not clear for the other datasets.

In the case of the ARI dataset, the median APE and QE decrease quickly within the first dozen iterations before pursuing the decrease more slowly. For instance, after 10 iterations the median QE is between 12 % and 14 % for any number of PCs p, i.e. about or below the median QEs of the Final 3 and Final 5 conditions. As a consequence, even if 20 or 40 PCs were to be retained for the tuning process, similar localization performance would be obtained after 10 iterations than with only 3 or 5 PCs.

At first glance, using many PCs thus appears to be desirable for practical HRTF tuning applications: it provides similar localization performance as with 3 or 5 PCs within the first dozen iterations of the optimization process, but allows the listener to spend more tuning time to further improve the rendering if he desires. However, a given number of iterations does not correspond to the same tuning time depending on the number of PCs p. The latter corresponds in fact to an offset in the number of localization tasks to be performed. Indeed, by construction, the Nelder-Mead's algorithm performs during the first iteration $p + 1$ cost function evaluations. For instance, for $p = 40$ PCs, 35 more localization tasks are to be performed during the first iteration than for $p = 5$ PCs. In our simulations, in the ARI case, 10 iterations corresponded to about 19 (between 18 and 20) localization tasks for $p = 5$ PCs, against about 50 (between 49 and 51) for $p = 40$ PCs. The corresponding estimated tuning times are 34 min and 90 min, respectively, a considerable difference for comparable localization performances.

4.4 Actual Listening Tests

For the tuning experiments with actual listening tests, we used the HRTF model trained on the ARI dataset, previously used in the tuning simulations (see Section 4.3). Indeed, unlike WiDESPREaD and FAST which are PRTF datasets, it includes the filtering contributions of the head and torso.

Aiming at a tuning session of about one hour, we set the number of tuning PCs to 5. As a reminder, we roughly estimated in Section 4.3 that the time needed to reach convergence with 5 PCs was 2 hours in median. With 5 PCs and the ARI HRTF model,

the localization performance was substantially and significantly improved compared to initialization and to the baseline condition, although it remained significantly higher than ground truth performance. After some informal trials of the tuning procedure, in order to reduce the duration of each localization task, we limited the median-plane test positions to 8 polar angles, at roughly every 30° and all present in the measurement grids of the ARI HRTF sets: {−30°, 0, 30°, 60°, 120°, 150°, 180°, 210°}.

4.4.1 Localization Task

Subjects

12 listeners (5 female and 7 male) participated in the experiment and were aged between 24 and 37 years old (28 on average). 9 were naive listeners with no experience with listening tests and 2 were experienced with localization experiments. All participants reported having normal hearing.

Localization task

Each listener participated in a rather large number of localization tasks (between 20 and 88). Due to the iterative nature of the tuning process, the listener is presented with a single HRTF set by localization task.

For the localization task, the listener was presented with each one of the 16 stimuli. After listening to a given stimulus as many times as he wanted, he reported his answer then moved on to the next one. There was no time limit for answering, although swift answers were encouraged due to the large numbers of HRTF sets to be evaluated in one session.

The participant was asked to report the perceived angle on a 2-D interface (see Figure 4.12). Such an exocentric method is known to be less accurate and less intuitive than an egocentric one [Bahu16a, Chap. 4; Katz19, pp. 359-361]. However, all directions are equally easy to report, while with egocentric methods the rear positions are more difficult to evaluate accurately, due to bio-mechanical limitations. Furthermore, it is materially easier to set up, as it does not require an additional tracking device for the head, hands or any other object used for pointing. Finally, this allowed us to use *as is* a user interface previously developed at 3D Sound Labs.

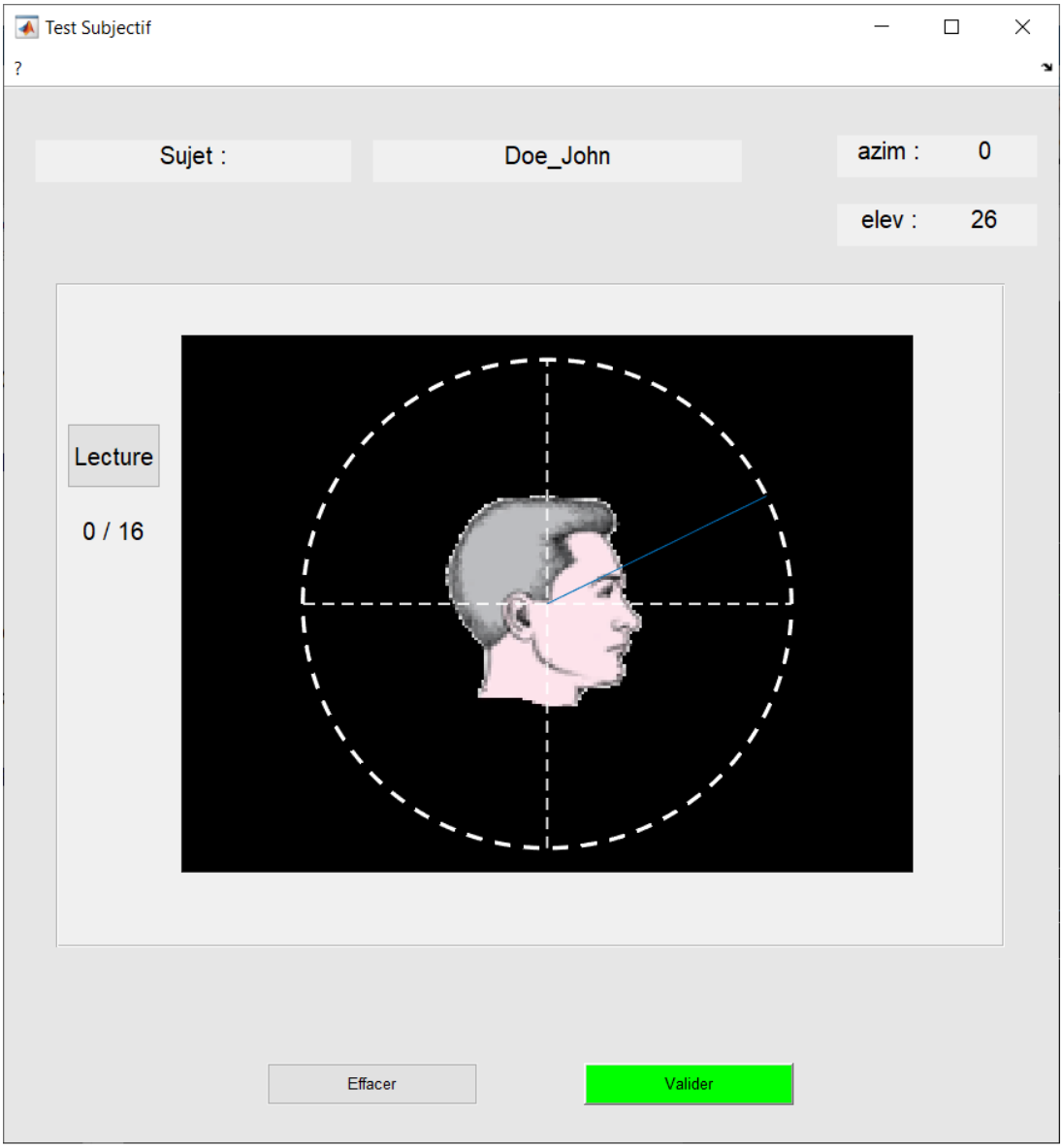


Figure 4.12 – Screenshot of the 2-D graphical user interface used to report the perceived direction of the stimuli in the median plane.

Stimuli presentation

The test stimuli that we used were a sequence of three white noise bursts of 40 ms, separated by silences of 30 ms [Andreopoulou17; Zagala20]. To avoid artifacts, each noise burst was faded in and out linearly in 2 ms. While white noise was chosen in order to include spectral cues over the whole audible frequency range, the bursts were kept short in order to limit the duration of the tests, to limit auditory fatigue and to encourage intuitive answers.

During each localization task – which corresponded to one HRTF set, and one cost function evaluation in the optimization scheme – the virtual sound source was presented at 8 different polar angles ($-30°$, 0, 30°, 60°, 120°, 150°, 180° and 210°), twice each, for a total of 16 stimuli, presented in random order.

The binauralized stimuli were played over a pair of Sennheiser HD 650 open circumaural headphones, *via* an Alesis iO2 sound card, and at a sampling rate of 48 kHz.

We performed no headphone equalization (HpEq) before presenting the binauralized stimuli. Indeed, while it is generally admitted in the literature that performing individual HpEq yields better VAS quality, such an equalization is independent of sound direction, thus being equivalent to source filtering, and has not been shown to have a significant impact on sound localization [Engel19].

Protocol

A session of localization experiments went as follows. After welcoming the participant, an operator (the author) read them instructions for the series of localization tasks. These instructions were also provided in the form of a written document.

It was explained to the listener that he or she was about to participate in about twenty listening tests. In each listening test, he or she would be prompted 16 times to indicate the perceived direction of an auditory stimulus.

Each stimulus was presented once by the software, and the listener could replay it any number of times before giving his answer. The user interface allowed to cancel an answer and go back to a previous one – if the participant had clicked by error, for instance.

The participant was asked to perform localization tasks during one hour, but could perform longer if he or she wanted. He or she was strongly encouraged to take breaks to limit auditory fatigue, at the end of every localization tasks if needed. In practice, most participants took one long break of about 10-15 min.

Orally, the operator indicated that it was normal to feel that the task was difficult and to not be able to localize some of the stimuli – free-field median-plane localization is harduous, especially with non-individual HRTF sets. In such cases, the listeners could give a random answer. Informally, intuitive responses were encouraged.

The sound level was set for listener comfort prior to the localization tasks, then remained untouched for the rest of the session.

In addition to the localization tasks that were part of the optimization scheme, localization performance with the baseline HRTF set (that of the Neumann KU-100 manikin, as measured by the ARI team) was evaluated by means of a localization task before the tuning session itself.

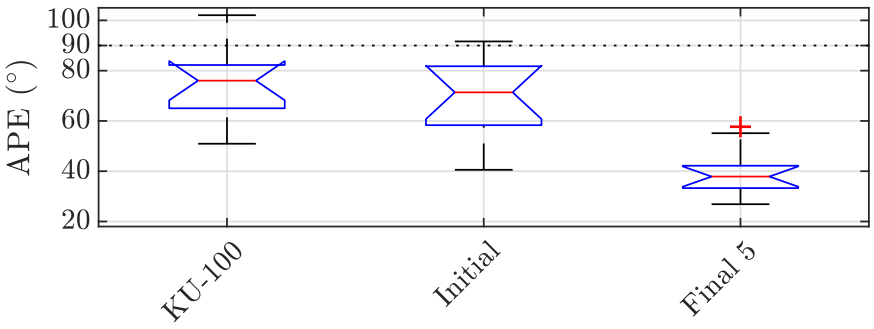
4.4.2 Results

Localization performance

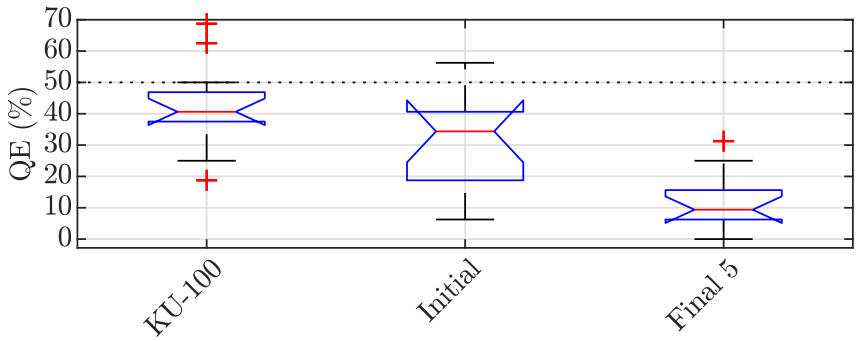
We herein compare the localization performances in three HRTF set conditions: baseline, initial and final. The baseline is the HRTF set of the KU-100 manikin as measured by the ARI. The initial condition is the HRTF set that was evaluated at the initialization of the tuning process. It corresponds to the average of the training set for the PCA HRTF model (all PC weights set to zero). The final HRTF set is the customized HRTF set provided by the proposed method. It corresponds to the solution retained by the Nelder-Mead simplex algorithm based on the various cost function evaluations throughout the tuning session.

The order of these evaluations was fixed for all subjects and not randomized, due to the constraints of the tuning method. Indeed, the perceptual evaluation of the baseline was performed before the start of the tuning session, and that of the initial condition was performed just after, when the tuning process started. As to the final condition, its perceptual evaluation occurred later throughout the tuning session.

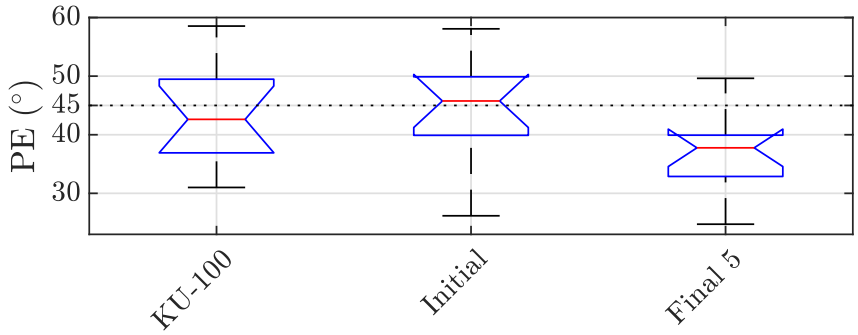
Initial and baseline conditions Similarly to what was observed and discussed in the case of the simulated localization tasks (see Section 4.3), the initial median APE (71°) is lower than the KU-100 one (76°). However, the difference between both median APEs is here not significant and is only of $76 - 71 = 5°$, against $66 - 54 = 12°$ in simulations. The difference between initial and baseline median APEs is mostly explained by the – non-significant – difference in quadrant error: the initial and KU-100 QEs are 34 % and



(a) APE



(b) QE



(c) PE

Figure 4.13 – Localization error outcome of the real tuning experiment during the tuning phase: notched box plots of the APEs (top), QEs (middle) and PEs (bottom) of the baseline (KU-100), initial and final (Final 5) conditions. The horizontal dotted line shows the localization error associated with random answers.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].

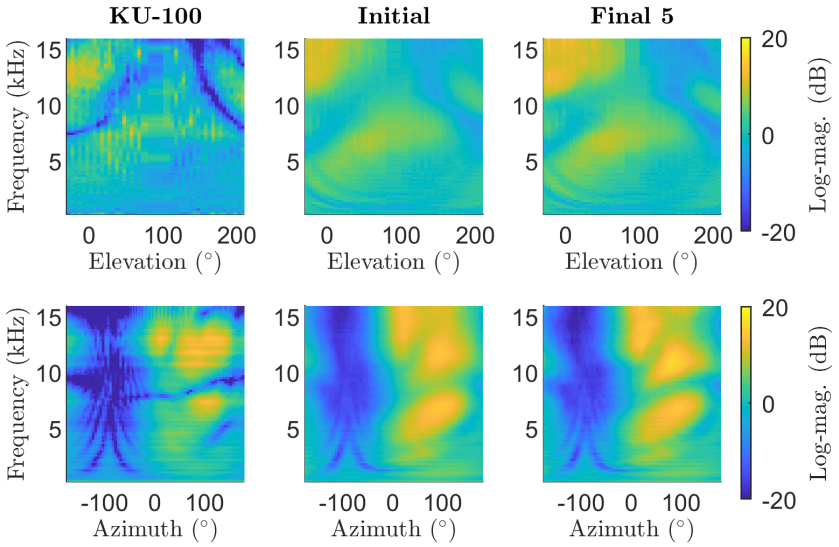


Figure 4.14 – Outcome of the tuning process for an exemplary subject. The KU-100 baseline, initial and final magnitude HRTF sets are plotted for directions of the median- (top) and horizontal-plane (bottom) directions. Matching APEs are 70.1°, 63.4°, 28.4°, respectively.

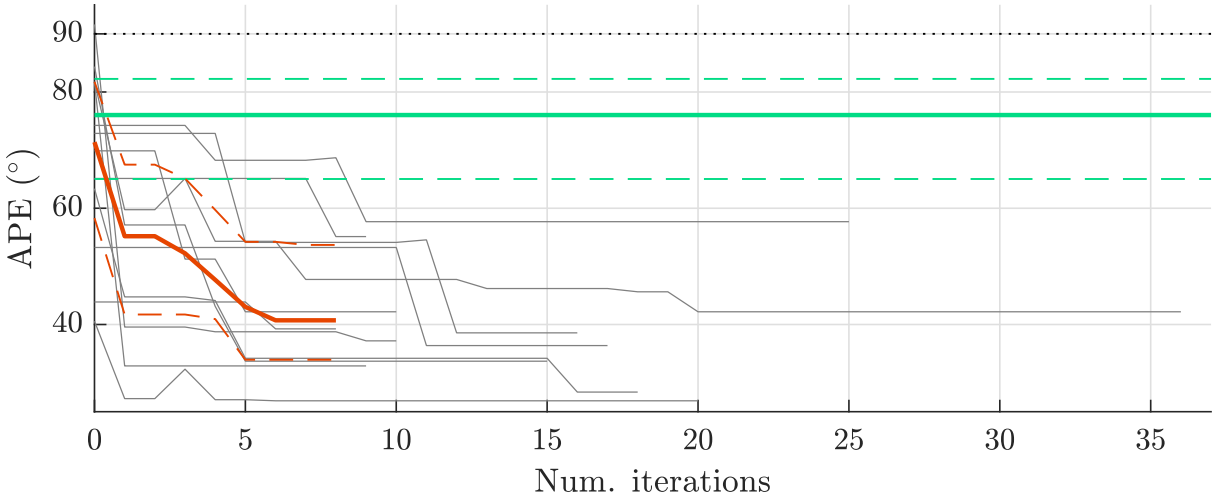


Figure 4.15 – APE throughout the tuning process based on real localization tasks for the 12 participants (gray). Median and quartiles of the APE across subjects are plotted as continuous and dashed red lines, respectively. The median and quartiles of the ground truth are plotted as horizontal blue lines, continuous and dashed, respectively. Finally, the median and quartiles of the baseline condition are plotted as horizontal green lines, continuous and dashed, respectively.

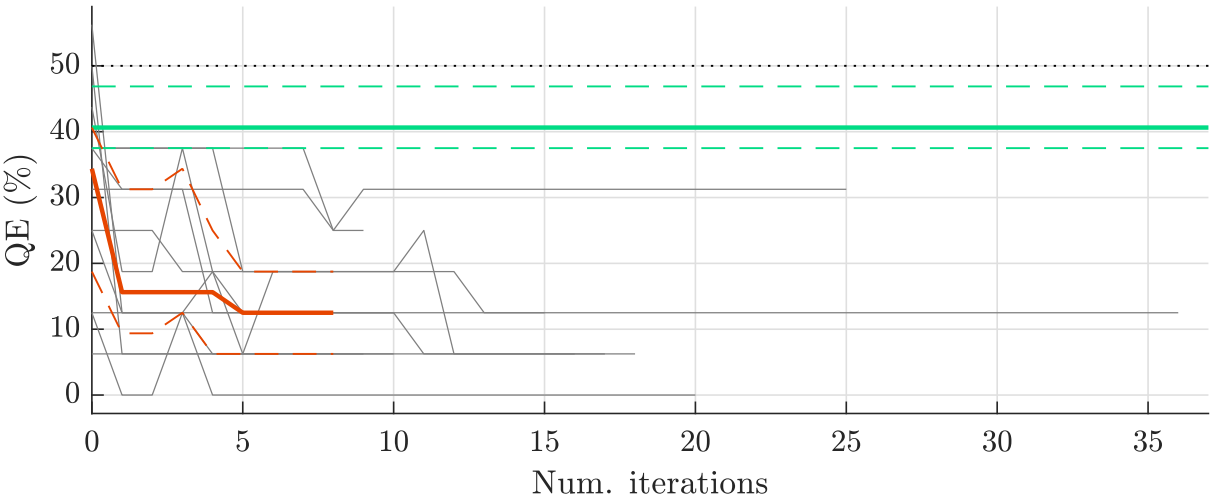


Figure 4.16 – QE throughout the tuning process based on real localization tasks for the 12 participants (gray). Median and quartiles of the QE across subjects are plotted as continuous and dashed red lines, respectively. The median and quartiles of the ground truth are plotted as horizontal blue lines, continuous and dashed, respectively. Finally, the median and quartiles of the baseline condition are plotted as horizontal green lines, continuous and dashed, respectively.

41 %, respectively. On the other hand, the tendency is reversed with the local angular errors (PEs). Indeed, the median initial PE of 46° is greater than the KU-100 one of 43°, although not significantly so. Neither median PE differs significantly from the chance PE. Let us note that the pointing method employed in these experiments is not very accurate, and that the results in terms of local polar error, PE, are thus to be considered in this light.

The initial and baseline localization errors were in general higher that those from the simulations. Indeed, in both conditions, the median APE is significantly greater in the actual experiments than in the simulated ones: 76° against 66° for KU-100, and 71° against 54° for the average HRTF set (initial condition). This trend is also found in QEs, with 41 % against 31 % for KU-100 (significant), and 34 % against 24 % for the initial HRTF set. Following the same trend, the actual median initial PE (46°) is significantly larger than the simulated one (37°). In contrast, the actual median KU-100 PE (43°) is equal to that of the simulations.

While the median QE of our initial condition (34 %) is comparable to the results of [Middlebrooks00] with non-individual human-subject HRTF sets (about 33 %), it is higher than the median QE reported in [Middlebrooks99b] in similar conditions (19 %).

The median QE is larger in the KU-100 condition (41 %) than in both studies. In terms of median PE, our initial (47°) and KU-100 (39°) conditions are respectively somewhat higher and comparable to the results of [Middlebrooks99b] (about 41°). These differences may be due to the different nature of the non-individual HRTF sets (mathematical average of measured HRTF sets / measurements of a manikin / measurements of other human subjects) or to differences in localization experiment methodology. In particular, in contrast with [Middlebrooks99b], in the present study the listeners did not go through any training phase before participating in the localization tasks. Indeed, Majdak *et al.* [Majdak10] find that training allows substantial improvement localization performance in an individualized VAS. For instance, the QEs that they reported in Table 4 are 21 ± 19 % (average $\pm$ standard deviation) and 22 ± 21 % in the two conditions without training, against 11 ± 7.8 % in the condition with training, which they found comparable to the 7.7 ± 8.0 % of [Middlebrooks99b].

Optimization outcome Localization performance has generally been substantially improved from the initial to the final HRTF set. As can be seen in Figure 4.13, the median APE significantly decreased by almost a factor of two (from 71° to 38°).

This decrease is mostly due to a large drop in QE. Indeed, the decrease in QE is significant as well, and constitutes a drop by almost a factor 4 – from 34 % to 9.4 %. The median final QE is the same as the median final QE (9.4 %) obtained in the ARI tuning simulations with 20 PCs, that is four times more PCs than in the present experiment (see Figure 4.4, Section 4.3). Moreover, the median final QE is in the order of the median QEs obtained with individual HRTF in previous studies, such as 10 % [Baumgartner14], 8 % [Middlebrooks00] and 4 % [Middlebrooks99b] (see Section 4.3 for more detail on these studies and the associated QEs).

Regarding the local polar errors (PEs), we observe a less spectacular but statistically significant decrease from the initial to the final condition: 46° to 38° in median. The final median PE is however not significantly lower than the baseline KU-100 one (43°).

Overall, the final localization performance is very good, with a median QE in the order of that of individual HRTF sets as reported in previous studies [Middlebrooks99b; Middlebrooks00; Baumgartner14], while the baseline and initial conditions seem to be comparable or poorer than the one reported for non-individual HRTF sets in two of these studies [Middlebrooks99b; Middlebrooks00]. This remarkably good result is likely partly due to training. Indeed, although the listeners had no visual feedback, it is likely that

they improved at the localization task over the course of their tuning session of 35 to 83 min. This would be a positive side-effect of the method. On another, the results might be partly overestimated by the fact that the Nelder-Mead algorithm always retains the best of all previously tested solutions. Indeed, this best solution might sometimes be more due to variability in the participant's answering that to the best suitability of the HRTF set. The existence and extent of this behavior would require further scrutiny.

Tuning time

As indicated in Section 4.4.1, the intended duration for a tuning session was about one hour. In practice, the operator adapted to the tiredness and motivation of the participants, resulting in tuning session durations between 35 and 83 min, and a median of 56 min (see Figure 4.17).

As can be seen in Figure 4.15, the APE generally decreased within the first 11 iterations, then plateaued, decreasing in a slower fashion afterwards. Similarly, the QE generally dropped within the first 13 iterations before plateauing, as shown in Figure 4.16. It is worth noting that the APE and QE sometimes re-increased, which was generally due to the optimization process finding an HRTF set that minimized the regularization cost (avoiding extreme PC weights) at the expanse of a small increase in APE.

At 6 iterations, the median APE and QE were already of 41° and 13 %, that is 92 % and 88 % of the decrease observed between the median initial and final APE and QE, respectively. Depending on the tuning experiment, these 6 iterations corresponded to a median of 14 cost function evaluations (minimum and maximum of 11 and 23, respectively), for a median tuning time of 21 min (minimum and maximum of 5.7 min and 43 min, respectively).

In the actual experiments, the average time spent per localization task over the tuning session was on average (across listeners) 1.5 min, and ranged from 26s to 3.1 min. This is rather consistent with the previous rough estimate of 2 s per answer (see Section 4.3) and the consequent estimation of 1.2 min $= 8 \times 2 \times 2$ s per localization task. Experience did not seem to be a very important factor in quickness to answer. Indeed, while, among the two experienced listeners, one of them was among the fastest (44s), the other was just slightly above average (1.2 min). On the other hand, although the two slowest participants – and outliers in this regard – were naive listeners, the fastest was a naive one as well.

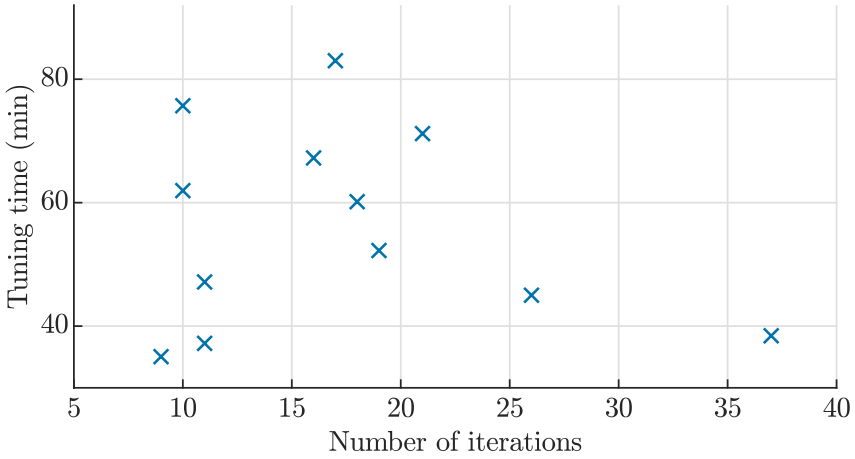


Figure 4.17 – Scatter plot of the durations of the tuning sessions – breaks excluded – as a function of the number of iterations, for all 12 participants.

Comparison with other HRTF individualization methods

As we have seen above, the proposed method allowed a significant and substantial reduction of localization errors compared to the baseline and initial HRTF sets, in about one hour of listening tests. In particular, the quadrant error rate was reduced by almost a factor 4 between initialization and the end of the tuning session. The final median QE of 9.4 % is of the same order as those observed in localization experiments with individual HRTF sets [Middlebrooks00; Baumgartner14], while the baseline and initial HRTF set yielded somewhat poorer performance (median QEs of 41 % and 34 %, respectively) than reported with non-individual HRTF sets in [Middlebrooks99b] (median QE of about 19 %, 11 listeners, 21 non-individual conditions) and [Middlebrooks00] (median QE of about 33 %, 5 listeners, 9 non-individual conditions). Let us compare these results to a few other perceptual feedback-based HRTF individualization techniques. For more detail on the studies mentioned below, please refer to Chapter 2, Section 2.3.

Selection Due the ever growing number and size of HRTF datasets, a common approach has been to select a best-fit non-individual HRTF set among a database. Katz *et al.* [Katz12], for instance, study the possibility of improving localization performance by selecting a best-fit non-individual HRTF set by means of judgment tasks. In a first experiment, 46 listeners each rated 46 HRTF sets from the LISTEN database (including their own) on a 3-point rating scale (bad/ok/excellent) based on the fidelity of rendered horizontal and vertical virtual trajectories. The duration of this task was approximately

	Precision (%)	Front-back (%)	Up-down (%)	Combined (%)
KU-100	46 (12)	28 (9)	17 (8)	10 (7)
Initial	55 (14)	28 (12)	11 (9)	6 (7)
Final	76 (11)	8 (6)	15 (8)	1 (2)
Individual [Katz12]	63 (4)	20 (3)	13 (3)	4 (2)
Best [Katz12]	46 (3)	32 (3)	15 (3)	6 (2)
Worst [Katz12]	38 (3)	35 (3)	19 (2)	8 (2)

Table 4.1 – Comparison of the results of our localization experiments and that of Katz *et al.* [Katz12] in terms of the classification employed in the latter: average *precision*, *front-back*, *up-down* and *combined* rates (standard deviations in parentheses).

35 minutes. This allowed the authors to identify a subset of 7 HRTF sets which satisfied most of the subjects. In a second experiment, 20 new listeners were asked to rate these 7 HRTF sets in a closely related judgment task, although it differed by the use of a continuous rating scale (from "bad" to "good"). Based on these results, a worst- and a best-fit HRTF set was identified for each subject.

The duration of this task was not reported. However, Zagala *et al.* [Zagala20] report a ranking time of 27 min using a similar approach – based on the rating of horizontal and vertical virtual trajectories – to rate the same 7 HRTF sets.

The worst- and best-fit non-individual HRTF sets were evaluated thanks to a localization task. 10 of the subjects (randomly selected) evaluated the former while the 10 others evaluated the latter. As a reference, 4 listeners (outside the aforementioned 20) evaluated their own HRTF sets. The results of the localization experiment were analyzed by means of the classification of errors introduced by Martin *et al.* [Martin01] (see Chapter 2, Section 2.2.2). We reproduce these results in Table 4.1, alongside the same classification applied to our own localization experiment results.

The localization performance in the individual condition was substantially superior than in both non-individual ones, with average precision rates of 63 % against 46 % and 38 % in the *individual*, *worst* and *best* conditions, respectively, and average front-back confusion rates of 20 % against 35 % and 32 %. The selection process seemed to allow an improvement in localization performance, the average *precision* rate increasing by 21 % between the worst- and best-fit non-individual HRTF sets.

In comparison, our proposed method appears to provide a greater improvement in localization performance. The average *precision* rate increased by 65 % from the KU-100 to the final condition, and by 38 % from the initial to the final condition.

The localization experiments in [Katz12] seem to have yielded higher localization errors than our own in general – i.e. regardless of the various conditions. In particular, the worst of our non-individual conditions, KU-100, is comparable to the precision rate of the best-fit HRTF set in [Katz12], while our final precision rate (77 %) is notably greater on average than that of the individual condition in [Katz12] (63 %). The authors note as well than their individual condition presented poorer localization performance than a previous study by Wightman *et al.* [Wightman89b].

Frequency scaling In [Middlebrooks00], Middlebrooks *et al.* propose a procedure in which a non-individual HRTF set is adjusted by means of a frequency scaling (identical for all directions), based on successive A/B judgments by the listener of various scaled HRTF sets in terms of localization accuracy of median-plane virtual sources. The resulting adapted non-individual HRTF sets were evaluated by means of a localization experiment for 5 participants (out of 20), each listening to one or two non-individual HRTF sets, for a total of 9 *non-individual* and *scaled non-individual* cases.

The tuning procedure took about one hour (including a 15-min training phase), and the resulting median QE was of about 13 %, against 8 % and 33 % in the individual and non-individual conditions.

In comparison, the proposed method seems to produce HRTF sets that provide better localization performance (median QE of 9.4 %) in a similar amount of time. The median QE obtained by Middlebrooks *et al.* is more comparable to the median QE that we observed after 6 iterations of the optimization process, that is a median tuning time of 21 min.

Synthesis Finally, let us compare our proposed method to more closely related approaches, which aim at synthesizing a customized HRTF set based on perceptual feedback from the listener.

Hwang *et al.* [Hwang08a] propose that the listener tune himself 3 PCWs of a *spectral* (see Chapter 2, Section 2.1.3) PCA model of HRIRs. This is a local approach, in the sense that the PCA model generates individual filters rather complete HRTF sets. The tuning was thus performed independently at each of 7 directions of interest – in the median plane. The tuning procedure was tested on three listeners, then its outcome was evaluated by means of a localization experiment. Three HRTF sets were under study: the customized one – produced by the individualization method, the listener's own, and that of the

	Front-back (%)
KU-100	40 (12)
Initial	39 (14)
Final	12 (6)
KEMAR [Hwang08a]	23 (3)
Custom [Hwang08a]	6 (7)
Individual [Hwang08a]	0.4 (0.6)
KEMAR [Shin08]	29 (14)
Custom [Shin08]	10 (3)
Individual [Shin08]	12 (4)

Table 4.2 – Comparison of the results of our localization experiments and that of Hwang *et al.* [Hwang08a, Table VII] and Shin *et al.* [Shin08, Table 1] in terms of average front-back confusion rates (standard deviation in parentheses), *as per* the traditional definition of front-back confusions by Wightman *et al.* [Wightman89b].

KEMAR manikin – a standard non-individual condition. Shin *et al.* [Shin08] propose a closely related approach in which the listener tunes himself 5 PCWs of a *spectral* PCA model of HRIRs. The performance of the resulting customized HRTF set is compared to that of their own and KEMAR HRTF sets by means of a localization experiment in which four listeners participated.

In both studies, front-back confusions were identified using the conventional definition by Wightman *et al.* [Wightman89b]. We report these results in Table 4.2, alongside the front-back confusion rates of our own localization experiments, calculated according to the same definition. In both studies, the customized HRTF set yields a rate of front-back confusion that is, on average, lower than the non-individual KEMAR condition by about 70 % (74 % $= 100 \cdot \frac{23-6}{23}$ for [Hwang08a], and 66 % $= 100 \cdot \frac{29-10}{29}$ for [Shin08]). According to Hwang *et al.*, the difference between the custom and KEMAR condition is significant. This is comparable to the difference between both our non-individual conditions (both KU-100 and initial) and our final condition, with reductions in the average front-back confusion rate of 70 % $= 100 \cdot \frac{40-12}{40}$ and 69 % $= 100 \cdot \frac{39-12}{39}$, respectively. In [Hwang08a], the average front-back confusion rate with the custom HRTF set (6 %) is higher than with the listener's own HRTF set (0.4 %), but Hwang *et al.* report that the difference is not statistically significant, while in [Shin08] the custom average front-back confusion rate (10 %) is slightly lower than the individual one (12 %). While both studies present comparable results for the KEMAR and custom conditions, there is a notable mismatch for

the individual condition. Finally, the front-back confusion rates of both our non-individual conditions (40 and 39 %) are notably higher on average than that of the KEMAR condition of [Hwang08a] and [Shin08], possibly suggesting that our localization experiment protocol yielded overall higher localization error.

Unfortunately, none of these studies reported the duration of the HRTF tuning procedure, although Hwang *et al.* indicated that they chose only 3 PCs precisely to keep it reasonable – after having determined in a first experiment that reconstructing HRTFs from 12 PCs yielded a localization performance indistinguishable from the original HRTFs. Furthermore, in both studies the tuning procedure needs to be performed at each direction of interest, which would likely result in an unpractical total duration for the tuning of even a sparsely spatially sampled HRTF set.

4.5 Conclusion & Perspectives

In this chapter, we proposed a method for low-cost HRTF individualization based on perceptual feedback. It consists in tuning the parameters of a statistical model of magnitude HRTF set based on the localization performance of the listener. Unlike most other similar approaches, the tuning is done globally, i.e. for all sound directions at once – a critical feature if we are to achieve reasonable tuning times. Furthermore, the optimization itself is performed by means of a Nelder-Mead simplex algorithm. The listener is thus solicited for localization performance evaluation only, not for tuning of the HRTF model's parameters.

As a first step, simulated localization experiments by means of the Baumgartner auditory model [Baumgartner14] allowed us to evaluate the proposed method under various configurations – three different datasets (FAST, WiDESPREaD and ARI) and five different numbers of tuning parameters from 3 to 40. In all conditions except one, the optimization process converged to a mag-HRTF set that significantly decreased localization errors (APE, QE and PE) compared to the training set's average HRTF set and to a baseline: the Neumann KU-100 manikin HRTF set. When more than 20 PCs were retained, the final localization errors (APE and QE) were not significantly different from the ground truth. For example, in the case of the ARI dataset, the median QE was reduced from 23 % (with the initial average HRTF set) to between 7.5 % and 15 % (with 40 and 3 PCs, respectively) with the customized HRTF sets. Comparatively, the baseline

non-individual (KU-100) and the individual HRTF sets yielded respective median QEs of 33 % and 6.3 %.

A large difference was observed between the FAST and WiDESPREaD PRTF datasets on the one hand and the ARI dataset on the other hand: the simulated localization performance were considerably higher in median in the former case in all conditions. Yet, results for ARI in the ground truth and in the non-individual (average and KU-100) conditions were consistent with localization errors reported in the literature with individual and non-individual HRTF sets, respectively [Middlebrooks99b; Middlebrooks00; Baumgartner14]. It seems likely that the absence of head- and torso- spectral features in PRTFs resulted in higher localization errors in the auditory model. As a consequence, we used the HRTF model based on the ARI dataset in the subsequent tuning experiments with actual listeners.

Regarding the duration of the procedure, the number of iterations required to converge increased with the number of tuning parameters, quickly reaching values impractical for a real-subject application: up to more than 500 iterations for 40 PCs, for a roughly estimated tuning time of 18 hours for a real listener. With only 3 or 5 PCs, however, the estimated tuning session duration was limited to one or two hours, whereas substantial improvement in localization performance was achieved – although more modest than when more PCs were retained. For example, with $p = 5$ PCs, the median APE was reduced by 56 % of the difference between initialization and ground truth.

While long, such sessions appeared to be feasible for a real listener, particularly if the number of test directions were to be reduced compared to the 27 considered in our estimation of total tuning time.

We put to the test this alleged feasibility as a second step by submitting the tuning procedure to 12 real listeners. As a compromise between expected final localization performance and tuning session duration, we retained 5 PCs for these experiments. The results somewhat differed from the simulations. Indeed, the customized HRTF sets produced by the procedure yielded substantial improvement of the localization performance compared to both non-individual conditions (average and KU-100 HRTF sets) in one hour of listening tests in median, thus confirming the feasibility of the procedure in that time frame. In particular, the median QE was reduced by nearly a factor 4 as a result of the tuning procedure, for a final value of 9.4 %. This is a good rate of quadrant errors, comparable to values reported in the literature for individual HRTF sets (10 %, 8 %,

4 %) [Baumgartner14; Middlebrooks00; Middlebrooks99b]. Yet, it does not appear that this particularly low final QE is due to a general underestimation of localization errors related to our localization experiment protocol. Indeed, the baseline and initial conditions yielded comparable or poorer localization performance (median QEs of 41 % and 34 %, respectively) than Middlebrooks *et al.* reported for non-individual HRTF sets (median QEs of 19 % and 33 %) [Middlebrooks99b; Middlebrooks00].

This notably large reduction in localization errors may be partly due to some training of the listener throughout the tuning session, which would represent a positive side-effect of the method. It is also possible that the final localization performance is somewhat overestimated due to the fact that the Nelder-Mead optimization algorithm systematically retains the lowest evaluation of the cost function, which may be due not only to HRTF set customization, but also to variation in the listener's answering.

Although the tuning sessions were quite long (one hour in median), we observed that most of the decrease in localization error (88 % and 92 % of total median decrease in QE and APE, respectively) occurred within the first 6 iterations: in about 20 min, a final median QE of 13 % was achieved. The present method thus offers flexibility in the form of a trade-off between HRTF tuning duration and localization performance.

Comparing our results to that of other HRTF individualization techniques, it appears that our proposed method improves localization performance substantially more than the selection of a best-fit non-individual HRTF set among 7 representatives HRTF sets [Katz12]. It however takes longer: one hour against 25 min [Zagala20]. The proposed approach also appears to reduce localization errors more than the self-tuning of a global frequency scaling parameter (in order to adapt a non-individual HRTF set to the user) [Middlebrooks00] in a comparable amount of time. Finally, our procedure yields a reduction in front-back confusions comparable to that of a related method which consists in tuning by ear the parameters of an HRTF PCA model in the median plane [Hwang08a; Shin08]. The duration of the tuning procedure was not reported in either study. It should be expected, however, to be substantially higher than ours, in particular when extended beyond the median plane. Indeed, in contrast with our global approach, their tuning procedure must be performed at each direction of the HRTF set.

Overall, in this chapter we proposed a method for low-cost HRTF individualization based on localization tasks, allowing considerable improvement in localization performance compared to non-individual conditions, up to a performance comparable to that

of individual HRTF sets found in the literature. Its main disadvantage is the length of a tuning session – one hour in median in the present experiments. In particular, extending the method to directions beyond the median plane is likely to lengthen the localization tasks. However, it offers flexibility in the form of a compromise between localization performance and tuning time: in the present experiments, most of the decrease in localization error occurred during the first 6 iterations, that is a median tuning time of 20 min.

As mentioned earlier, the large improvement in localization performance observed with the proposed method may be partly due to factors other than HRTF set customization itself, such as training for the localization task throughout the procedure. While this would be a positive side-effect, it may be interesting in the future to investigate the existence and part of such an effect in localization performance improvement throughout the tuning session. Future work also includes evaluating the customized HRTF sets in a separate listening test, so as to gain further understanding of their perceptual quality, while escaping potential biases such as the aforementioned selection bias of the Nelder-Mead algorithm. Moreover, sound source directions other than the ones used for tuning should be evaluated.

Going further, the proposed approach ought to be extended to positions beyond the median plane. While it already produces a whole-sphere magnitude HRTF set, it does so only based on median-plane localization performance. ITD thus needs to be included in the HRTF model. For instance, an ITD model could be tuned alongside the magnitude HRTF set model. This would allow, on the one hand, the production of an HRTF set that includes ITD and, on the other hand, the tuning of the HRTF set based on localization tasks at positions throughout the whole sphere. In order to limit tuning time, an important subject of study would be the identification of a minimal viable spatial sampling of the sphere for the tuning procedure to work.

Conclusion & Perspectives

Summary

The work presented in this thesis falls within the context of binaural synthesis, a technology that allows the rendering of immersive audio in headphones or earbuds. In contrast with loudspeaker-based techniques (such as wave field synthesis), an inescapable advantage of binaural synthesis lies in its simplicity of implementation. Indeed, only a standard pair of headphones and a little computing power are needed to summon a convincing virtual audio scene (VAS). Thanks to the omnipresence of smartphones, tablets and laptops, and to the democratization of virtual and augmented realities (VR and AR), binaural synthesis has known a growing popularity. Providing an optimal experience to the listener, which involves using individual head-related transfer functions (HRTFs), is thus more and more important. However, in most applications a generic set of HRTFs is used. Indeed, providing individualized HRTFs to the public has proved challenging and remains an open issue, which this thesis addresses.

Background / state of the art

In Chapter 1 and Chapter 2, we provided background knowledge to the work presented in this thesis. In the former, notions regarding human auditory localization, localization cues and binaural synthesis were introduced. In the latter, we laid down a state of the art on various subjects such as HRTF modeling, evaluation, individualization and databases. In particular, we established in Section 2.4 a state of the art of contemporary HRTF databases, noting that they include very few subjects (less than 201) compared to the dimensionality of an HRTF set (in the order of 10^4 to 10^5 degrees of freedom, see Table 3.1). Furthermore, we presented in Section 2.3 a survey of HRTF individualization techniques, which was the subject of an article presented at the 145$^{\text{th}}$ Audio Engineering Society Convention [Guezenoc18]. Four approaches were distinguished: acoustic measurement, numerical simulation, and less direct yet user-friendly methods either based on morphological data or perceptual feedback. We remarked that, with respect to constraints of user-friendliness and perceptual assessment of the resulting HRTFs, the latter approach

represents a particularly interesting option.

HRTF Individualization based on perceptual feedback

In Chapter 4, we presented such a method which consists in tuning the weights of a PCA model of magnitude HRTF set based on listener localization performance. Unlike many approaches, the tuning is performed globally i.e. for all directions at once. Furthermore, the listener is prompted for subjective evaluation but is not asked to tune the model, the optimization being performed by a Nelder-Mead simplex algorithm [Nelder65]. In the present work, the listening tests were restricted to the median plane, where the ITD and ILD are almost zero, thus focusing on the monaural spectral cues which are the most crucial for HRTF individualization (see Chapter 1, Section 1.3.2).

As a first step, psycho-acoustic simulation of the listening tests by means of an auditory model (see Chapter 2, Section 2.2.3 and [Baumgartner14]) allowed us to perform a preliminary evaluation of the proposed method under various settings: 3 different training datasets for the PCA model, and 5 different numbers of tuning parameters ranging from 3 to 40. Testing these different configurations would have represented a prohibitive amount of time with actual subjective evaluation. In all conditions except one, the optimization process converged to a mag-HRTF set that yielded localization errors significantly lower than the two non-individual HRTF sets under test, i.e. the training set's average and the HRTF set of the Neumann KU-100 manikin. The final localization error tended to decrease with the number of PCs, notably for the ARI dataset: the final median QE varied from 15 % to 7.5 % for 3 to 40 PCs. In comparison, for the same dataset, the median QEs for the average and KU-100 HRTF sets were 23 % and 33 %, respectively, whereas it was 6.3 % for the individual HRTF sets. While the estimated duration of the tuning procedure was prohibitive when many PCs were used for tuning, it appeared feasible (about one or two hours) when only 3 or 5 were retained, while substantial yet more modest localization performance improvement could be obtained.

We thus put to the test this alleged feasibility by submitting the tuning procedure to 12 actual listeners. Based on the results of the previous tuning simulations, we used the mag-HRTF model trained on the ARI dataset, limited to its first 5 PCs. The results somewhat differed from the simulations. Indeed, we found that the proposed method allowed considerable and significant improvement in localization performance over non-individual conditions, up to a performance comparable to that of individual HRTF sets reported in the literature [Middlebrooks99b; Middlebrooks00; Baumgartner14], with a

median quadrant error rate of 9.4 % for the customized HRTF sets. In comparison, the two non-individual conditions, i.e. the average and KU-100 HRTF sets, yielded respective median QEs of 34 % and 41 %, respectively.

Comparing our results to that of other HRTF individualization techniques, it appears that our proposed method improves localization performance substantially more than the selection of a best-fit non-individual HRTF set among 7 representatives HRTF sets [Katz12]. It however takes longer: one hour (in median) against 25 min [Zagala20]. The proposed approach also appears to reduce localization errors more than the self-tuning of a global frequency scaling parameter (in order to adapt a non-individual HRTF set to the user) [Middlebrooks00] in a comparable amount of time. Finally, our procedure yields a reduction in front-back confusions comparable to that of a related method which consists in tuning by ear the parameters of an HRTF PCA model in the median plane [Hwang08a; Shin08]. The duration of the tuning sessions was not reported in the latter studies, but should be expected to be considerably higher than ours: in contrast with our global approach, their tuning procedure must be performed at each direction of the HRTF set.

Although the main weakness of the proposed approach is the duration of a tuning session – one hour is far too high for a practical consumer-grade application, it can be largely decreased at the cost of a minimal increase in localization error. Indeed, within the first 20 minutes (in median), 88 % of the total decrease in median QE and and 92 % of the total decrease in APE was already achieved. Even though such a duration is not negligible, a playful calibration phase (in the form of a small video game, for instance) may very well make it acceptable, if not fun, to the end-user. Furthermore, this duration is of the same order as that of one of the simplest perceptual-feedback-based methods, i.e. selecting a non-individual HRTF set among a representative subset, while yielding substantially better localization performance.

Dimensionality reduction and data augmentation of HRTFs

As mentioned above, HRTF sets are a high-dimensionality data. It is thus highly desirable for the aforementioned approach – and many other statistical model-based ones – to reduce the dimensionality of the problem, i.e. the inter-individual variations of HRTF sets. In Chapter 3, we investigated this matter of HRTF dimensionality reduction and data augmentation.

In particular, in Section 3.2, we studied the dimensionality reduction performance

of PCA on log-magnitude HRTF sets from 9 datasets including FAST, using an *inter-individual* approach that has barely been touched on in the literature. Corroborating the initial observation that current HRTF datasets are small compared to the dimensionality of the data, we found that they are indeed too small to be representative of log-magnitude HRTF sets in general, which constitutes another contribution of this thesis.

In Section 3.3, we turned to 3-D morphology, and compared the respective dimensionality reduction performances of PCA on ear point clouds and on log-magnitude PRTF sets computed from them. We found that PCA performs considerably better at reducing the dimensionality of the former. Based on this, we presented in Section 3.4 a data augmentation process that allows the generation of an arbitrarily large synthetic dataset of PRTFs by means of random 3-D ear shapes generations and FM-BEM numerical simulations. The resulting dataset, named WiDESPREaD[2], comprises over a thousand registered pinna meshes and matching computed PRTF sets, and is freely available online[3]. This work constitutes one of the major contributions of this thesis and was published in the Journal of the Acoustical Society of America [Guezenoc20a].

In Section 3.5, the dimensionality reduction performance of PCA on WiDESPREaD log-magnitude PRTF sets was compared to that of other datasets. We found that such a model generalizes much better to new data, suggesting that a satisfactory number of examples was reached by means of 3-D morphology-based data augmentation. This final contribution was published and presented at the 148$^{\text{th}}$ Audio Engineering Society Convention [Guezenoc20b].

Perspectives

Despite the progress that has been made during this thesis, much work remains to be done towards HRTF individualization for the public. In particular, the approach that we proposed in Chapter 4 – tuning a statistical model of HRTF set based on localization performance – is a proof of concept that ought to be taken further.

Beyond the median plane Firstly, for the HRTF sets produced by our proposed method to be audible at directions beyond the median plane, it needs to include ITD. This could be done, for example, by tuning an ITD model based on lateral localization

[2] A Wide Dataset of Ear Shapes and Pinna-related transfer functions based on Random Ear Drawings
[3] `https://sofacoustics.org/data/database/widespread/`

error, alongside the magnitude HRTF model.

Furthermore, while the proposed approach produces a whole-sphere magnitude HRTF set, it does so based only on median-plane localization performance. It thus remains to be determined if and how well these magnitude HRTFs generalize to other cones of confusions in terms of intra-conic localization performance.

Regardless, in the future, the tuning should be based on positions beyond the median plane, in order to tune an ITD model and to possibly improve localization performance within lateral cones of confusions. Hence, in order to limit tuning time, an important subject of study should be the identification of a minimal viable spatial sampling of the sphere for the tuning procedure to work.

Further perceptual assessment To further establish the relevance of our proposed method, performing a separate subjective study may be desirable. Indeed, the localization tasks performed throughout the procedure were constrained in terms of duration and allowed a limited number of repetitions of the stimuli and test directions. Furthermore, this would allow us to evaluate the customized HRTF sets at positions other than the ones the tuning was based on. In particular, as discussed in the previous paragraph, provided that we include the listener's ITD, we could evaluate how the tuning generalizes to lateral cones of confusions.

On another level, the large improvement in localization performance that we observed in this work may be partly due to factors other than HRTF set customization itself, such as listener training throughout the tuning procedure. While this would constitute a positive side-effect, it may be interesting to investigate the existence and part of such an effect in localization performance improvement throughout the tuning session. Furthermore, evaluating the customized HRTF sets in a separate subjective study might allow us to avoid potential biases due, for example, to the Nelder-Mead algorithm.

HRTF model Finally, the model of magnitude HRTF set may be further improved. For instance, auto-encoder neural networks may be able to encode the inter-individual variations of magnitude HRTF sets into fewer parameters than PCA, resulting in a lower tuning duration. This would be the case, for instance, if the magnitude HRTF sets spanned a non-linear manifold of their high-dimensional space.

A secondary but nonetheless interesting advantage of neural networks is that there is a lot of freedom in the choice of the error metric that underlies their training – unlike

PCA which is inherently based on the mean squared error. One could thus imagine using an HRTF metric that is based on psycho-acoustics, such as the average difference between positive gradients of magnitude spectra that underlies the Baumgartner model (see Section 2.2.3), for example.

However, neural networks generally require a lot of data and, as we have seen in Chapter 3, currently available HRTF datasets are small compared to the dimensionality of the data. In this regard, the method that we proposed for randomly generating PRTF sets – and the resulting 1000-example dataset – may prove useful. However, in the future, supplementing the PRTFs with the contribution of a head and torso remains an indispensable next step in order to obtain "listenable" HRTFs. Although this could be approximated *a posteriori* by means of structural composition [Algazi01b], the ideal solution would be to include a statistical shape model of the head and torso into the data generation process (at the cost of much additional computing power). Finally, seeing that the quality of state-of-the-art computed HRTFs is still in question (see Chapter 2, Section 2.3), potential upgrades to HRTF numerical simulation ought to be included in the approach.

One Last Perceptual Experiment

In order to address some of the points we raised concerning the perceptual evaluation of our HRTF individualization method in the perspectives, we carried out a final campaign of listening tests six months after the tuning experiment presented in Section 4.4 of Chapter 4. 11 subjects out of the 12 from the first experiment participated in this new round of perceptual evaluations.

Method

We performed a double-blind evaluation of the three HRTF set conditions: Initial (average HRTF set), KU-100 and Final 5 i.e. customized HRTF set.

In order to be able to test sound source directions beyond the median plane, the ITDs of the KU-100 HRTF set were injected into the two other HRTF sets – which were minimum-phase. These ITDs were estimated using a threshold of -10 dB relative to the maximum peak of the low-passed HRIRs (with a cut-off frequency of 3 kHz), an approach among the most perceptually relevant according to the work by Andreopoulou *et al.* [Andreopoulou17]. After this step, the three HRTF sets shared the same ITDs.

In this experiment, the HRTF sets were evaluated at 16 sound source directions, which were different from ones used in the tuning experiment, at the exception of the frontal and rear directions (elevations of 0 and respective azimuths of 0 and 180°). 8 were located in the median plane, with elevations of −15°, 0, 20°, 70°, 110°, 160°, 180° and 195°, while the other 8 were slightly lateralized, in the ±10 later-angle cones of confusion, equally distributed to the left and to the right of the listener. Their elevations (identical for both left and right cones of confusion) were −30°, 40°, 140° and 210°. Listeners were asked to report the slightly lateralized positions onto the median plane, using the same 2-D interface as in the first experiment.

In one localization task, 32 stimuli – two repetitions for each one of the 16 positions – were presented in random order. In each of two successive blocks of evaluation, the 3 HRTF set conditions were presented in random order.

Results

The localization errors for the three HRTF sets are reported in box plots in Figure 4.18.

A first observation we can make is that the localization performances associated with both non-individual conditions are in overall coherence with the results of the tuning experiment. Indeed, the median KU-100 and initial APEs are 73° and 74°, against 76° and 71° in the first experiment. Regarding QEs, the median QE of the initial condition is identical in both experiments (34 %), while the median QE of the KU-100 HRTF set is lower (not significantly) in the second experiment (30 % against 41 %). As to the median PEs, they are comparable to chance in all cases, which was somewhat expected since the exocentric method that we used for sound localization reporting is not the most accurate (see Section 4.4.1 of Chapter 4 and [Bahu16a, Chap. 4; Katz19, pp. 359-361]). Overall, it seems that there was little to no effect of training between the first experiment (in which the KU-100 and initial conditions were evaluated at the beginning of the tuning procedure) and the second experiment.

In contrast, when looking at the customized HRTF set condition, we observe that the median APE and QE (65° and 30 %, respectively) are significantly larger in the second experiment than in the first one (38° and 9.4 %, respectively). Furthermore, the median PE for the customized condition is significantly lower than chance in the first experiment, while it is close to chance (like the KU-100 and initial conditions in both experiments) in the second experiment. It thus seems that the systematic selection by the Nelder-Mead simplex algorithm of the solution with the lowest localization error led

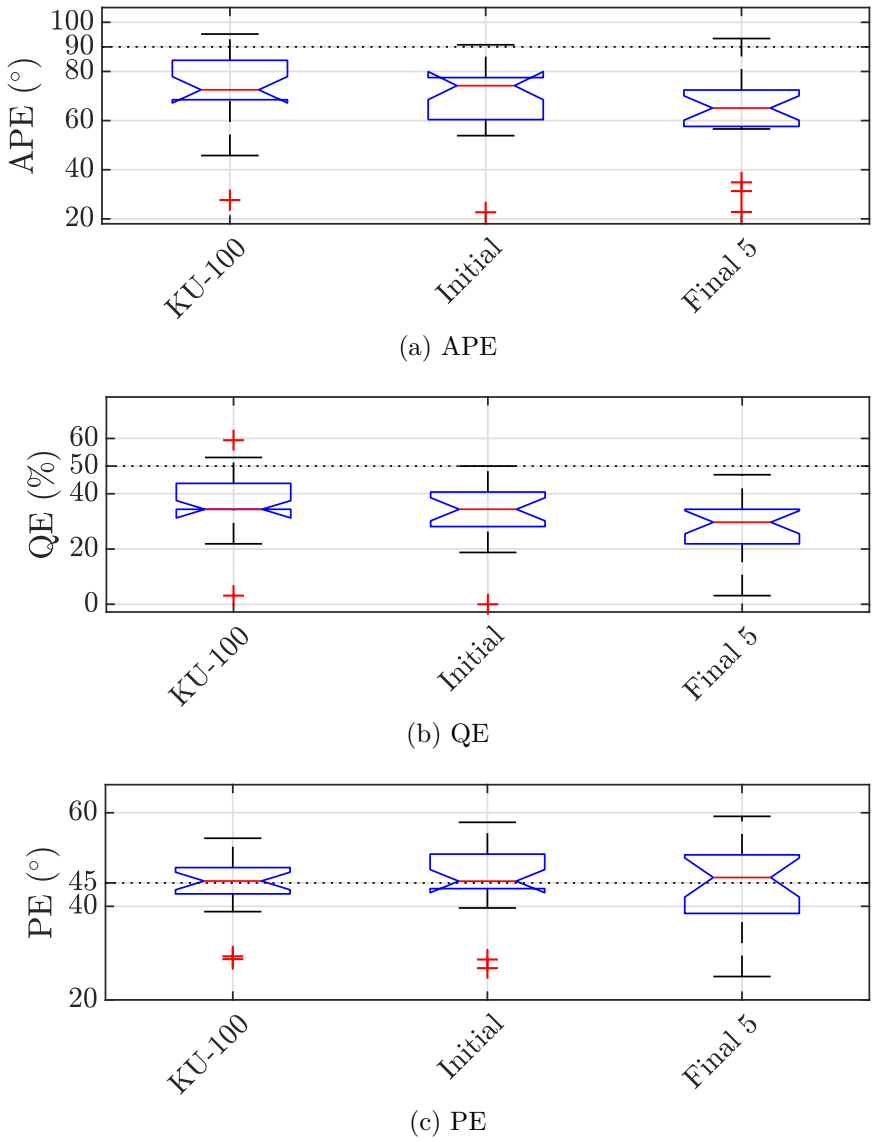
to an underestimation of that error in the first experiment, and that it had an influence on the statistical significance of the previously observed drop in localization error.

However, the overall trend is preserved: the localization error with the customized HRTF sets is lower than with the KU-100 or average HRTF sets. Indeed, the median APE and QE for the customized HRTF sets are 65° and 30 %, respectively, against 73° and 34 % for KU-100, and 74° and 34 % for the average HRTF set. Unlike in the first experiment, this difference is not statistically significant, due to considerable variability between subjects and blocks.

Further work

This last perceptual experiment highlighted the fact that the proposed HRTF tuning method can be further improved. In particular, it seems that the variability in listener answering should be taken into account in the optimization process, so as to avoid the selection of a good performance that could be more due to chance than to the HRTF set at hand itself.

Using another type of perceptual evaluation altogether might help to reduce variability in answering, to reduce tuning time, and to improve listener comfort. Indeed, locating a static non-reverberated sound signal within a cone of confusion is an arduous task, which was often reported by the participants in the listening experiments. An interesting alternative is a task that consists in rating a horizontal and/or vertical virtual trajectory. Indeed, it has been shown by [Zagala20] that ranking HRTF sets using this method correlates well to ranking HRTF sets based on localization tasks. Furthermore, judging a trajectory is arguably more playful and user-friendly than reporting the location of a number of static stimuli. However, in this type of listening tests, HRTF sets are evaluated relatively to one another, which would require to adapt the optimization process. For instance, the evaluations could take the form of A/B comparisons, which would likely feel easier and more comfortable to the listener than absolute judgments.



(a) APE

(b) QE

(c) PE

Figure 4.18 – Localization error outcome of the second perceptual evaluation: notched box plots of the APEs (top), QEs (middle) and PEs (bottom) of the baseline (KU-100), initial and final (Final 5) conditions. The horizontal dotted line shows the localization error associated with random answers.
On each box, the central red mark indicates the median, the bottom and top edges of the box the quartiles. Whiskers extend to the most extreme data points not considered as outliers, which are plotted as red crosses, and defined as the values that are away from the top or bottom of the box by more than 1.5 times the interquartile range. Two medians are significantly different at the 5 % significance level if their notches do not overlap [Mathworks18].

BIBLIOGRAPHY

[Algazi01a] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. "Elevation Localization and Head-Related Transfer Function Analysis at Low Frequencies". In: *The Journal of the Acoustical Society of America* 109.3 (Feb. 27, 2001), pp. 1110–1122. DOI: 10.1121/1.1349185.

[Algazi01b] V. Ralph Algazi, Richard O. Duda, Reed P. Morrison, and Dennis M. Thompson. "Structural Composition and Decomposition of HRTFs". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 20, 2001, pp. 103–106. DOI: 10.1109/ASPAA.2001.969553.

[Algazi01c] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. "The CIPIC HRTF Database". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Platz, NY, USA, 2001, pp. 99–102. DOI: 10.1109/ASPAA.2001.969552.

[Algazi02] V. Ralph Algazi, Richard O. Duda, Ramani Duraiswami, Nail A. Gumerov, and Zhihui Tang. "Approximating the Head-Related Transfer Function Using Simple Geometric Models of the Head and Torso". In: *The Journal of the Acoustical Society of America* 112.5 (Oct. 25, 2002), pp. 2053–2064. DOI: 10.1121/1.1508780.

[AlSheikh09] Bahaa W. Al-Sheikh, Mohammad A. Matin, and Daniel J. Tollin. "All-Pole and All-Zero Models of Human and Cat Head Related Transfer Functions". In: *Proceedings of SPIE 7444*. Vol. 7444. San Diego, CA, USA: International Society for Optics and Photonics, Aug. 2009, p. 74440X. DOI: 10.1117/12.829872.

[Andreopoulou11] Areti Andreopoulou and Agnieszka Roginska. "Towards the Creation of a Standardized HRTF Repository". In: *Proceedings of the*

131th Audio Engineering Society Convention. New York, NY, USA: Audio Engineering Society, 2011.

[Andreopoulou15] Areti Andreopoulou, Durand R. Begault, and Brian F. G. Katz. "Inter-Laboratory Round Robin HRTF Measurement Comparison". In: *IEEE Journal of Selected Topics in Signal Processing* 9.5 (Aug. 2015), pp. 895–906. DOI: 10.1109/JSTSP.2015.2400417.

[Andreopoulou16] Areti Andreopoulou and Brian F. G. Katz. "Investigation on Subjective HRTF Rating Repeatability". In: *Proceedings of the 140th Audio Engineering Society Convention*. Paris, France: Audio Engineering Society, June 4, 2016.

[Andreopoulou17] Areti Andreopoulou and Brian F. G. Katz. "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation". In: *The Journal of the Acoustical Society of America* 142.2 (Aug. 1, 2017), pp. 588–598. DOI: 10.1121/1.4996457.

[Armstrong18] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database". In: *Applied Sciences* 8.11 (Nov. 2018), p. 2029. DOI: 10.3390/app8112029.

[Asano90] Futoshi Asano, Yoiti Suzuki, and Toshio Sone. "Role of Spectral Cues in Median Plane Localization". In: *The Journal of the Acoustical Society of America* 88.1 (1990), pp. 159–168. DOI: 10.1121/1.399963.

[Augenbaum85] Jeffrey M. Augenbaum and Charles S. Peskin. "On the Construction of the Voronoi Mesh on a Sphere". In: *Journal of Computational Physics* 59.2 (1985), pp. 177–192. DOI: 10.1016/0021-9991(85)90140-8.

[Aytekin08] Murat Aytekin, Cynthia F. Moss, and Jonathan Z. Simon. "A Sensorimotor Approach to Sound Localization". In: *Neural Computation* 20.3 (Mar. 1, 2008), pp. 603–635. DOI: 10.1162/neco.2007.12-05-094.

[Bahu16a] Hélène Bahu. « Localisation auditive en contexte de synthèse binaurale non-individuelle [Auditory Localization in the Context of Non-Individual Binaural Synthesis] ». PhD Thesis. Universite Pierre et Marie Curie / IRCAM, 14 déc. 2016.

[Bahu16b] Hélène Bahu, Thibaut Carpentier, Markus Noisternig, and Olivier Warusfel. "Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments". In: *Acta Acustica united with Acustica* 102.1 (Jan. 1, 2016), pp. 107–118. DOI: 10.3813/AAA.918928.

[Barumerli20] Roberto Barumerli, Piotr Majdak, Jonas Reijniers, Robert Baumgartner, and Michele Geronazzo and Federico Avanzini. "Predicting Directional Sound-Localization of Human Listeners in Both Horizontal and Vertical Dimensions". In: *Proceedings of the 148th Audio Engineering Society Convention*. Vienna, Austria: Audio Engineering Society, May 28, 2020.

[Baskind12] Alexis Baskind, Thibaut Carpentier, Markus Noisternig, Olivier Warusfel, and Jean-Marc Lyzwa. "Binaural and Transaural Spatialization Techniques in Multichannel 5.1 Production". In: *Proceedings of the 27th Tonmeistertagung, VDT International Convention*. Köln, Germany, Nov. 2012.

[Baumgartner13] Robert Baumgartner, Piotr Majdak, and Bernhard Laback. "Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications". In: *The Technology of Binaural Listening*. Ed. by Jens Blauert. Springer, 2013, pp. 93–119. ISBN: 978-3-642-37761-7.

[Baumgartner14] Robert Baumgartner, Piotr Majdak, and Bernhard Laback. "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners". In: *The Journal of the Acoustical Society of America* 136.2 (Aug. 2014), pp. 791–802. DOI: 10.1121/1.4887447.

[Behnke12] Robert S. Behnke. *Kinetic Anatomy*. 3rd Edition. Human Kinetics, 2012. 329 pp. ISBN: 978-1-4504-1055-7.

[Beranek93] Leo L. Beranek. *Acoustical Measurements*. Revised Edition. Acoustical Society of America, 1993. 850 pp. ISBN: 0-88318-590-3.

[Bilinski14] Piotr Bilinski, Jens Ahrens, Mark R. P. Thomas, Ivan J. Tashev, and John C. Platt. “HRTF Magnitude Synthesis via Sparse Representation of Anthropometric Features”. In: *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy, 2014, pp. 4468–4472.

[Blauert97] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997. 514 pp. ISBN: 978-0-262-02413-6.

[Blauert98] Jens Blauert, Marc Brueggen, Adelbert W. Bronkhorst, Rob Drullman, Gerard Reynaud, Lionel Pellieux, Winfried Krebber, and Roland Sottek. “The AUDIS Catalog of Human HRTFs”. In: *The Journal of the Acoustical Society of America* 103.5 (May 1, 1998), pp. 3082–3082. DOI: 10.1121/1.422910.

[Blommer97] Michael A. Blommer and Gregory H. Wakefield. “Pole-Zero Approximations for Head-Related Transfer Functions Using a Logarithmic Error Criterion”. In: *IEEE Transactions on Speech and Audio Processing* 5.3 (1997), pp. 278–287.

[Bomhardt16a] Ramona Bomhardt, Hark Braren, and Janina Fels. “Individualization of Head-Related Transfer Functions Using Principal Component Analysis and Anthropometric Dimensions”. In: *Proceedings of the 172nd Meeting on Acoustics*. Vol. 29. Honolulu, HI, USA: Acoustical Society of America, Dec. 2016, p. 050007. DOI: 10.1121/2.0000562.

[Bomhardt16b] Ramona Bomhardt, Matias de la Fuente Klein, and Janina Fels. “A High-Resolution Head-Related Transfer Function and Three-Dimensional Ear Model Database”. In: *Proceedings of the 172nd Meeting on Acoustics*. Vol. 29. Honolulu, HI, USA: Acoustical Society of America, Nov. 28, 2016, p. 050002. DOI: 10.1121/2.0000467.

[Bomhardt16c] Ramona Bomhardt, Marcia Lins, and Janina Fels. “Analytical Ellipsoidal Model of Interaural Time Differences for the Individualization of Head-Related Impulse Responses”. In: *Journal of the Audio Engineering Society* 64.11 (2016), pp. 882–894.

[Bomhardt17] Ramona Bomhardt. “Anthropometric Individualization of Head-Related Transfer Functions Analysis and Modeling”. PhD Thesis. Aachen, Germany: Aachener Beiträge zur Akustik, 2017. 143 pp.

[Braren19] Hark Braren and Janina Fels. “Objective Differences between Individual HRTF Datasets of Children and Adults”. In: *Proceedings of the 23rd International Congress on Acoustics (ICA)*. Aachen, Germany, Sept. 9, 19, pp. 5220–5224.

[Breebaart01] Jeroen Breebaart and Armin Kohlrausch. “The Perceptual (Ir)Relevance of HRTF Magnitude and Phase Spectra”. In: *Proceedings of the 110th Audio Engineering Society Convention*. Amsterdam, Netherlands: Audio Engineering Society, May 12, 2001.

[Breebaart10] Jeroen Breebaart, Fabian Nater, and Armin Kohlrausch. “Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing”. In: *Journal of the Audio Engineering Society* 58.3 (Apr. 3, 2010), pp. 126–140.

[Brinkmann17] Fabian Brinkmann, Alexander Lindau, Stefan Weinzerl, Steven van de Par, Markus Müller-Trapet, Rob Opdam, and Michael Vorländer. “A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations”. In: *Journal of the Audio Engineering Society* 65.10 (Oct. 30, 2017), pp. 841–848. DOI: 10.17743/jaes.2017.0033.

[Brinkmann19] Fabian Brinkmann, Manoj Dinakaran, Robert Pelzer, Peter Grosche, Daniel Voss, and Stefan Weinzierl. “A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses”. In: *Journal of the Audio Engineering Society* 67.9 (Sept. 21, 2019), pp. 705–718. DOI: 10.17743/jaes.2019.0024.

[Bronkhorst95] Adelbert W. Bronkhorst. “Localization of Real and Virtual Sound Sources”. In: *The Journal of the Acoustical Society of America* 98.5 (Nov. 1, 1995), pp. 2542–2553. DOI: 10.1121/1.413219.

[Busson06] Sylvain Busson. « Individualisation d’indices acoustiques pour la synthèse binaurale [Individualization of Acoustic Cues for Binaural Synthesis] ». PhD Thesis. Université de la Méditerranée-Aix-Marseille II, 2006.

[Carlile97] Simon Carlile, Philip Leong, and Stephanie Hyams. “The Nature and Distribution of Errors in Sound Localization by Human Listeners”. In: *Hearing Research* 114.1 (Dec. 1, 1997), pp. 179–196. DOI: 10.1016/S0378-5955(97)00161-5.

[Carlile98] Simon Carlile, Craig Jin, and Vaughn Harvey. “The Generation and Validation of High Fidelity Virtual Auditory Space”. In: *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 20. Hong Kong, China: IEEE, Nov. 1, 1998, pp. 1090–1095. DOI: 10.1109/IEMBS.1998.747061.

[Carpentier14] Thibaut Carpentier, Hélène Bahu, Markus Noisternig, and Olivier Warusfel. “Measurement of a Head-Related Transfer Function Database with High Spatial Resolution”. In: *Proceedings of the 7th Forum Acusticum*. Kraków, Poland: European Acoustics Association, Sept. 7, 2014.

[Casadamont18] Amandine Casadamont et Alexandre Plank. *Welcome to Nay Pyi Taw*. Hyperradio, France Culture / Deutschlandradio Kultur. 20 avr. 2018. URL : https://hyperradio.radiofrance.fr/son-3d/welcome-to-nay-pyi-taw/.

[Chen20] Wei Chen, Ruimin Hu, Xiaochen Wang, and Dengshi Li. “HRTF Representation with Convolutional Auto-Encoder”. In: *Proceedings of the 26th International Conference on Multimedia Modeling (MMM)*. Ed. by Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve. Lecture Notes in Computer Science. Seoul, South Korea: Springer International Publishing, Jan. 5, 2020, pp. 605–616. DOI: 10.1007/978-3-030-37731-1_49.

[Cooper89] Duane H. Cooper and Jerald L. Bauck. “Prospects for Transaural Recording”. In: *Journal of the Audio Engineering Society* 37.1/2 (1989), pp. 3–19.

[Cootes95] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. “Active Shape Models - Their Training and Applica-

tion". In: *Computer Vision and Image Understanding* 61.1 (1995), pp. 38–59. DOI: 10.1006/cviu.1995.1004.

[Deleforge15] Antoine Deleforge, Florence Forbes, and Radu Horaud. "Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds". In: *International Journal of Neural Systems* 25.1 (Feb. 1, 2015), 21p. DOI: 10.1142/S0129065714400036.

[Denk17] Florian Denk, Jan Heeren, Stephan D. Ewert, Birger Kollmeier, and Stephan M.A. Ernst. "Controlling the Head Position During Individual HRTF Measurements and Its Effect on Accuracy". In: *Proceedings of the Annual German Conference on Acoustics (DAGA)*. Kiel, Germany, 2017.

[Dinakaran18] Manoj Dinakaran, Fabian Brinkmann, Stine Harder, Robert Pelzer, Peter Grosche, Rasmus R. Paulsen, and Stefan Weinzierl. "Perceptually Motivated Analysis of Numerically Simulated Head-Related Transfer Functions Generated By Various 3D Surface Scanning Systems". In: *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, Apr. 2018, pp. 551–555. DOI: 10.1109/ICASSP.2018.8461789.

[Duraiswami04] Ramani Duraiswami, Dmitry N. Zotkin, and Nail A. Gumerov. "Interpolation and Range Extrapolation of HRTFs". In: *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4. Montréal, QE, Canada, May 17, 2004, pp. 45–48. DOI: 10.1109/ICASSP.2004.1326759.

[Duraiswami05] Ramani Duraiswami and Vikas C. Raykar. "The Manifolds of Spatial Hearing". In: *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 3. Philadelphia, PA, USA: IEEE, 2005, pp. iii/285–iii/288. DOI: 10.1109/ICASSP.2005.1415702.

[Durant02] E. A. Durant and G. H. Wakefield. "Efficient Model Fitting Using a Genetic Algorithm: Pole-Zero Approximations of HRTFs". In: *IEEE Transactions on Speech and Audio Processing* 10.1 (Jan. 2002), pp. 18–27. DOI: 10.1109/89.979382.

[Ehret78] Günter Ehret. “Stiffness Gradient along the Basilar Membrane as a Basis for Spatial Frequency Analysis within the Cochlea”. In: *The Journal of the Acoustical Society of America* 64.6 (Dec. 1, 1978), pp. 1723–1726. DOI: 10.1121/1.382153.

[Engel19] Isaac Engel, David Lou Alon, Philip W. Robinson, and Ravish Mehra. “The Effect of Generic Headphone Compensation on Binaural Renderings”. In: *Proceedings of the 2019 AES International Conference on Immersive and Interactive Audio*. York, UK: Audio Engineering Society, Mar. 17, 2019.

[Enzner08] Gerald Enzner. “Analysis and Optimal Control of LMS-Type Adaptive Filtering for Continuous-Azimuth Acquisition of Head Related Impulse Responses”. In: *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Las Vegas, NV, USA: IEEE, 2008, pp. 393–396. DOI: 10.1109/ICASSP.2008.4517629.

[Fan19] Ziqi Fan, Terek Arce, Chenshen Lu, Kai Zhang, T. W. Wu, and Kyla McMullen. “Computation of Head-Related Transfer Functions Using Graphics Processing Units and a Perceptual Validation of the Computed HRTFs against Measured HRTFs”. In: *Proceedings of the 2019 Audio Engineering Society International Conference on Headphone Technology*. San Francisco, CA, USA: Audio Engineering Society, Aug. 27, 2019.

[Farahikia17] Mahdi Farahikia and Quang T. Su. “Optimized Finite Element Method for Acoustic Scattering Analysis With Application to Head-Related Transfer Function Estimation”. In: *Journal of Vibration and Acoustics* 139.3 (June 2017), p. 034501. DOI: 10.1115/1.4035813.

[Fayek17] Haytham Fayek, Laurens van der Maaten, Griffin Romigh, and Ravish Mehra. “On Data-Driven Approaches to Head-Related-Transfer Function Personalization”. In: *Proceedings of the 143rd Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 8, 2017.

[Fink15] Kimberly J. Fink and Laura Ray. “Individualization of Head Related Transfer Functions Using Principal Component Analysis”. In: *Applied Acoustics* 87 (Jan. 2015), pp. 162–173. DOI: 10.1016/j.apacoust.2014.07.005.

[Furness90] Roger K. Furness. “Ambisonics - An Overview”. In: *Proceedings of the 8th AES International Conference on The Sound of Audio*. Washington D.C., USA: Audio Engineering Society, May 1, 1990.

[Gardner97] William Grant Gardner. “3-D Audio Using Loudspeakers”. PhD Thesis. Massachusetts Institute of Technology, Sept. 1997.

[Geronazzo18] M. Geronazzo, S. Spagnol, and F. Avanzini. “Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Apr. 2, 2018), pp. 1247–1260. DOI: 10.1109/TASLP.2018.2821846.

[Ghorbal16] Slim Ghorbal, Renaud Séguier, and Xavier Bonjour. “Process of HRTF Individualization by 3D Statistical Ear Model”. In: *Proceedings of the 141st Audio Engineering Society Convention*. Los Angeles, CA, USA: Audio Engineering Society, Sept. 20, 2016.

[Ghorbal17] Slim Ghorbal, Théo Auclair, Catherine Soladié, and Renaud Séguier. “Pinna Morphological Parameters Influencing HRTF Sets”. In: *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*. Edinburgh, UK, Sept. 9, 2017.

[Ghorbal19] Slim Ghorbal, Renaud Séguier, and Xavier Bonjour. “Method for Establishing a Deformable 3D Model of an Element, and Associated System”. U.S. pat. Patent 16/300, 044. May 16, 2019.

[Ghorbal20] Slim Ghorbal. « Personnalisation de l’écoute binaurale par modèle déformable d’oreille [Personnalization of Binaural Listening by means of a Deformable Ear Model] ». PhD Thesis. To be published : CentraleSupélec, 2020.

[Glasberg90] Brian R. Glasberg and Brian C. J. Moore. “Derivation of Auditory Filter Shapes from Notched-Noise Data”. In: *Hearing Research* 47.1-2 (Aug. 1990), pp. 103–138. DOI: 10.1016/0378-5955(90)90170-T.

[Gower75] John C. Gower. “Generalized Procrustes Analysis”. In: *Psychometrika* 40.1 (Mar. 1975), pp. 33–51.

[Greff07] Raphaël Greff and Brian F. G. Katz. “Round Robin Comparison of HRTF Simulation Systems: Preliminary Results”. In: *Proceedings of the 123rd Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 1, 2007.

[Grijalva14] Felipe Grijalva, Luiz Martini, Siome Goldenstein, and Dinei Florencio. “Anthropometric-Based Customization of Head-Related Transfer Functions Using Isomap in the Horizontal Plane”. In: *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 4473–4477. DOI: 10.1109/ICASSP.2014.6854448.

[Grijalva16] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein. “A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.3 (Mar. 2016), pp. 559–570. DOI: 10.1109/TASLP.2016.2517565.

[Guezenoc18] Corentin Guezenoc and Renaud Séguier. “HRTF Individualization: A Survey”. In: *Proceedings of the 145th Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 7, 2018. DOI: 10.17743/aesconv.2018.978-1-942220-25-1.

[Guezenoc20a] Corentin Guezenoc and Renaud Séguier. “A Wide Dataset of Ear Shapes and Pinna-Related Transfer Functions Generated by Random Ear Drawings”. In: *The Journal of the Acoustical Society of America* 147.6 (June 23, 2020), pp. 4087–4096. DOI: 10.1121/10.0001461.

[Guezenoc20b] Corentin Guezenoc and Renaud Séguier. “Dataset Augmentation and Dimensionality Reduction of Pinna-Related Transfer Functions”. In: *Proceedings of the 148th Audio Engineering Society Convention*. Vienna, Austria: Audio Engineering Society, May 28, 2020. DOI: 10.17743/aesconv.2020.978-1-942220-32-9.

[Guillon08] Pierre Guillon, Rozenn Nicol, and Laurent 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 Audio Engineering Society Convention*. San Francisco, CA, USA: Audio Engineering Society, Oct. 1, 2008.

[Gumerov05] Nail A. Gumerov and Ramani Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier Series in Electromagnetism. Elsevier Science, Jan. 27, 2005. 426 pp. ISBN: 978-0-08-053159-5.

[Gumerov07] Nail A. Gumerov, Ramani Duraiswami, and Dmitry N. Zotkin. “Fast Multipole Accelerated Boundary Elements for Numerical Computation of the Head Related Transfer Function”. In: *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Honolulu, HI, USA: IEEE, 2007, pp. I-165–I-168. DOI: 10.1109/ICASSP.2007.366642.

[Gumerov10] Nail A. Gumerov, Adam E. O’Donovan, Ramani Duraiswami, and Dmitry N. Zotkin. “Computation of the Head-Related Transfer Function via the Fast Multipole Accelerated Boundary Element Method and Its Spherical Harmonic Representation”. In: *The Journal of the Acoustical Society of America* 127.1 (Jan. 2010), pp. 370–386. DOI: 10.1121/1.3257598.

[Haneda99] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. “Common-Acoustical-Pole and Zero Modeling of Head-Related Transfer Functions”. In: *IEEE Transactions on Speech and Audio Processing* 7.2 (Mar. 1999), pp. 188–196. DOI: 10.1109/89.748123.

[Hebrank74] Jack Hebrank and Donald Wright. “Spectral Cues Used in the Localization of Sound Sources on the Median Plane”. In: *The Journal of the Acoustical Society of America* 56.6 (Dec. 1, 1974), pp. 1829–1834. DOI: 10.1121/1.1903520.

[Hirahara10] Tatsuya Hirahara, Hiroyuki Sagara, Iwaki Toshima, and Makoto Otani. “Head Movement during Head-Related Transfer Function Measurements”. In: *Acoustical Science and Technology* 31.2 (2010), pp. 165–171. DOI: 10.1250/ast.31.165.

[Hoffmann08] Pablo F. Hoffmann and Henrik Møller. “Audibility of Differences in Adjacent Head-Related Transfer Functions”. In: *Acta Acustica united with Acustica* 94.6 (Nov. 1, 2008), pp. 945–954. DOI: 10.3813/AAA.918111.

[Hold17] Christoph Hold, Fabian Seipel, Fabian Brinkmann, Athanasios Lykartsis, and Stefan Weinzierl. “Eigen-Images of Head-Related Transfer Functions”. In: *Proceedings of the 143rd Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 8, 2017.

[Hölzl14] Josef Hölzl. “A Global Model for HRTF Individualization by Adjustment of Principal Component Weights”. Master Thesis. Graz, Austria: Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz University of Technology, 2014. 135 pp.

[Hu06] Hongmei Hu, Lin Zhou, Jie Zhang, Hao Ma, and Zhenyang Wu. “Head Related Transfer Function Personalization Based on Multiple Regression Analysis”. In: *Proceedings of the 2006 International Conference on Computational Intelligence and Security*. Vol. 2. Guangzhou, China, Nov. 2006, pp. 1829–1832. DOI: 10.1109/ICCIAS.2006.295380.

[Hu08] Hongmei Hu, Lin Zhou, Hao Ma, and Zhenyang Wu. “HRTF Personalization Based on Artificial Neural Network in Individual Virtual Auditory Space”. In: *Applied Acoustics* 69.2 (Feb. 2008), pp. 163–172. DOI: 10.1016/j.apacoust.2007.05.007.

[Hu16] Shichao Hu, Jorge Trevino, Cesar Salvador, Shuichi Sakamoto, Junfeng Li, and Yôiti Suzuki. “A Local Representation of the Head-Related Transfer Function”. In: *The Journal of the Acoustical Society of America* 140.3 (Sept. 21, 2016), EL285–EL290. DOI: 10.1121/1.4962805.

[Hu19] Shichao Hu, Jorge Trevino, César Salvador, Shuichi Sakamoto, and Yôiti Suzuki. “Modeling Head-Related Transfer Functions with Spherical Wavelets”. In: *Applied Acoustics* 146 (Mar. 1, 2019), pp. 81–88. DOI: 10.1016/j.apacoust.2018.10.026.

[Huang09a] Qinghua Huang and Yong Fang. “Modeling Personalized Head-Related Impulse Response Using Support Vector Regression”. In: *Journal of Shanghai University (English Edition)* 13.6 (2009), p. 428. DOI: 10.1007/s11741-009-0602-2.

[Huang09b] Qinghua Huang and Qi-lei Zhuang. “HRIR Personalisation Using Support Vector Regression in Independent Feature Space”. In: *Electronics Letters* 45.19 (Sept. 2009), pp. 1002–1003. DOI: 10.1049/el.2009.1865.

[Hugeng10] Hugeng Hugeng, Wahab Wahidin, and Dadag Gunawan. “A Novel Individualization of Head-Related Impulse Responses on Median Plane Using Listener’s Anthropometries Based On Multiple Regression Analysis”. In: *Jurnal Penelitian dan Pengembangan Telekomunikasi* 15.1 (June 2010).

[Huopaniemi99] Jyri Huopaniemi, Nick Zacharov, and Matti Karjalainen. “Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design”. In: *Journal of the Audio Engineering Society* 47.4 (Apr. 1, 1999), pp. 218–239.

[Huttunen07] Tomi Huttunen, Eira T. Seppälä, Ole Kirkeby, Asta Kärkkäinen, and Leo Kärkkäinen. “Simulation of the Transfer Function for a Head-and-Torso Model over the Entire Audible Frequency Range”. In: *Journal of Computational Acoustics* 15.04 (Dec. 1, 2007), pp. 429–448. DOI: 10.1142/S0218396X07003469.

[Huttunen13] Tomi Huttunen, Kimmo Tuppurainen, Antti Vanne, Pasi Ylä-Oijala, Seppo Järvenpää, Asta Kärkkäinen, and Leo Kärkkäinen. “Simulation of the Head-Related Transfer Functions Using Cloud Computing”. In: *Proceedings of the 21st International Congress on Acoustics (ICA)*. Vol. 19. Montréal, QE, Canada: Acoustical Society of America, June 2, 2013, p. 050168. DOI: 10.1121/1.4800138.

[Hwang08a] Sungmok Hwang, Youngjin Park, and Youn-sik Park. “Modeling and Customization of Head-Related Impulse Responses Based on General Basis Functions in Time Domain”. In: *Acta Acustica united with Acustica* 94.6 (Nov. 1, 2008), pp. 965–980. DOI: 10.3813/AAA.918113.

[Hwang08b] Sungmok Hwang, Youngjin Park, and Youn-sik Park. “Modeling and Customization of Head-Related Transfer Functions Using Principal Component Analysis”. In: *Proceedings of the 2008 International Conference on Control, Automation and Systems (ICCAS 2008)*. Seoul, South Korea: IEEE, 2008, pp. 227–231. DOI: 10.1109/ICCAS.2008.4694554.

[Inoue05] Naoya Inoue, Toshiyuki Kimura, Takanori Nishino, Katsunobu Itou, and Kazuya Takeda. “Evaluation of HRTFs Estimated Using Physical Features”. In: *Acoustical Science and Technology* 26.5 (Apr. 6, 2005), pp. 453–455. DOI: 10.1250/ast.26.453.

[Iwaya06] Yukio Iwaya. “Individualization of Head-Related Transfer Functions with Tournament-Style Listening Test: Listening with Other’s Ears”. In: *Acoustical Science and Technology* 27.6 (2006), pp. 340–343. DOI: http://dx.doi.org/10.1250/ast.27.340.

[Jin00] Craig Jin, Philip H. W. Leong, Johahn Leung, Anna Corderoy, and Simon Carlile. “Enabling Individualized Virtual Auditory Space Using Morphological Measurements”. In: *Proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia*. Sydney, NSW, Australia: IEEE, 2000, pp. 235–238.

[Jin14] Craig Jin, Pierre Guillon, Nicolas Epain, Reza Zolfaghari, André van Schaik, Anthony I. Tew, Carl Hetherington, and Jonathan Thorpe. “Creating the Sydney York Morphological and Acoustic Recordings of Ears Database”. In: *IEEE Transactions on Multimedia* 16.1 (Jan. 2014), pp. 37–46. DOI: 10.1109/TMM.2013.2282134.

[Jolliffe02] Ian T. Jolliffe. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. Springer-Verlag, 2002. ISBN: 978-0-387-95442-4.

[Kahana06] Yuvi Kahana and Philip A. Nelson. “Numerical Modelling of the Spatial Acoustic Response of the Human Pinna”. In: *Journal of Sound and Vibration* 292.1-2 (Oct. 2006), pp. 148–178. DOI: 10.1016/j.jsv.2005.07.048.

[Kahana99] Yuvi Kahana, Philip A. Nelson, Maurice Petyt, and Sunghoon Choi. “Numerical Modelling of the Transfer Functions of a Dummy-Head and of the External Ear”. In: *Proceedings of the 16th AES International Conference on Spatial Sound Reproduction*. Rovaniemi, Finland: Audio Engineering Society, Apr. 10, 1999.

[Kaneko16a] Shoken Kaneko, Tsukasa Suenaga, Mai Fujiwara, Kazuya Kumehara, Futoshi Shirakihara, and Satoshi Sekine. “Ear Shape Modeling for 3D Audio and Acoustic Virtual Reality: The Shape-Based Average HRTF”. In: *Proceedings of the 61st AES International Conference on Audio for Games*. London, UK: Audio Engineering Society, Feb. 10, 2016. ISBN: 978-1-942220-08-4.

[Kaneko16b] Shoken Kaneko, Tsukasa Suenaga, and Satoshi Sekine. “DeepEar-Net: Individualizing Spatial Audio with Photography, Ear Shape Modeling, and Neural Networks”. In: *Proceedings of the 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Los Angeles, CA, USA: Audio Engineering Society, Sept. 30, 2016.

[Kapralos08] Bill Kapralos, Nathan Mekuz, Agnieszka Kopinska, and Saad Khattak. “Dimensionality Reduced HRTFs: A Comparative Study”. In: *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology (ACE)*. Yokohama, Japan: Association for Computing Machinery, Dec. 5, 2008, pp. 59–62. DOI: 10.1145/1501750.1501763.

[Katz00] Brian F. G. Katz. “Acoustic Absorption Measurement of Human Hair and Skin within the Audible Frequency Range”. In: *The Journal of the Acoustical Society of America* 108.5 (Nov. 1, 2000), pp. 2238–2242. DOI: 10.1121/1.1314319.

[Katz01] Brian F. G. Katz. “Boundary Element Method Calculation of Individual Head-Related Transfer Function. I. Rigid Model Calculation”. In: *The Journal of the Acoustical Society of America* 110.5 (Oct. 29, 2001), pp. 2440–2448. DOI: 10.1121/1.1412440.

[Katz12] Brian F. G. Katz and Gaëtan Parseihian. “Perceptually Based Head-Related Transfer Function Database Optimization”. In: *The Journal of the Acoustical Society of America* 131.2 (Jan. 13, 2012), EL99–EL105. DOI: 10.1121/1.3672641.

[Katz14] Brian F. G. Katz and Markus Noisternig. “A Comparative Study of Interaural Time Delay Estimation Methods”. In: *The Journal of the Acoustical Society of America* 135.6 (June 1, 2014), pp. 3530–3540. DOI: 10.1121/1.4875714.

[Katz19] Brian F. G. Katz and Rozenn Nicol. “Binaural Spatial Reproduction”. In: *Sensory Evaluation of Sound*. Ed. by Nick Zacharov. CRC Press, 2019. ISBN: 978-0-429-76991-7.

[Kearney15] Gavin Kearney and Tony Doyle. “An HRTF Database for Virtual Loudspeaker Rendering”. In: *Proceedings of the 139th Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 2015.

[Kim05] Sang-Myeong Kim and Wonjae Choi. “On the Externalization of Virtual Sound Images in Headphone Reproduction: A Wiener Filter Approach”. In: *The Journal of the Acoustical Society of America* 117.6 (May 31, 2005), pp. 3657–3665. DOI: 10.1121/1.1921548.

[Kimura14] Masateru Kimura, Jason Kunio, Andreas Schuhmacher, and Yunseon Ryu. “A New High-Frequency Impedance Tube for Measuring Sound Absorption Coefficient and Sound Transmission Loss”. In: *Proceedings of Inter-Noise*. Melbourne, Australia: Institute of Noise Control Engineering, Nov. 16, 2014.

[Kistler92] Doris J. Kistler and Frederic L. Wightman. “A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction”. In: *The Journal of the Acoustical Society of America* 91.3 (Mar. 1, 1992), pp. 1637–1647. DOI: 10.1121/1.402444.

[Kreuzer09] Wolfgang Kreuzer, Piotr Majdak, and Zhengsheng Chen. “Fast Multipole Boundary Element Method to Calculate Head-Related Transfer Functions for a Wide Frequency Range”. In: *The Journal of the Acoustical Society of America* 126.3 (Sept. 1, 2009), pp. 1280–1290. DOI: 10.1121/1.3177264.

[KRoll18] Kristoff K.Roll. *Petite Suite A l'Ombre des Ondes*. Avec la coll. de Valérie Lavallart, Laure Jung-Lancrey, Francesco Cameli, Claire Bergerault, Isabelle Duthoit, Patrice Soletti, Didier Aschour et Edward Perraud. Hyperradio, Radio France. 10 déc. 2018. URL : https://hyperradio.radiofrance.fr/son-3d/creation-mondiale-kristoff-k-roll-petites-suites-a-lombres-des-ondes-dans-la-bibliotheque-de-recits-de-reves/.

[Kuhn77] George F. Kuhn. “Model for the Interaural Time Differences in the Azimuthal Plane”. In: *The Journal of the Acoustical Society of America* 62.1 (July 1, 1977), pp. 157–167. DOI: 10.1121/1.381498.

[Kulkarni99] Abhijit Kulkarni, Scott K. Isabelle, and H. Steven Colburn. “Sensitivity of Human Subjects to Head-Related Transfer-Function Phase Spectra”. In: *The Journal of the Acoustical Society of America* 105.5 (1999), pp. 2821–2840. DOI: 10.1121/1.426898.

[Lagarias98] Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright, and Paul E. Wright. “Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions”. In: *SIAM Journal on Optimization* 9.1 (Jan. 1, 1998), pp. 112–147. DOI: 10.1137/S1052623496303470.

[Langendijk02] Erno H. A. Langendijk and Adelbert W. Bronkhorst. “Contribution of Spectral Cues to Human Sound Localization”. In: *The Journal of the Acoustical Society of America* 112.4 (Sept. 27, 2002), pp. 1583–1596. DOI: 10.1121/1.1501901.

[Langendijk99] Erno H. A. Langendijk and Adelbert W. Bronkhorst. “Fidelity of Three-Dimensional-Sound Reproduction Using a Virtual Auditory Display”. In: *The Journal of the Acoustical Society of America* 107.1 (Dec. 29, 1999), pp. 528–537. DOI: 10.1121/1.428321.

[Larcher00] Véronique Larcher, Olivier Warusfel, Jean-Marc Jot, and Jérome Guyard. "Study and Comparison of Efficient Methods for 3D Audio Spatialization Based on Linear Decomposition of HRTF Data". In: *Proceedings of the 108th Audio Engineering Society Convention*. Paris, France: Audio Engineering Society, Feb. 19, 2000.

[Larcher97] Véronique LARCHER et Jean-Marc JOT. « Techniques d'interpolation de filtres audio-numériques : Application à la reproduction spatiale des sons sur écouteurs [Interpolation Techniques for Audio-Digital Filters : Application to Sound Spatial Reproduction through Ear-buds] ». In : *Actes du Congrès Français d'Acoustique (CFA) 1997*. Marseille, France, avr. 1997, p. 97-100.

[Le Bagousse10] Sarah Le Bagousse, Catherine Colomes, and Mathieu Paquier. "State of the Art on Subjective Assessment of Spatial Sound Quality". In: *Proceedings of the 38th AES International Conference on Sound Quality Evaluation*. Piteå, Sweden: Audio Engineering Society, June 13, 2010.

[Li13] Lin Li and Qinghua Huang. "HRTF Personalization Modeling Based on RBF Neural Network". In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 3707–3710. DOI: 10.1109/ICASSP.2013.6638350.

[Liu19a] Huaping Liu, Yong Fang, and Qinghua Huang. "Efficient Representation of Head-Related Transfer Functions With Combination of Spherical Harmonics and Spherical Wavelets". In: *IEEE Access* 7 (June 27, 2019), pp. 78214–78222. DOI: 10.1109/ACCESS.2019.2921388.

[Liu19b] Xuejie Liu, Hao Song, and Xiaoli Zhong. "A Hybrid Algorithm for Predicting Median-Plane Head-Related Transfer Functions from Anthropometric Measurements". In: *Applied Sciences* 9.11 (11 Jan. 2019), p. 2323. DOI: 10.3390/app9112323.

[Macpherson02] Ewan A. Macpherson and John C. Middlebrooks. "Listener Weighting of Cues for Lateral Angle: The Duplex Theory of Sound Localiza-

tion Revisited". In: *The Journal of the Acoustical Society of America* 111.5 (May 1, 2002), pp. 2219–2236. DOI: 10.1121/1.1471898.

[Macpherson07] Ewan A. Macpherson and Andrew T. Sabin. "Binaural Weighting of Monaural Spectral Cues for Sound Localization". In: *The Journal of the Acoustical Society of America* 121.6 (June 1, 2007), pp. 3677–3688. DOI: 10.1121/1.2722048.

[Majdak07] Piotr Majdak, Peter Balazs, and Bernhard Laback. "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions". In: *Journal of the Audio Engineering Society* 55.7/8 (July 2007), pp. 623–637.

[Majdak10] Piotr Majdak, Matthew J. Goupell, and Bernhard Laback. "3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training". In: *Attention, Perception & Psychophysics* 72.2 (Feb. 1, 2010), pp. 454–469. DOI: 10.3758/APP.72.2.454.

[Maki05] Katuhiro Maki and Shigeto Furukawa. "Reducing Individual Differences in the External-Ear Transfer Functions of the Mongolian Gerbil". In: *The Journal of the Acoustical Society of America* 118.4 (Oct. 1, 2005), pp. 2392–2404. DOI: 10.1121/1.2033571.

[Mäkivirta20] Aki Mäkivirta, Matti Malinen, Jaan Johansson, Ville Saari, and Aapo Karjalainen and Poorang Vosough. "Accuracy of Photogrammetric Extraction of the Head and Torso Shape for Personal Acoustic HRTF Modeling". In: *Proceedings of the 148th Audio Engineering Society Convention*. Vienna, Austria: Audio Engineering Society, May 28, 2020.

[Marburg02] Steffen Marburg. "Six Boundary Elements per Wavelength: Is That Enough?" In: *Journal of Computational Acoustics* 10.01 (Mar. 1, 2002), pp. 25–51. DOI: 10.1142/S0218396X02001401.

[Martin01] Russell L. Martin, Ken I. McAnally, and Melis A. Senova. "Free-Field Equivalent Localization of Virtual Audio". In: *Journal of the Audio Engineering Society* 49.1/2 (Feb. 1, 2001), pp. 14–22.

[Mathworks18] Mathworks. *MATLAB Statistics and Machine Learning Toolbox Release 2018b: User's Guide*. 2018. URL: https://fr.mathworks.com/help/stats/boxplot.html.

[Matsunaga10] Noriyuki Matsunaga and Tatsuya Hirahara. "Reexamination of Fast Head-Related Transfer Function Measurement by Reciprocal Method". In: *Acoustical Science and Technology* 31.6 (2010), pp. 414–416. DOI: 10.1250/ast.31.414.

[Mehrgardt77] S. Mehrgardt and V. Mellert. "Transformation Characteristics of the External Human Ear". In: *The Journal of the Acoustical Society of America* 61.6 (June 1, 1977), pp. 1567–1576. DOI: 10.1121/1.381470.

[Meshram14] Alok Meshram, Ravish Mehra, Hongsheng Yang, Enrique Dunn, Jan-Michael Franm, and Dinesh Manocha. "P-HRTF: Efficient Personalized HRTF Computation for High-Fidelity Spatial Sound". In: *Proceedings of the 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Munich, Germany: IEEE, Sept. 2014, pp. 53–61. DOI: 10.1109/ISMAR.2014.6948409.

[Middlebrooks00] John C. Middlebrooks, Ewan A. Macpherson, and Zekiye A. Onsan. "Psychophysical Customization of Directional Transfer Functions for Virtual Sound Localization". In: *The Journal of the Acoustical Society of America* 108.6 (Nov. 21, 2000), pp. 3088–3091. DOI: 10.1121/1.1322026.

[Middlebrooks90] John C. Middlebrooks and David M. Green. "Directional Dependence of Interaural Envelope Delays". In: *The Journal of the Acoustical Society of America* 87.5 (May 1, 1990), pp. 2149–2162. DOI: 10.1121/1.399183.

[Middlebrooks92] John C. Middlebrooks and David M. Green. "Observations on a Principal Components Analysis of Head-related Transfer Functions". In: *The Journal of the Acoustical Society of America* 92.1 (July 1, 1992), pp. 597–599. DOI: 10.1121/1.404272.

[Middlebrooks99a] John C. Middlebrooks. “Individual Differences in External-Ear Transfer Functions Reduced by Scaling in Frequency”. In: *The Journal of the Acoustical Society of America* 106.3 (Aug. 23, 1999), pp. 1480–1492. DOI: 10.1121/1.427176.

[Middlebrooks99b] John C. Middlebrooks. “Virtual Localization Improved by Scaling Nonindividualized External-Ear Transfer Functions in Frequency”. In: *The Journal of the Acoustical Society of America* 106.3 (Aug. 23, 1999), pp. 1493–1510. DOI: 10.1121/1.427147.

[Mills58] A. W. Mills. “On the Minimum Audible Angle”. In: *The Journal of the Acoustical Society of America* 30.4 (Apr. 1, 1958), pp. 237–246. DOI: 10.1121/1.1909553.

[Mills60] A. W. Mills. “Lateralization of High-Frequency Tones”. In: *The Journal of the Acoustical Society of America* 32.1 (Jan. 1, 1960), pp. 132–134. DOI: 10.1121/1.1907864.

[Möbius10] Jan Möbius and Leif Kobbelt. “OpenFlipper: An Open Source Geometry Processing and Rendering Framework”. In: *Proceedings of Curves and Surfaces 2010*. Ed. by Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2010, pp. 488–500. DOI: 10.1007/978-3-642-27413-8_31.

[Mokhtari07] Parham Mokhtari, Hironori Takemoto, Ryouichi Nishimura, and Hiroaki Kato. “Comparison of Simulated and Measured HRTFs: FDTD Simulation Using MRI Head Data”. In: *Proceedings of the 123rd Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 5, 2007.

[Mokhtari08] Parham Mokhtari, Ryouichi Nishimura, and Hironori Takemoto. “Toward HRTF Personalization: An Auditory-Perceptual Evaluation of Simulated and Measured HRTFs”. In: *Proceedings of the 14th International Conference on Auditory Display*. Paris, France, 2008.

[Mokhtari19] Parham Mokhtari, Hiroaki Kato, Hironori Takemoto, Ryouichi Nishimura, Seigo Enomoto, Seiji Adachi, and Tatsuya Kitamura. "Further Observations on a Principal Components Analysis of Head-Related Transfer Functions". In: *Scientific Reports* 9.7477 (May 16, 2019). DOI: 10.1038/s41598-019-43967-0.

[Møller92] Henrik Møller. "Fundamentals of Binaural Technology". In: *Applied Acoustics* 36.3 (Jan. 1, 1992), pp. 171–218. DOI: 10.1016/0003-682X(92)90046-U.

[Møller96] Henrik Møller, Michael Friis Sørensen, Clemen Boje Jensen, and Dorte Hammershøi. "Binaural Technique: Do We Need Individual Recordings?" In: *Journal of the Audio Engineering Society* 44.6 (June 1, 1996), pp. 451–469.

[Morimoto01] Masayuki Morimoto. "The Contribution of Two Ears to the Perception of Vertical Angle in Sagittal Planes". In: *The Journal of the Acoustical Society of America* 109.4 (Mar. 30, 2001), pp. 1596–1603. DOI: 10.1121/1.1352084.

[Morimoto84] Masayuki Morimoto and Hitoshi Aokata. "Localization Cues of Sound Sources in the Upper Hemisphere". In: *Journal of the Acoustical Society of Japan* 5.3 (1984), pp. 165–173. DOI: 10.1250/ast.5.165.

[Nelder65] John A. Nelder and Roger Mead. "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4 (Jan. 1, 1965), pp. 308–313. DOI: 10.1093/comjnl/7.4.308.

[Nishino07] Takanori Nishino, Naoya Inoue, Kazuya Takeda, and Fumitada Itakura. "Estimation of HRTFs on the Horizontal Plane Using Physical Features". In: *Applied Acoustics* 68.8 (Aug. 1, 2007), pp. 897–908. DOI: 10.1016/j.apacoust.2006.12.010.

[Oppenheim09] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. 3rd Edition. Prentice Hall, 2009. 1108 pp. ISBN: 978-0-13-198842-2.

[Otani03] Makoto Otani and Shiro Ise. “A Fast Calculation Method of the Head-Related Transfer Functions for Multiple Source Points Based on the Boundary Element Method”. In: *Acoustical Science and Technology* 24.5 (2003), pp. 259–266. DOI: 10.1250/ast.24.259.

[Patterson92] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. “Complex Sounds and Auditory Images”. In: *Proceedings of the 9th International Symposium on Hearing*. Ed. by Y. Cazals, K. Horner, and L. Demany. Carcens, France, Jan. 1, 1992, pp. 429–446. DOI: 10.1016/B978-0-08-041847-6.50054-X.

[Plogsties00] Jan Plogsties, Pauli Minnaar, S. Krarup Olesen, Flemming Christensen, and Henrik Møller. “Audibility of All-Pass Components in Head-Related Transfer Functions”. In: *Proceedings of the 108th Audio Engineering Society Convention*. Paris, France: Audio Engineering Society, Feb. 19, 2000.

[Pollow14] Martin Pollow and Michael Vorländer. “Efficient Quality Assessment of Spatial Audio Data of High Resolution”. In: *Proceedings of the 40th German Annual Conference on Acoustics (DAGA)*. Oldenburg, Germany, 2014.

[Prepeliță16] Sebastian Prepeliță, Michele Geronazzo, Federico Avanzini, and Lauri Savioja. “Influence of Voxelization on Finite Difference Time Domain Simulations of Head-Related Transfer Functions”. In: *The Journal of the Acoustical Society of America* 139.5 (May 1, 2016), pp. 2489–2504. DOI: 10.1121/1.4947546.

[Qi18] Xiaoke Qi and Jianhua Tao. “Sparsity-Constrained Weight Mapping for Head-Related Transfer Functions Individualization from Anthropometric Features”. In: *Proceedings of Interspeech 2018*. Hyderabad, India, Sept. 2, 2018, pp. 841–845. DOI: 10.21437/Interspeech.2018-1615.

[Rajamani07] Kumar T. Rajamani, Martin A. Styner, Haydar Talib, Guoyan Zheng, Lutz P. Nolte, and Miguel A. González Ballester. “Statistical Deformable Bone Models for Robust 3D Surface Extrapolation from Sparse Data”. In: *Medical Image Analysis* 11.2 (Apr. 1, 2007), pp. 99–109. DOI: 10.1016/j.media.2006.05.001.

[Ranjan16] Rishabh Ranjan, JianJun He, and Woon-Seng Gan. “Fast Continuous Acquisition of HRTF for Human Subjects with Unconstrained Random Head Movements in Azimuth and Elevation”. In: *Proceedings of the 2016 AES International Conference on Headphone Technology*. Aalborg, Denmark: Audio Engineering Society, Aug. 19, 2016.

[Rayleigh07] Lord Rayleigh. “On Our Perception of Sound Direction”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. Series 6 13.74 (1907), pp. 214–232.

[Reichinger13] Andreas Reichinger, Piotr Majdak, Robert Sablatnig, and Stefan Maierhofer. “Evaluation of Methods for Optical 3-D Scanning of Human Pinnas”. In: *Proceedings of the 2013 International Conference on 3D Vision (3DV)*. Seattle, WA, USA: IEEE, June 2013, pp. 390–397. DOI: 10.1109/3DV.2013.58.

[Reiss05] Lina A. J. Reiss and Eric D. Young. “Spectral Edge Sensitivity in Neural Circuits of the Dorsal Cochlear Nucleus”. In: *Journal of Neuroscience* 25.14 (Apr. 6, 2005), pp. 3680–3691. DOI: 10.1523/JNEUROSCI.4963-04.2005. pmid: 15814799.

[Richter19] Jan-Gerrit Richter. “Fast Measurement of Individual Head-Related Transfer Functions”. PhD Thesis. Aachen, Germany: Aachener Beiträge zur Akustik, 2019. 172 pp.

[Riederer98] Klaus A. J. Riederer. “Repeatability Analysis of Head-Related Transfer Function Measurements”. In: *Proceedings of the 105th Audio Engineering Society Convention*. San Francisco, CA, USA: Audio Engineering Society, Sept. 26, 1998.

[Röber06] Niklas Röber, Sven Andres, and Maic Masuch. *HRTF Simulations through Acoustic Raytracing*. Technical Report 4. Fakultät für Informatik, Otto-von-Guericke Universität: Magdeburg Germany, 2006, 2006.

[Romigh14] Griffin D. Romigh and Brian D. Simpson. “Do You Hear Where I Hear?: Isolating the Individualized Sound Localization Cues”. In: *Frontiers in Neuroscience* 8 (2014). DOI: 10.3389/fnins.2014.00370.

[Rothbucher13] Martin Rothbucher, Kajetan Veprek, Philipp Paukner, Tim Habigt, and Klaus Diepold. “Comparison of Head-Related Impulse Response Measurement Approaches”. In: *The Journal of the Acoustical Society of America* 134.2 (July 15, 2013), EL223–EL229. DOI: 10.1121/1.4813592.

[Royston83] J. Patrick Royston. “Some Techniques for Assessing Multivarate Normality Based on the Shapiro-Wilk W”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 32.2 (1983), pp. 121–133. DOI: 10.2307/2347291. JSTOR: 2347291.

[Rueff20] Pascal Rueff. *Barrow-Madec-Turnbull Trio*. 3D Radio. 2020. URL: https://www.binaural.fr/binaural?p=1533.

[Rugeles Ospina14] Felipe Rugeles Ospina, Marc Emerit et Brian F. G. Katz. « Évaluation Objective et Subjective de Différentes méthodes de Lissage des HRTF [Objective and Subjective Evaluation of Various HRTF Smoothing Methods] ». In : *Actes du Congrès Français d’Acoustique (CFA)*. Poitiers, France, 25 avr. 2014.

[Rugeles Ospina15] Felipe Rugeles Ospina, Marc Emerit, and Jérôme Daniel. “A Fast Measurement of High Spatial Resolution Head Related Transfer Functions for the BiLi Project”. In: *Proceedings of the 3rd International Conference on Spatial Audio (ICSA)*. Graz, Austria, Sept. 2015.

[Rugeles Ospina16] Felipe Rugeles Ospina. « Individualisation de l’écoute binaurale : création et transformation des indices spectraux et des morphologies des individus [Individualization of Binaural Listening : Creation and Transformation of the Spectral Cues and Morphologies of Individuals] ». PhD Thesis. Universite Pierre et Marie Curie / Orange Labs, juil. 2016. 207 p.

[Rui13] Yuanqing Rui, Guangzheng Yu, Bosun Xie, and Yu Liu. “Calculation of Individualized Near-Field Head-Related Transfer Function Database Using Boundary Element Method”. In: *Proceedings of the 134th Audio Engineering Society Convention*. Rome, Italy: Audio Engineering Society, May 4, 2013.

[Runkle00] Paul Runkle, Anastasia Yendiki, and Gregory H. Wakefield. “Active Sensory Tuning for Immersive Spatialized Audio”. In: *Proceedings of the 2000 International Conference on Auditory Display (ICAD)*. Atlanta, GA, USA, Apr. 2000.

[Sandvad94] Jesper Sandvad and Dorte Hammershøi. “Binaural Auralization, Comparison of FIR and IIR Filter Representation of HIRs”. In: *Proceedings of the 96th Audio Engineering Society Convention*. Amsterdam, Netherlands: Audio Engineering Society, Feb. 26, 1994.

[Savioja99] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Ritta Väänänen. “Creating Interactive Virtual Acoustic Environments”. In: *Journal of the Audio Engineering Society* 47.9 (1999), pp. 675–705.

[Schönstein10] David Schönstein and Brian F. G. Katz. “HRTF Selection for Binaural Synthesis from a Database Using Morphological Parameters”. In: *International Congress on Acoustics (ICA)*. 2010.

[Schönstein12a] David Schönstein. “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”. PhD Thesis. Université Pierre et Marie Curie - Paris VI, Sept. 2012.

[Schönstein12b] David Schönstein and Brian F. G. Katz. “Variability in Perceptual Evaluation of HRTFs”. In: *Journal of the Audio Engineering Society* 60.10 (Nov. 26, 2012), pp. 783–793.

[Schroeder70] M. R. Schroeder. “Digital Simulation of Sound Transmission in Reverberant Spaces”. In: *The Journal of the Acoustical Society of America* 47 (2A Feb. 1, 1970), pp. 424–431. DOI: 10.1121/1.1911541.

[Seeber03] Bernhard U. Seeber and Hugo Fastl. “Subjective Selection of Non-Individual Head-Related Transfer Functions”. In: *Proceedings of the 2003 International Conference on Auditory Display (ICAD)*. Boston, MA, USA, July 6, 2003.

[Shaw68] E. A. G. Shaw and R. Teranishi. “Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source”. In: *The Journal of the Acoustical Society of America* 44.1 (July 1, 1968), pp. 240–249. DOI: 10.1121/1.1911059.

[Shin08] Ki Hoon Shin and Youngjin Park. “Enhanced Vertical Perception through Head-Related Impulse Response Customization Based on Pinna Response Tuning in the Median Plane”. In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E91-A.1 (Jan. 1, 2008), pp. 345–356. DOI: 10.1093/ietfec/e91-a.1.345.

[Simon16] Laurent S. R. Simon, Areti Andreopoulou, and Brian F. G. Katz. “Investigation of Perceptual Interaural Time Difference Evaluation Protocols in a Binaural Context”. In: *Acta Acustica united with Acustica* 102.1 (2016), pp. 129–140.

[Smith07] Julius O. Smith. *Introduction to Digital Filters with Audio Applications*. http://ccrma.stanford.edu/ jos/filters, online book, 2007.

[Søndergaard13] P. L. Søndergaard and P. Majdak. “The Auditory Modeling Toolbox”. In: *The Technology of Binaural Listening*. Ed. by Jens Blauert. Springer, 2013, pp. 33–56. ISBN: 978-3-642-37761-7.

[Spagnol11] Simone Spagnol, Marko Hiipakka, and Ville Pulkki. “A Single-Azimuth Pinna-Related Transfer Function Database”. In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*. Paris, France, Sept. 19, 2011, pp. 209–212.

[Spagnol20] Simone Spagnol. “Auditory Model Based Subsetting of Head-Related Transfer Function Datasets”. In: *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2020, pp. 391–395. DOI: 10.1109/ICASSP40776.2020.9053360.

[Stitt19] Peter Stitt, Lorenzo Picinali, and Brian F. G. Katz. “Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning”. In: *Scientific Reports* 9.1 (1 Jan. 31, 2019), p. 1063. DOI: 10.1038/s41598-018-37873-0.

[Takane15] Shouichi Takane. “Effect of Domain Selection for Compact Representation of Spatial Variation of Head-Related Transfer Function in All Directions Based on Spatial Principal Components Analysis”. In: *Applied Acoustics* 101 (Aug. 24, 2015), pp. 64–77. DOI: 10.1016/j.apacoust.2015.07.018.

[Takemoto12] Hironori Takemoto, Parham Mokhtari, Hiroaki Kato, Ryouichi Nishimura, and Kazuhiro Iida. “Mechanism for Generating Peaks and Notches of Head-Related Transfer Functions in the Median Plane”. In: *The Journal of the Acoustical Society of America* 132.6 (Dec. 1, 2012), pp. 3832–3841. DOI: 10.1121/1.4765083.

[Tan98] Chong-Jin Tan and Woon-Seng Gan. “User-Defined Spectral Manipulation of HRTF for Improved Localisation in 3D Sound Systems”. In: *Electronics Letters* 34.25 (Dec. 10, 1998), pp. 2387–2389. DOI: 10.1049/el:19981629.

[Tao03] Yufei Tao, Anthony I. Tew, and Stuart J. Porter. “The Differential Pressure Synthesis Method for Efficient Acoustic Pressure Estimation”. In: *Journal of the Audio Engineering Society* 51.7/8 (July 15, 2003), pp. 647–656.

[Tipping99] Michael E. Tipping and Christopher M. Bishop. “Probabilistic Principal Component Analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622. DOI: 10.1111/1467-9868.00196.

[Tsui18] Benjamin Tsui and Gavin Kearney. “A Head-Related Transfer Function Database Consolidation Tool For High Variance Machine Learning Algorithms”. In: *Proceedings of the 145th Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, 2018.

[Turku08] Julia Turku, Miikka Vilermo, Eira Seppälä, Monika Pölönen, Ole Kirkeby, Asta Kärkkäinen, and Leo Kärkkäinen. “Perceptual Evaluation of Numerically Simulated Head-Related Transfer Functions”. In: *Proceedings of the 124th Audio Engineering Society Convention*. Amsterdam, Netherlands: Audio Engineering Society, May 1, 2008.

[Wade08] Nicholas J. Wade and Diana Deutsch. “Binaural Hearing – Before and After the Stethophone”. In: *Acoustics Today* 4.3 (July 2008), pp. 16–27.

[Wallach40] Hans Wallach. “The Role of Head Movements and Vestibular and Visual Cues in Sound Localization.” In: *Journal of Experimental Psychology* 27.4 (1940), p. 339. DOI: 10.1037/h0054629.

[Warusfel03] Olivier Warusfel. *Listen HRTF Database*. IRCAM and AK. 2003. URL: http://recherche.ircam.fr/equipes/salles/listen/index.html.

[Watanabe14] Kanji Watanabe, Yukio Iwaya, Yôiti Suzuki, Shouichi Takane, and Sojun Sato. “Dataset of Head-Related Transfer Functions Measured with a Circular Loudspeaker Array”. In: *Acoustical Science and Technology* 35.3 (Mar. 1, 2014), pp. 159–165. DOI: 10.1250/ast.35.159.

[Wenzel93] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. “Localization Using Nonindividualized Head-related Transfer Functions”. In: *The Journal of the Acoustical Society of America* 94.1 (July 1, 1993), pp. 111–123. DOI: 10.1121/1.407089.

[Wightman89a] Frederic L. Wightman and Doris J. Kistler. “Headphone Simulation of Free-field Listening. I: Stimulus Synthesis”. In: *The Journal of the Acoustical Society of America* 85.2 (Feb. 1, 1989), pp. 858–867. DOI: 10.1121/1.397557.

[Wightman89b] Frederic L. Wightman and Doris J. Kistler. “Headphone Simulation of Free-field Listening. II: Psychophysical Validation”. In: *The Journal of the Acoustical Society of America* 85.2 (Feb. 1, 1989), pp. 868–878. DOI: 10.1121/1.397558.

[Wightman92] Frederic L. Wightman and Doris J. Kistler. “The Dominant Role of Low-frequency Interaural Time Differences in Sound Localization”. In: *The Journal of the Acoustical Society of America* 91.3 (Mar. 1, 1992), pp. 1648–1661. DOI: 10.1121/1.402445.

[Wightman99] Frederic L. Wightman and Doris J. Kistler. “Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement”. In: *The Journal of the Acoustical Society of America* 105.5 (Apr. 27, 1999), pp. 2841–2853. DOI: 10.1121/1.426899.

[Woodworth54] Robert Woodworth and Harold Schlosberg. *Experimental Psychology*. Revised Edition. Holt, Rinehart and Winston, 1954. ISBN: 030074401.

[Xiao03] Tian Xiao and Qing Huo Liu. “Finite Difference Computation of Head-Related Transfer Function for Human Hearing”. In: *The Journal of the Acoustical Society of America* 113.5 (May 1, 2003), pp. 2434–2441. DOI: 10.1121/1.1561495.

[Xie10] Bosun Xie and Tingting Zhang. “The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency”. In: *Acta Acustica united with Acustica* 96.2 (Mar. 1, 2010), pp. 328–339. DOI: 10.3813/AAA.918282.

[Xie12] Bo-Sun Xie. “Recovery of Individual Head-Related Transfer Functions from a Small Set of Measurements”. In: *The Journal of the Acoustical Society of America* 132.1 (July 1, 2012), pp. 282–294. DOI: 10.1121/1.4728168.

[Xie15] Bosun Xie, Xiaoli Zhong, and Nana He. “Typical Data and Cluster Analysis on Head-Related Transfer Functions from Chinese Subjects”. In: *Applied Acoustics* 94 (July 1, 2015), pp. 1–13. DOI: 10.1016/j.apacoust.2015.01.022.

[Xu08] Song Xu, Zhizhong Li, and Gavriel Salvendy. “Improved Method to Individualize Head-Related Transfer Function Using Anthropometric Measurements”. In: *Acoustical Science and Technology* 29.6 (2008), pp. 388–390.

[Yamamoto17] Kazuhiko Yamamoto and Takeo Igarashi. “Fully Perceptual-Based 3D Spatial Sound Individualization with an Adaptive Variational Autoencoder”. In: *Association for Computing Machinery (ACM) Transactions on Graphics* 36.6 (Nov. 20, 2017), pp. 1–13. DOI: 10.1145/3130800.3130838.

[Yao17] Shu-Nung Yao, Tim Collins, and Chaoyun Liang. “Head-Related Transfer Function Selection Using Neural Networks”. In: *Archives of Acoustics* 42.3 (2017), pp. 365–373. DOI: 10.1515/aoa-2017-0038.

[Younes20] Lara Younes, Corentin Guezenoc, and Renaud Séguier. “Method for Producing a 3D Scatter Plot Representing a 3D Ear of an Individual, and Associated System”. U.S. pat. 10,818,100. 3D Sound Labs, Mimi Hearing Technologies GmbH. Feb. 13, 2020.

[Zagala20] Franck Zagala, Markus Noisternig, and Brian F. G. Katz. “Comparison of Direct and Indirect Perceptual Head-Related Transfer Function Selection Methods”. In: *The Journal of the Acoustical Society of America* 147.5 (May 1, 2020), pp. 3376–3389. DOI: 10.1121/10.0001183.

[Zhang20] Mengfan Zhang, Zhongshu Ge, Tiejun Liu, Xihong Wu, and Tianshu Qu. “Modeling of Individual HRTFs Based on Spatial Principal Component Analysis”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 785–797. DOI: 10.1109/TASLP.2020.2967539.

[Ziegelwanger13] Harald Ziegelwanger, Andreas Reichinger, and Piotr Majdak. “Calculation of Listener-Specific Head-Related Transfer Functions: Effect of Mesh Quality”. In: *Proceedings of the 21st International Congress on Acoustics (ICA)*. Vol. 19. Montréal, QE, Canada: Acoustical Society of America, June 2, 2013, p. 050017. DOI: 10.1121/1.4799868.

[Ziegelwanger14a] Harald Ziegelwanger and Piotr Majdak. “Modeling the Direction-Continuous Time-of-Arrival in Head-Related Transfer Functions”. In: *The Journal of the Acoustical Society of America* 135.3 (Mar. 1, 2014), pp. 1278–1293. DOI: 10.1121/1.4863196.

[Ziegelwanger14b] Harald Ziegelwanger, Piotr Majdak, and Wolfgang Kreuzer. “Efficient Numerical Calculation of Head-Related Transfer Functions”. In: *Proceedings of the 7th Forum Acusticum*. Kraków, Poland: European Acoustics Association, Sept. 7, 2014.

[Ziegelwanger14c] Harald Ziegelwanger, Piotr Majdak, and Wolfgang Kreuzer. "Non-Uniform Sampling of Geometry for the Numeric Simulation of Head-Related Transfer Functions". In: *Proceedings of the 21st International Congress on Sound and Vibration (ICSV)*. Beijing, China, July 13, 2014.

[Ziegelwanger15a] Harald Ziegelwanger, Wolfgang Kreuzer, and Piotr Majdak. "Mesh2HRTF: Open-Source Software Package for the Numerical Calculation of Head-Related Transfer Functions". In: *Proceedings of the 22nd International Congress on Sound and Vibration (ICSV)*. Florence, Italy, July 16, 2015.

[Ziegelwanger15b] Harald Ziegelwanger, Piotr Majdak, and Wolfgang Kreuzer. "Numerical Calculation of Listener-Specific Head-Related Transfer Functions and Sound Localization: Microphone Model and Mesh Discretization". In: *The Journal of the Acoustical Society of America* 138.1 (July 1, 2015), pp. 208–222. DOI: 10.1121/1.4922518.

[Ziegelwanger16] Harald Ziegelwanger, Wolfgang Kreuzer, and Piotr Majdak. "A Priori Mesh Grading for the Numerical Calculation of the Head-Related Transfer Functions". In: *Applied Acoustics* 114 (Dec. 15, 2016), pp. 99–110. DOI: 10.1016/j.apacoust.2016.07.005.

[Zotkin02] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. "Customizable Auditory Displays". In: *Proceedings of the 2002 International Conference on Auditory Display (ICAD)*. Kyoto, Japan, July 2, 2002.

[Zotkin06] Dmitry N. Zotkin, Ramani Duraiswami, Elena Grassi, and Nail A. Gumerov. "Fast Head-Related Transfer Function Measurement via Reciprocity". In: *The Journal of the Acoustical Society of America* 120.4 (Oct. 1, 2006), pp. 2202–2215. DOI: 10.1121/1.2207578.

Appendix A

Abbreviations

APE	Absolute polar error
BEM	Boundary element method
CAPZ	Common acoustical poles and zeros
CTF	Common transfer function
DFEQ	Diffuse-field equalization
DTF	Directional transfer function
ERB	Equivalent rectangular bandwidth
FDTD	Finite difference time domain
FEM	Finite element method
FM-BEM	Fast-multipole boundary element method
HRTF	Head-related transfer function
HRIR	Head-related impulse response
ICA	Independent component analysis
ILD	Interaural level difference
ITD	Interaural time difference
JND	Just-noticeable difference
PCA	Principal component analysis
PRTF	Pinna-related transfer function
PRIR	Pinna-related impulse response
QE	Quadrant error
PE	Polar error
SFRS	Spatial frequency response surface
SH	Spherical harmonic
SHD	Spherical harmonics decomposition
SWD	Spherical wavelets decomposition
TOA	Time of arrival
VAS	Virtual acoustic scene

WiDESPREaD	Wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings

Appendix B

Publications

Peer-Reviewed Journals

- Corentin Guezenoc and Renaud Séguier. "A Wide Dataset of Ear Shapes and Pinna-Related Transfer Functions Generated by Random Ear Drawings". In: *The Journal of the Acoustical Society of America* 147.6 (June 23, 2020), pp. 4087–4096. DOI: 10.1121/10.0001461

Peer-Reviewed International Conferences

- Corentin Guezenoc and Renaud Séguier. "Dataset Augmentation and Dimensionality Reduction of Pinna-Related Transfer Functions". In: *Proceedings of the 148th Audio Engineering Society Convention*. Vienna, Austria: Audio Engineering Society, May 28, 2020. ISBN: 978-1-942220-32-9. DOI: 10.17743/aesconv.2020.978-1-942220-32-9.

- Corentin Guezenoc and Renaud Séguier. "HRTF Individualization: A Survey". In: *Proceedings of the 145th Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, Oct. 7, 2018. DOI: 10.17743/aesconv.2018.978-1-942220-25-1.

Patents

- Lara Younes, Corentin Guezenoc, and Renaud Séguier. "Method for Producing a 3D Scatter Plot Representing a 3D Ear of an Individual, and Associated System". U.S. pat. 10,818,100. 3D Sound Labs, Mimi Hearing Technologies GmbH. Feb. 13, 2020.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

Titre : Individualisation de la synthèse binaurale par retours perceptifs d'auditeur

Mot clés : audio spatiale, synthèse binaurale, individualisation, HRTF

Résumé : En synthèse binaurale, fournir à l'auditeur des HRTFs (fonctions de transfert relatives à la tête) personnalisées est un problème clef, traité dans cette thèse. D'une part, nous proposons une méthode d'individualisation qui consiste à régler automatiquement les poids d'un modèle statistique ACP (analyse en composantes principales) de jeu d'HRTF à partir des performances de localisation de l'auditeur. Nous examinons la faisabilité de l'approche proposée sous différentes configurations grâce à des simulations psycho-acoustiques des tests d'écoute, puis la testons sur 12 auditeurs. Nous constatons qu'elle permet une amélioration considérable des performances de localisation comparé à des conditions d'écoute non-individuelles, atteignant des performances comparables à celles rapportées dans la littérature pour des HRTF individuelles. D'autre part, nous examinons une question sous-jacente : la réduction de dimensionnalité des jeux d'HRTF. Après avoir comparé la réduction de dimensionalité par ACP de 9 bases de données contemporaines d'HRTF et de PRTF (fonctions de transfert relatives au pavillon de l'oreille), nous proposons une méthode d'augmentation de données basée sur la génération aléatoire de formes d'oreilles 3D et sur la simulation des PRTF correspondantes par méthode des éléments frontières.

Title: Binaural Synthesis Individualization based on Listener Perceptual Feedback

Keywords: spatial audio, binaural synthesis, individualization, HRTF

Abstract: In binaural synthesis, providing individual HRTFs (head-related transfer functions) to the end user is a key matter, which is addressed in this thesis. On the one hand, we propose a method that consists in the automatic tuning of the weights of a principal component analysis (PCA) statistical model of the HRTF set based on listener localization performance. After having examined the feasibility of the proposed approach under various settings by means of psycho-acoustic simulations of the listening tests, we test it on 12 listeners. We find that it allows considerable improvement in localization performance over non-individual conditions, up to a performance comparable to that reported in the literature for individual HRTF sets. On the other hand, we investigate an underlying question: the dimensionality reduction of HRTF sets. After having compared the PCA-based dimensionality reduction of 9 contemporary HRTF and PRTF (pinna-related transfer function) databases, we propose a dataset augmentation method that relies on randomly generating 3-D pinna meshes and calculating the corresponding PRTFs by means of the boundary element method.