POLITECNICO DI MILANO
DEPARTMENT OF ELECTRONICS, INFORMATION AND BIOENGINEERING
DOCTORAL PROGRAMME IN INFORMATION TECHNOLOGY

A MORPHOACOUSTIC APPROACH TOWARDS HEAD RELATED TRANSFER FUNCTION PERSONALIZATION

Doctoral Dissertation of:
Muhammad Shahnawaz

Supervisor:
Prof. Augusto Sarti

Tutor:
Prof. Andrea Monti Guarnieri

The Chair of the Doctoral Program:
Prof. Barbara Pernici

Academic Year : **2018-2019**
PhD Cycle : **XXX**

To my parents, my siblings, and my wife.

Acknowledgement

First of all, I would like to thank ALLAH for giving me strength, chance, and endurance to go through this long journey. After that, I will thank everybody at the Erasmus-Mundus INTACT project team, especially the organizers and coordinators, to make it possible for me to pursue my Ph.D. in Europe.

After, that the first person I will thank is my Ph.D. supervisor Prof. Augusto Sarti, for all the guidance, support, the motivation he provided me throughout my Ph.D. and giving me enough self-confidence to believe that I can do it and prepared me to be an independent researcher. I am also very grateful to Prof. Marcon for the multiple discussions we had regarding shape space modeling and introducing me to industry contacts to arrange an internship and financial support for me. It is him who is the reason behind me finding an open-ended contract and working for a big company like STMicroelectronics today.

I would also express my sincerest and deepest gratitude to Prof. Craig Jin from the University of Sydney, Australia, for not just offering me an opportunity to spend some time at his lab but also funding it. During this time, I was able to use the tools, databases, and lab facilities, which were crucial and played a big part in creating and formulating this Ph.D. study. Without his help and guidance, it would be tough for me to graduate. I would also like to thank him for having in-depth discussions and working on my research skills and personal development during my stay at CARLab, which did not just help me during my Ph.D. but will help me for the rest of my life. Through him, I also met Prof. Tony Tew, of The University of York, UK, and Prof. Joan Glaunes, from Paris, who also contributed towards some studies in this work and provided useful insights.

I would also like to thank the reviewers for this thesis, professor Maximo Cobos and professor Federico Avanzini, for providing valuable feedback to improve the thesis.

After thanking all the professors, I will like to express my deepest gratitude to everyone at the DEIB admin team who made this my second home. Special thanks to Mme. Cortiana, Mme. Brambila, Mme. Rosa, Mme. Parada and Mme. Clemenza for all their bureaucratic support, Gianfranco and Resmini from IT support team for providing all the IT related support whenever needed.

After thanking everyone in universities, I will like to thank people who had more personal and selfless contributions. I would like to thank every single one who helped me through this long journey. There were countless people, and it is hard to mention every single one, but I would like to thank especially, Bruno, Michele, Antonio, Max, Paolo, Fabio, Federico, Alberto, from Italy, and Reza, Abdullah, Sam, and Duy from Australia to help me in technical aspects and discussions. I will also thank Farooq, Saleem, Naveed from Italy, Aaminah, Basit, Hammad, Tanvir, Jeff, and Graham from Australia, Naumana, Ayesha, and Sanober from the UK to motivate me. I will also thank, Naeem, Israr, Akram, Sara, Silvia, from Italy, and Nadim, Rashid, Junaid, Aasim, Waseem, Mubashir, Sagheer, Abdullah Alazzawi, from Australia for making my life fun during this time. I will also express sincere thanks to my eldest friend in this world, Dr. Ahsan Zafarullah, whom I had very deep discussions with on philosophy of life, religion, love, and whatnot.

Finally, I would like to extend my special thanks to my family: starting from my parents, Malik Imtiaz Ahmed and Mrs. Tasneem Akhtar, for all the motivation, support, and selfless love. I am also very thankful to my siblings, Muhammad Umar Sajjad and Muhammad Talha Imtiaz, and my sister in law Mrs. Sonia Bashir for taking care of all the family matters back in the home and letting me focus on my studies, to my nephews, Muhammad Asadullah and Muhammad Alam for being super cute and asking super-smart questions, and my niece to be the light of our house.

Finally, I would like to mention a new addition to my life, my precious, lovely, caring, and kind wife, Huma Hatun, who has not just contributed towards the completion of this thesis but has also completed my life.

Abstract

The head-related transfer function (HRTF) for a location describes the transfer characteristics for the sound waves as they travel from a sound source at that location to the ear canal in free space conditions. These transfer functions depend significantly on the individual's head, torso, and ear morphology and are highly idiosyncratic. The knowledge of these individualized acoustic transfer functions is crucial to present personalized 3D audio through binaural rendering. This thesis builds on the currently available knowledge on the HRTF personalization and aims to widen this knowledge space by presenting some studies. These studies can aid in modeling and understanding the relationship between the morphology of an individual and corresponding HRTFs and facilitate one to create a simple HRTF personalization method to estimate individualized HRTFs without performing acoustic measurements or long numerical simulations. This thesis work is a composite of many studies and concepts from different fields. These studies include primary signal processing techniques such as spectral analysis, notch extraction, principal component analysis (PCA), and sparse representation based modeling, the physics of numerical simulations like Fast-Multipole Boundary Element Methods (FM-BEM), and functional space analysis of shapes like large deformation diffeomorphic metric mapping (LDDMM), and kernel principal component analysis (KPCA) on LDDMM data. The studies performed in this thesis can be divided into two groups. The first set of works provides some preliminary studies which can be used to personalize the HRTFs based on anthropometric data. These studies are mainly performed on the CIPIC database, and focus on the personalization methods based on the anthropometric data. On the other hand, the second set of works presents the studies based on the morphoacoustic approach and considers 3D morphology data for subjects. This work aims to widen the understanding of the relationship between the outer ear shapes and the corresponding acoustics by studying the variations in both spaces separately and then finding a mapping between the variations in two spaces. All the studies presented in this group are performed on the SYMARE database. There are two studies in the first group. The first study provides a statistical analysis of the center frequencies of first notches in the HRTFs of CIPIC and SYMARE databases. The notches for the HRTFs in the median plane are extracted for both databases and clustered into three clusters using k-means. Each cluster represents

the notch created due to one of the three main contours of the ear shape, as suggested in past studies. The centroids of the clusters show the evolution of notches in frequency as a function of elevation angles. The results are compared for two databases showing almost the same results. The results show that the mean value for three notch frequencies in both databases evolves monotonically for the first two notches as a function of elevation angle from -45° to 45°. In contrast, for the third notch, this frequency almost stays flat. The mean notch frequency for first, second, and third notches range from 6 kHz to 8.5 kHz, 10 kHz to 12 kHz, and 13.5 kHz to 14 kHz respectively. This study also compares these frequencies for left and right ears in both databases. The results show that these frequencies are not symmetric in both ears. This asymmetry suggests either the possible effectiveness of the binaural cues in the median plane or could be simply due to the asymmetry in the ear shapes of the involved subjects or measurement setup. The second study provides a preliminary HRTF personalization method based on weighted sparse representation based modeling. Like past sparse representation-based methods this method also relies on two strong assumptions, 1) the anthropometric features of the available subject set are rich enough to model the anthropometric features of any new subject, and 2) a same sparse modeling (linear combination) can be used to model both the anthropometric features as well as the corresponding HRTFs. However, the study presented in this work is different from the past sparse representation based HRTF personalization studies for two reasons. The first difference is that it uses a separate sparse representation for both left and right ears, while the past studies used the same model for both left and right ears. The reason to do so is the findings of our previous studies on notch analysis, which showed asymmetry in HRTF of both ears. The second difference and contribution of this work is the use of weighted sparse representation. The previous studies considered all the anthropometric parameters to be equally relevant while calculating the sparse representation. However, our work calculates the relevance of each of the available anthropometric parameters and use these relevance metrics as the weights to the sparse representation. Hence the name weighted sparse representation. Furthermore, this compares the results of the method with some famous closest-matching based personalization schemes and shows that it outperforms the previous techniques. In the second group of studies, the first work analyzes the effects of affine transformations of the ear shapes on the corresponding HRTFs. As a counter product, this study creates a synthetic database from SYMARE (one of its kind), which we call affine models for the SYMARE population. For the affine models, the ear shapes are affine matched with the template ear shape to have the same scale, orientation, and position. The affine matched ears are then attached to the template head and torso shapes to create a 3D model of the head, torso, and (affine matched ears), called an affine model. The benefits of creating an affine model can be multifold. The first and most important benefit of this is that it creates a simplistic paradigm to study the morphoacoustics of the ear shape, by limiting the variations to only ear shape variations, and removing all the variations due to different head and torso shapes, ear sizes, ear rotations, and position of the ears on the head. The second benefit is that it simplifies the process of modeling the ear shape as one has to model the shape variations only using LDDMM and KPCA, not the scale and rotation. Third, it supposedly simplifies the modeling process of the acoustics, as all the ear shapes are at the same scale, position, and rotation and are placed on the same head and torso shape.

However, this may end up creating artifacts that outweigh all these benefits. This work investigates all these questions. In this work, we present a study that provides an analysis of how simple corrections such as frequency scaling of the HRTFs (to correct for the scales) and rotation of HRTF directivity patterns (to correct for the rotations) can significantly compensate for these affine transformations. This also studies and calculate the amount of inter-subject variations coming from affine matching vs. the original shape. Finally, the study calculates the optimal frequency scaling factor from a purely acoustic point of view, which matches the affine modeled HRTFs to the original HRTFs in the best way. These optimal scaling factors are then related to the physical scaling factors by using linear regression. The results show these scaling factors can be inferred simply by knowing the ear shape scaling factors coming from the affine matching process. The second study in this group provides a simple Spatial Principal Component Analysis (SPCA) based modeling method to analyze the variations in the acoustic directivity patterns of the HRTFs as a function of frequency. The directivity patterns of different frequencies are modeled separately, and the number of principal components required to model the directivity patterns for a given frequency is quantified for all the frequency bins in the frequency range from 0.2-17 kHz. This study reasserts the importance of the affine models by showing that the directivity patterns of the affine models can be described by using only eight principal components at even high frequencies up to 17 kHz, keeping the average standard spectral difference (SDD) of less than 3 dBs. Using the existing morphable model of the ear shapes this work model the ear shapes with just first eight principal components and showing results for some ears. Finally, using the eight principal components of the shape space, it estimates the acoustic principal components through linear regression to provide a simple personalization method for HRTFs.

The last study in this work provides a novel idea of morphological weighting to create a weighted morphable model for ear shapes. This study proposes to assign different weights to different ear portions and use a weighted kernel for KPCA on LDDMM data to create a weighted morphable model. The results of this preliminary work show a better prediction for the acoustic principal components is achieved when weighted KPCA is used compared to traditional KPCA on LDDMM data. These insights are very interesting and suggest that with further work, this tool can be used to not just better prediction of personalized HRTFs but also could be an effective way to understand the contributions of different parts of the ear shapes as a variant of morphoacoustic perturbation analysis.

Sommario

La Head-Related Transfer Function (HRTF) esprime la funzione di trasferimento delle onde sonore che viaggiano da una sorgente audio, posta ad una certa posizione nello spazio, fino al canale uditivo, in condizioni di spazio aperto. Queste funzioni di trasferimento, per tutte le posizioni, dipendono in modo significativo dalla morfologia della testa, del busto e dell'orecchio dell'individuo, e sono perciò molto idiosincrasiche, ovvero uniche da individuo a individuo. Lo studio di queste funzioni di trasferimento indivuali è cruciale per poter generare audio 3D attraverso rendering binaurale. Questa tesi parte dalle conoscenze odierne sulla personalizzazione della HRTF e mira ad allargare questa conoscenza. Gli studi presentati in questa tesi possono aiutare a capire e modellare la relazione che intercorre tra la morfologia di un individuo e la sua HRTF, e facilitare la creazione di un metodo semplice per la personalizzazione della HRTF, o stimare HRTF personalizzate, senza il bisogno di condurre misurazioni acustiche o lunghe simulazioni numeriche. Questa tesi si compone di diversi studi e concetti presi da diversi ambiti. Gli studi includono tecniche primarie di elaborazione dei segnali, come analisi spettrale, estrazione di notch, analisi delle componenti principali (principal component analysis, PCA), e modellazione basata su rappresentazioni ridotta (sparse), la fisica dietro a simulazioni numeriche come i metodi di compuazione veloce di elementi finiti con vincoli (Fast-Multipole Boundary Element Methods, FM-BEM), e analisi funzionale di forme come large deformation diffeomorphic metric mapping (LDDMM) e kernel principal component analysis (KPCA) su dati LDDMM. Gli studi condotti in questa tesi possono essere divisi in due gruppi. Il primo gruppo di lavori fornisce degli studi preliminari che possono essere utilizzati per personalizzare the HRTF a partire da dati antropometrici. Questi studi sono stati condotti principalmente sul database CIPIC, e si concentrano sui metodi di personalizzazione basati su dati antropometrici. Il secondo gruppo presenta studi basati sull'approccio morfoacustico e considera la morfologia 3D degli individui. Questo lavoro mira ad ampliare la comprensione della relazione tra le forme del padiglione auricolare e la relativa acustica studiando le variazioni in entambi gli spazi separatamente e poi trovando un collegamento tra le variazioni nei due spazi. Gli studi presentati nel secondo gruppo sono stati condotti sul database SYMARE. Il primo gruppo comprende due studi. Il primo studio fornisce un'analisi statistica delle frequenze centrali

dei primi notch nelle HRTF dei database CIPIC e SYMARE. I notch delle HRTF sul piano mediano sono estratti da entrambi i database e raggruppati in tre cluster usando l'algoritmo k-means. Ogni cluster rappresenta il notch creato da ognuno dei tre contorni principali della forma dell'orecchio, come suggerito dalla letteratura. I centroidi dei cluster mostrano l'evoluzione dei notches in frequenza in funzione dell'angolo di elevazione. I risultati sono confrontati per i due database, mostrando praticamente gli stessi risultati. Questi risultati mostrano che il valore medio per le tre frequenze di notch in entrambi i database evolve monotonicamente per i primi due notch come in funzione dell'angolo di elevazione da -45° a 45°. Il terzo notch, invece, presenta una frequenza praticamente piatta. La frequenza di notch media per i primi tre notch Ã¨ compresa rispettivamente: tra 6 kHz a 8.5 kHz, da 10 kHz a 12 kHz, e da 13.5 kHz a 14 kHz. Questo studio inoltre confronta queste frequenze per l'orecchio destro e sinistro in entrambi i database. I risultati mostrano che queste frequenze non sono simmetriche per entrambe le orecchio. L'asimmetria potrebbe essere causata o da una possibile efficacia di "indizi" binaurali sul piano mediano, o semplicemente dall'asimmetria delle forme delle orecchie dei soggetti coinvolti nella misurazione.

Il secondo studio presenta un metodo preliminare per la personalizzazione della HRTF basata su una modellazione di una rappresentazione ridotta pesata. Come altri metodi basati su rappresentazione ridotta della letteratura, questo metodo assume che: 1) le caratteristiche antropometriche dei soggetti coinvolti siano sufficientemente informative da modellare le caratteristiche di nuovi soggetti e 2) la stessa modellazione ridotta (tramite combinazione lineare) può essere usata per modellare sia le caratteristiche antropometriche che le HRTF corrispondenti. Questo studio però si discosta dai metodi presentati in letterature in due modi. La prima differenza è l'utilizzo di due spazi di rappresentazione ridotta diversi per orecchio destro e sinistro, anzichÃ© un unico spazio. La scelta è motivata dalla suddetta analisi delle frequenze di notch che aveva mostrato un certo grado di asimmetria tra le orecchie. La seconda differenza è l'uso di una rappresentazione ridotta pesata, mentre gli studi in letteratura consideravano i parametri parametrici come equamente rilevanti nel calcolo della rappresentazione ridotta. Invece, il nostro lavoro calcola la rilevanza dei vari parametri e la utilizza come pesi della rappresentazione ridotta, da cui il nome di rappresentazione ridotta pesata. I risultati di questo approccio sono comparabili con alcuni metodi di personalizzazione basati sul closest-matching e sono migliori di molte tecniche della letteratura. Nel secondo gruppo di studi, il primo lavoro analizza gli effetti delle trasformazioni affini delle forme dell'orecchio sulle HRTF corrispondenti. Per realizzarlo, questo studio crea un database sintetico a partire dal (SYMARE), che abbiamo chiamato "modelli affini per la popolazione del SYMARE". Per i modelli affini, le forme dell'orecchio sono combinate in modo affine con la forma base dell'orecchio in modo da avere stesso orientamento, dimensione e posizione. Gli orecchi così composti sono attaccati alle forme base di busto e testa per creare un modello 3D chiamato modello affine. I benefici di creare un modello affine sono molteplici. Il primo e principale è di creare un paradigma semplicistico per studiare la morfoacustica dell'orecchio, limitando le variazioni solo alle variazioni della forma dell'orecchio e rimuovendo quelle relative alla forma di testa e busto, o di dimensione, orientamento e posizione delle orecchie sulla testa. Il secondo vantaggio Ã¨ che questo semplifica fortemente il processo di modellamento della forma dell'orecchio, in quanto bisogna modellare solo le variazioni usando LDDMM

e KPCA, non dimensioni e rotazioni. Terzo beneficio è che semplifica il processo di modellazione dell'acustica, se tutte le forme dell'orecchio sono della stessa scala, posizione e rotazione, e sono posti sulle stesse forme di testa e busto. Ad ogni modo, questo potrebbe creare artefatti che superano i benefici. Questo lavoro affronta queste domande. In questo lavoro, presentiamo uno studio che fornisce un'analisi di come semplici correzioni come un ridimensionamento delle frequenze delle HRTF e una rotazione dei loro pattern di direttivitÃ possono significativamente compensare per tutte queste trasformazioni affini. Inoltre, presentiamo la quantità di variazioni inter-soggetti provenienti dall'abbinamento affine confrontati con la forma originaria. Infine, Questo studio calcola il fattore di scala ottimale per la frequenza a partire da un punto di vista puramente acustico, che abbina nel modo migliore le HRTF modellate in modo affine a quelle originarie. Questi fattori di scala ottimali sono poi collegati ai fattori di scala fisici usando una regressione lineare. I risultati mostra che i fattori di scala possono essere dedotti semplicemente conoscendo la forma dell'orecchio e i fattori di scala che arrivano dal processo di abbinamento affine. Il secondo studio in questo gruppo fornisce un semplice metodo di modellazione basato sull'analisi dei componenti principali spaziali (SPCA) per analizzare le variazioni dei modelli di direttivitÃ acustica delle HRTF in funzione della frequenza. I modelli di direttivitÃ di frequenze diverse sono modellati separatamente e il numero di componenti principali richiesti per modellare i modelli di direttivitÃ per una data frequenza è quantificato per tutti i bin di frequenza nella gamma di frequenza da 0,2-17 kHz. Questo studio riafferma l'importanza dei modelli affini dimostrando che i modelli di direttività dei modelli affini possono essere descritti usando solo otto componenti principali a frequenze anche elevate fino a 17 kHz, mantenendo la differenza spettrale standard media (SDD) inferiore a 3 dB. Utilizzando il modello morfologico esistente delle forme dell'orecchio, questo lavoro modella le forme dell'orecchio con solo i primi otto componenti principali e mostrando risultati per alcuni padiglioni auricolari. Infine, utilizzando le otto componenti principali dello spazio della forma, stima i componenti acustici principali attraverso la regressione lineare per fornire un semplice metodo di personalizzazione delle HRTF. L'ultimo studio in questo lavoro fornisce una idea innovativa di ponderazione morfologica per creare un modello morfologico pesato per le forme dell'orecchio. Questo studio propone di assegnare pesi diversi a diverse porzioni dell'orecchio e di utilizzare un kernel pesato per eseguire una KPCA su dati LDDMM per creare un modello misurabile ponderato. I risultati di questo lavoro preliminare mostrano una migliore previsione per le componenti acustiche principali quando si utilizza KPCA pesato rispetto al KPCA tradizionale su dati LDDMM. Queste intuizioni sono molto interessanti e suggeriscono che con un ulteriore lavoro, questo strumento può essere utilizzato non solo per una migliore predizione delle HRTF personalizzati, ma potrebbe anche essere un modo efficace per comprendere i contributi di diverse parti delle forme dell'orecchio come una variante dell'analisi delle perturbazioni morfoacoacustiche.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This thesis work aims to develop simple HRTF personalization methods for personalized binaural reproduction of the spatial audio over headphones. Mainly this thesis studies the relationship between the morphology of human head, ear, and torso shapes. It performs various studies to understand the underlying phenomenons which are responsible for generating the individualized acoustic transfer function. This thesis is a composite study of multiple works from many fields, such as human acoustics, computational anatomy, signal processing, and data analysis. Using the existing knowledge from these fields, it describes a simple framework to map the morphology of a subject to the estimate of the corresponding set of HRTFs. It starts by providing a simple study to analyze the notch features in two of the most famous existing databases, namely CIPIC [14] and SYMARE [16] and provides insights on the evolution of notch features of HRTFs in the median plane. This knowledge can aid in the development of a preliminary HRTF personalization method. Continuing the exploration studies for HRTFs, this work analyzes the existing sparse representation based HRTF modeling and personalization methods. This thesis also reports on exploratory research that evaluates the relevance of various anthropometric features to the HRTF personalization process. Based on these relevance metrics, a novel method for weighted sparse representation based HRTF personalization is proposed, which provides comparable or even slightly improved results using fewer parameters.

The main contributions of this thesis are three morphoacoustic studies performed on the SYMARE database. These studies greatly involve the concepts and knowledge base built in [1, 17]. Using the foundations laid in these two works, the first study proposes an affine model for the head, torso, and ear shapes of the listers. In [1], authors have shown for few subjects that this affine model can provide simplifications in terms of modeling the corresponding acoustics and morphology of the listeners. In this work, we extend this study to all the subjects. Furthermore, the possible corrections

and compensations for these affine transformations are also studied. The second study analyzes the inter-subject variations in the corresponding HRTFs of these affine models using spatial principal component analysis SPCA as a function of frequency. Using the parameters for the ear shapes derived through the morphable ear model, the variations in the shape and acoustic spaces are related using linear regression. This study showed that the inter-subject variations for the spectral content at even very high frequencies can be captured using very few principal components again highlighting the benefits of the affine models. In the final study in this thesis, reports on a novel method for ear shape modeling using weighted kernel principal component analysis (W-KPCA). The results of this model show an improvement when used for the linear regression for the prediction of personalized HRTFs.

Following this chapter provides a brief background and the motivation to pursue this study. Sec. 1.2, highlights the problem statement. Sec. 1.3 and Sec. 1.4, provides contributions and the structure for the rest of the thesis. While Sec. 1.5 reports on the research articles and conference papers published and under progress, which are stemmed from this work. Spatial hearing is an ability of listeners to perceive the spread, distance, and direction of the incoming sound, as well as enables one to understand its surroundings, such as the size or properties of the room (cave, glass room, open area, etc.).

This ability comes as the result of the interaction of the sound signal, the environment, and mainly the anatomy of the listener, as the sound travels from the sound source to the eardrums of the listener. This interaction results in various physical phenomenons such as scattering, reflection, and refraction of sound energy waves, manipulating them differently depending on the direction of arrival, frequency, and distance from the eardrums. These spectral modifications provide psycho-perceptual cues to the brain to map the incoming sound to a particular position in the space. This ability to sound localization is developed in mammals through the evolution process of thousands of years and is essential in terms of both predatory and prey senses for survival. The generated psycho-perceptual cues fall into two categories. Cues inferred from the signal received at one ear cues or monaural cues and cues which are generated based on the signal received at both ears (as the difference in level or time of arrival of the signals) or binaural cues (refer to section 2.1.1 for more details). All these spectral coloration or transfer properties of the sound signals can be described by a set of impulse responses called head-related impulse responses (HRIRs). These HRIRs are different for both left and right ears. The frequency counterpart of these impulse responses is called head-related transfer functions (HRTFs) (refer to section 2.2 for details). Hence, HRTFs are the mathematical functions of distance, the direction of arrival, and frequency, which describe the spatial filter properties of the head, torso, and external ears for a sound source of a given frequency at a given distance and direction. Mostly distance is kept fixed, and the HRTFs are considered to be a function of frequency and direction of arrival only.

Knowledge of these transfer functions in hand enables one to present the spatial hearing experience to a listener by simply filtering the audio signal with both transfer functions and presenting on left and right ears, hence the term binaural hearing. However, as these transfer functions are the results of the interaction between morphology and anatomy of the listeners, these are not just dependent on the direction and frequency

but also the body of the listener and are unique for every listener. In the past, a lot of studies and experiments have focused on understanding the underlying psychophysical principals of the spatial hearing and on investigating the phenomenons which generate these transfer functions or spatial cues to create a personalized spatial hearing experience. Thanks to these studies, the importance and relevance of these cues are known these days, along with the frequency regions where they are relevant. The past studies show that binaural cues are mainly responsible for the sound localization in the horizontal plane and primarily work for the lower frequencies (i.e., up to 5 kHz for ILDs). These cues are mainly attributed to head shapes. While the monaural cues, which are generally the complex features of frequency in the form of deep notches and peaks in the high-frequency regions, are mainly contributed by the outer ear shapes (refer to section 2.1.1 for further details). This fact is further supported by the studies conducted in [18], which showed that the HRTFs for the occluded ear shapes, results in significant localization errors in the elevation plane.

Although there have been many studies to understand the reproduction of the binaural cues, the understanding of the generation of monaural cues is minimal in comparison. The possible reason for this is that the binaural cues mainly depend on the head shape, which is relatively more straightforward than the intricate and complex shape of the outer ear. Simple models can reproduce binaural cues by just knowing the head width, height, and depth. While there are no explicit models of such kind which relate the ear shapes to the corresponding complex monaural cues. The mechanisms that generate the monaural cues are much complicated to understand and require a more in-depth study to understand better and model the mapping between the ear shapes and monaural cues. This enforces the importance of the outer ear morphology and its effects on the HRTFs, and demands for a better understanding of the mapping between two to create a better HRTF personalization method.

1.1 Motivation

The listening experience through the traditional headphones lags in this aspect and fails to offer a quality 3D listening experience. Often hearing through headphones gives a notion as if the sound is coming from the center of the head. Even when the stereo, it can only pan the sound towards the left or right on the inter-aural axis, failing to provide the sense of true externalization and do not contain any spatial cues.

In a relatively sophisticated setup, a generic set of HRTFs measured on the manikins with average anthropometry for a given population, such as KEMAR or B&K are used through binaural rendering to provide a sensation of spatial hearing over headphones. The sound signal is convolved with the left and right HRTFs and presented as a two-channel signal at both ears. This reproduction of spatial audio over headphones is also called virtual auditory space generation [19, 20]. The psycho-perceptual experiments conducted in the past show that when spatial audio is generated using non-individualized HRTFs, it creates a bad listening experience resulting in problems such as lack of externalization, sound localization errors, up-down reversals, and front-back reversals. These studies suggested that to reproduce high-fidelity spatial audio over headphones; personalized HRTFs are to be used [20, 21].

The personalized HRTFs are traditionally acquired either through empirical measurements that require a big and expensive setup along with an expert and skilled

scientist or audio engineer. Alternatively, through running numerical simulations on the high-resolution 3D models of the human head, ear and torso morphology (refer to Sec. 2.3 for more details), which avoid the requirement of big measurement system but shifts the burden to the acquisition of the high-resolution mesh and a very power-hungry computation setup which can very well spend a day or more to acquire the HRTF of a single ear of the listener. This makes both of these modes of acquisitions for public mass usage unpractical, creating a bottleneck for the commercialization of personalized binaural rendering for the mass market.

This demands a simple and effective method to acquire and personalize the HRTFs for a listener based on easily gatherable information, such as low-resolution 3D models, ear images, and anthropometric features. The main driving force for this study is the ability to provide a personalization method. To the best of the knowledge of the author, currently, there are not any fast, accurate, and comprehensive methods to obtain the individualized HRIRs or HRTFs.

This study is a part of a large ongoing Australian Research Council (ARC) Discovery project, which aims at providing a framework that seeks to provide a comprehensive framework to understand the underlying phenomenons of HRTF generation and its relationship to various components of the ear shape. This limitation demands an in-depth study that can unveil the underlying functions and mechanisms that relate the ear, head, and torso morphology to the HRIRs of an individual. An accurate understanding of this will enable us to provide a method to obtain the individualized HRIRs for a listener in a fast and precise manner. The obtained HRIRs can then be of great use in medical and commercial applications (refer to section 2.4.1).

1.2 Problem statement and goals

Finding the relationship between the anatomy of the listener and corresponding HRTFs is essential to provide a personalized 3D hearing experience over headphones. However, due to the complex shape of the ears and the complexity of the corresponding acoustic features makes it challenging. Modeling and mapping the variations in the acoustic and shapes spaces is not trivial. The outer ear shape has many cavities and ridges which interact with the sound field in a unique way for every individual. These interactions are greatly dependent on the frequency and direction of arrival of the sound. Due to these interactions, the HRTF functions end up having complicated features in the form of peaks and sharp notches. The center frequency, width, and height/depth of these features play a vital role in generating psycho-perceptual cues for the listeners enabling them to localize a sound source in space. This complexity in both domains makes the task of finding a general model for both domains and a relationship between these models very difficult. However, it is necessary to understand to create a personalization method for HRTFs which work well for the mass market.

Even though the importance of this is well known, to the best of author's knowledge, there is no well defined and comprehensive framework available for the personalization of the HRTFs based on morphology. This framework should be able to understand the cues in HRTFs and the underlying phenomenons creating them to provide a personalized HRTF efficiently.

Although there are many existing methods for HRTF personalization which provide a great deal of understanding on the personalization process providing with some

knowledge about the underlying features in the HRTFs, which makes them unique for an individual, how these are generated, and basic mappings between the some of the features in the morphology to these acoustic features. However, they all have their limitations when it comes to using them for VAS (refer to Sec. 3.4 for details). For example, in the study [8], authors modeled the underlying mechanisms for first and third notch generation for a given HRTF using a simple parabolic sheet. In [11, 22] authors introduced micro perturbations on the surface of the outer ear and studied its effects on the notches and peaks in the HRTFs. The aim of these studies was to understand which features are sensitive to the perturbations in which parts of the ear shape. Similarly, in [23], authors took a rather sophisticated approach; they first analyzed the notches in the HRTFs for the median plan. They reasserted the fact that the center frequencies for these notches increase by increasing the elevation. Furthermore, they suggested that the three primary notches in the HRTFs are the result of the reflection of the sound waves from the three main contours of the ear shape, and proposed a mapping based on 2D ray tracing on these contours. While these findings are significant and relevant, considering the complexity of the ear shape and its significant variations amongst different listeners, an in-depth study is required to understand the relationship between these variations and the corresponding acoustics. To solve this problem, we propose a divide and conquer approach.

The problem of HRTF personalization can be divided into three smaller problems.

1. Modeling of the variations of ear shapes in a parametric way.

2. Modeling of the variations of the HRTFs in a parametric way.

3. Creating a map between the parameters of morphology and acoustics to create a personalization method for HRTFs.

The ear shape variations are modeled using the morphable ear shape model proposed in [17]. Using the powerful LDDMM framework, this work models the variations in the SYMARE database using KPCA. This simple yet powerful model lets one model any ear in the given ear population with only a few numbers as parameters. However, this works has a twist. As we are mainly interested in modeling the shape of the ears, all the ear shapes are affine transformed to the to match with the template ear shape [5] in size, position, and orientation. The motivation behind doing this is to simplify the modeling process of shape as well as the corresponding acoustics. The affine matchings used are scaling, rotation, and translation.

The first work analyzes the newly created synthetic database and explores the answers to the following research questions: 1) Does affine transformation simplify the modeling process of acoustics and/or morphology? If yes, can we quantify the simplifications? 2) Is this divide and conquer approach works, i.e., is it easy to model and compensate for the artifacts created because of these affine transformations? 3) Finally, can the correction parameters can be be obtained from the affine transformations [24–26].?

The second big step is to model the variations in the HRTFs. For this, we investigate the application of a frequency-dependent spatial principal component analysis based approach? This analysis examines the following questions: 1) Is the amount of variation captured using a given number of principal components is frequency dependent?

2) If yes, how many principal components are required to model the data for each of the frequency?

The final missing piece of solving the HRTF personalization problem is the mapping between both morphology and acoustic parametric models. So the last research question of this thesis work would be, can we use simple linear regression to relate the parameters in the morphology with the parameters in the acoustics?

Once this whole framework is in place, we also investigated if there is a way to explore the use of morphological weighting of the ear shape to create a better morphable model for the ear shapes. More specifically, can we put more emphasis on certain parts of the ears while creating a morphable model to model these parts better compared to the rest of the ear shape? Using this weighted KPCA model, we can improve the personalization of the HRTFs. Furthermore, this tool can be used to understand the relative contributions of each of the ear portions in the HRTFs?

1.3 Contributions

The ultimate goal of this study is to aid in creating a comprehensive and straightforward framework for HRTF personalization. This framework can be used to provide personalized sets of HRTFs for any subject avoiding the cumbersome and exhaustive numerical simulations. More specifically, in this thesis, we propose a simple method which uses parametric models for both outer ear shapes and the HRIRs/HRTFs, to efficiently and compactly represent the morphology and corresponding acoustics and enable one to obtain the personalized HRTFs for a listener without running the cumbersome simulations or going through laborious measurements.

This work has five major contributions:

1. It performs a simple statistical analysis of the median plane HRTFs for two databases CIPIC [14] and SYMARE [16]. This study uses simple signal and data processing techniques to analyze how the notch frequencies evolve as a function of the elevation angle. The findings suggest three things, 1) It verifies and reasserts the claim that the notch center frequencies are directly proportional to the elevation angle in the median plane. 2) The evolution of the center frequencies for three main notches is the same in two databases. 3) The notch frequencies for both left and right ears are not symmetric, suggesting that there might be some binaural cues that can help one to localize in the median plane (a novel finding). The results were published in [27].

2. Starting from the existing sparse representation based HRTF personalization technique [28], and knowing that not all the anthropometric features are equally relevant in creating the HRTFs, in this work we proposed a simple weighted-sparse representation based approach for HRTF personalization. This study finds the relative importance for each of the anthropometric features using an exhaustive search. Furthermore, this study compares the performance of the weighted-sparse representation with previously available approaches and some famous closest-matching based solutions. The results show that our approach outperforms the previous approaches as well as all the closest matching based approaches when spectral distortion based evolution is used. The results are published in [29].

3. A synthetic data-set is created to study the acoustic effects of the affine transformations on the ear shapes. In this work, a simple affine model is proposed, and the effects of the affine transformation on the ear shapes are studied and modeled. Then a simple correction method for the HRTFs of the affine models to obtain the HRTFs of the individual shapes is studied through frequency axis scaling like [25]. However, instead of using a single scaling factor, two scaling factors separate for head and ear regions are created, and a simple mapping between the ear and head scaling, and the optimal scaling factor derived from a pure acoustic point of view is created. The results are being prepared as an article to be submitted to the IEEE transaction of Audio, Speech, and Language Processing.

4. A simple method to analyze the variations in the acoustic directivity patterns of the HRTFs based on principal spatial components is proposed. This model analyzes the inter-subject variations as a function of frequency and models the directivity patterns of every frequency separately. Furthermore, this work analyzes how many principal components are required to model the directivity patterns of a given frequency. This work also analyzes how many variations in the ear shapes can be captured using the first eight principal components. Once the model is created, this work uses this model to create a simple mapping between the ear shape model parameters and the HRTF parameters. Part of this work is published in [30]. With some more analysis and further refinement, the work is to be submitted to some journals.

5. The final contribution of this work is the use of morphological weighting to understand the ear shape model better. In this work, we explore the potential for morphological weighting of different regions of the pinna (outer ear) to improve the prediction of acoustic directivity patterns associated with head-related transfer functions. Using a large deformation diffeomorphic metric mapping framework, we apply kernel principal component analysis to model the pinna morphology. Different regions of the pinna can be weighted differently before the kernel principal component analysis. By varying the weights applied to the various regions of the pinna, we begin to learn the relative importance of the various regions to the acoustic directivity of the ear as a function of frequency. The pinna is divided into nine parts comprising the helix, scaphoid fossa, triangular fossa, concha rim, cymbal concha, cavum concha, conchal ridge, ear lobe, and back of the ear. Results indicate that weighting the conchal region (concha rim, cavum, and cymbal concha) improves the predicted acoustic directivity for frequency bands centered around 3 kHz, 7 kHz, 10 kHz, and 13 kHz. Similarly, weighting the triangular and scaphoid fossa improves the prediction of acoustic directivity in frequency bands centered around 7 kHz, 13 kHz and, 15.5 kHz. The results are published in proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2019 [31].

1.4 Thesis Structure

This thesis consists of seven chapters in total. The first three chapters are introductory and provide the reader with the introduction, background, and an overview of the

existing literature and state of the art for this study; the next three chapters are the contributing chapters that contain the research work performed by the author in the course of this Ph.D. Finally, the seventh chapter sums it all up and provide a conclusion of the dissertation. Following is an overview of what each chapter contains.

Ch. 2 equips the reader with the basic and necessary background knowledge to understand the work. It starts by providing the introduction to the spatial audio and explains what it is in Sec. 2.1. In Sec. 2.1.1 it talks about the localization cues used by the listener for spatial hearing. In Sec. 2.2 a simple overview on the HRIRs and HRTFs is provided. Sec. 2.3 comments on the methods to obtain the HRTFs. This is followed by a section on the Virtual Auditory Space(VAS) and its applications in Sec. 2.4. The purpose of discussing this here is to give the reader a general idea of how this study fits within a larger prospectus of the spatial audio field and why it is important to solve the research problem under consideration. In Sec. 2.5, 2.6, and 2.7 the details on Principal Component Analysis (PCA), Large Deformation Diffeomorphic Mapping Metric (LDDMM) framework and Kernel Principal Component Analysis (KPCA) are provided. These details are essential to understand the work performed in this work. The concepts provided in this chapter are widely used in the rest of the thesis.

Ch. 3 reports the literature reviewed. This chapter is aimed to equip the reader with a brief review over the state of the art shape on shape modeling in Sec. 3.1, use of LDDMM on the ear shapes in Sec. 3.2, morphoacoustic perturbation analysis (MPA) in Sec. 3.3 and HRTF modeling, and personalization techniques in Sec. 3.4. Finally, it provides some of the different measures used in shape and acoustic space in Sec. 3.5.

Ch. 4 provides the details on the preliminary studies conducted in this thesis. In Sec. 4.1 the details on a simple notch analysis technique along with the results and findings are provided while in Sec. 4.2 presents the weighted-sparse representation based personalization technique developed in this work.

Ch. 5 presents a simple model for 3D shapes to obtain the HRTFs either through BEM simulations or some kind of modeling. Sec. 5.1 provides the details on how the affine models for a subject are created using LDDMM and shows how the scale factors for head and ear shapes are measured. Furthermore, it shows how these factors sit with respect to the template head and ear size. Sec. 5.2 quantifies the simplifications caused by the affine matching of the ear shapes. Sec. 5.3 shows how to find the optimal scale factors for the head, ear, and whole frequency range. Finally, it concludes the chapter explaining how these optimal scale factors are related to the physical scale factors.

Ch. 6 provides details on the spatial principal component analysis performed to study the variations in acoustic transfer functions as a function of frequency. Sec. 6.1 shows how the directivity patterns are preprocessed to perform this analysis. Sec. 6.2 describes how these directivity patterns are modeled frequency by frequency. Sec 6.3 quantifies the number of principal components required to model the directivity patterns for a given frequency. Sec. 6.4 provides a simple analysis of using SPCA in the view of personalization. Finally, the Sec. 6.5 provides the details on weighted KPCA and shows how the results of this study can be used to understand the contributions of each part of the ear in HRTFs.

Finally, Chapter 7 provides the concluding remarks for the thesis revisiting the contributions along with indicating some future works and challenges faced during this work.

1.5 Publications

1.5.1 Published

These papers are provided at the end of the thesis.

- **M. Shahnawaz**, L. Bianchi, A. Sarti, S. Tubaro, "Analyzing Notch Patterns of Head-related Transfer Functions in CIPIC and SYMARE Databases", 24th European Signal Processing Conference (EUSIPCO), (pp. 101-105), 29 Aug.-2 Sept. 2016, Budapest, Hungary.
- M. Zhu, **M. Shahnawaz**, A. Sarti, "HRTF Personalization Based on Weighted Sparse Representation of Anthropometric Features", International Conference on 3D Immersion (IC3D), (pp. 1-7), 11 Dec. 2017, Brussels, Belgium.
- C. Jin, R. Zolfeghari, X. Long, A. Sebastian, S. Hossain, J. Glaunes, A. Tew, **M. Shahnawaz**, A. Sarti, "Considerations regarding personalization of head-related transfer functions", International Conference on Audio, Acoustic and Speech Processing (ICAASP), (pp. 6787-6791), 15-20 April 2018, Calgary, AB, Canada.
- **M. Shahnawaz**, C. Jin, J. Glaunes, A. Tew, A. Sarti, "Morphological Weighting Improves Individualized Predictions of HRTF Directivity Patterns". IEEE Workshop on Application of Signal Processing to Audio and Acoustics (WASPAA), (pp. 75-79), 20-23 Oct. 2019, New Paltz, NY, USA, USA.

1.5.2 In Preparation

- **M. Shahnawaz**, C. Jin, X. Long, R. Zolfaghari, J. Glaunes, A. Tew, A. Sarti, "An Affine Model of 3D Models for HRTF Modeling", being prepared. and most probably to be submitted to IEEE/ACM Transaction on Audio, Speech, and Language Processing.

1.5.3 To be Written

- **M. Shahnawaz**, C. Jin, A. Sarti, J. Glaunes, A. Tew, "SPCA based Modeling of HRTFs", work under progress and to be written to submit as a journal or transaction.

CHAPTER 2

Background

This chapter provides a brief overview of the key concepts, techniques, and methods from past studies that are extensively used in this work. It is essential to revisit these details to understand this thesis. The main aim of this work is to broaden the understanding of the HRTF personalization process, which will enable every individual to have an immersive and high-quality spatial hearing experience over headphones. As mentioned in the previous chapter, a substantial part of this thesis work is the continuation of the work presented in [1], so some of the sections in this chapter are extensively borrowed from [1]. This chapter covers the topics which will assist a reader in understanding the work conducted in this study. Moreover, it will get the reader familiarized with the relevant concepts and techniques widely used in the field of spatial hearing, human acoustics, statistical shape, and data processing.

At the start, a brief overview of the spatial hearing and the sound localization cues is provided. These cues enable an individual to perceive the size and volume of the sound sources and localize them in 3D space (Sec. 2.1). Following this section is the description of the Head-Related Impulse Responses (HRIRs) and Head-Related Transfer Functions (HRTFs) in Sec. 2.2. Subsequently, the details on the processes to acquire individualized HRTFs are provided in Sec. 2.3. Sec. 2.4 discusses what virtual auditory space (VAS) is and highlights the applications of VAS in various fields. Sec. 2.5 provides the introduction to the principal component analysis (PCA) and shows its application to a set of HRTFs.

This morphological modeling in this study and largely in this project is revolving around Large Deformation Diffeomorphic Mapping Metric (LDDMM), so Sec. 2.6 provides a detailed introduction to LDDMM. Furthermore, it also highlights the application of the LDDMM framework in the context of ear shape modeling [32]. In Sec. 2.7 we provide some introduction to the kernel principal component analysis (KPCA).

Although it is advised to go through this chapter to review the concepts and used

notations in the thesis, readers already familiar with these topics can skip this chapter and go to Ch. 3 right away to see the literature review.

2.1 Spatial Audio

Thanks to the powerful and complex human vision system, humans are mainly visual-oriented. However, with all its might and benefit visual sense is limited when it comes to observe or feel objects behind the head or at situations when the lighting is not adequate. In contrast, humans can perceive and localize sound originating from all directions in space. This ability of humans listeners to perceive the location (direction and distance), the spaciousness of the sound sources in space, and the acoustic properties of the environment is called "spatial hearing" [33]. The spatial hearing compliments the sense of vision and helps humans to interact with their surroundings more effectively. On top of it, just like the visual sense, the spatial hearing also lets the listener focus and concentrate on a single audio source in a particular position in the space.

A real-life example of seeing spatial hearing in action is the conversations in cocktail parties [34]. In such cases, there are multiple talkers, and many competing sound sources exist around an individual listener. In the absence of the spatial hearing, the listener will not be able to distinguish the direction of sound sources, and it will put him in an awkward situation by making it very difficult to respond to the speakers appropriately. This example shows that without the spatial hearing, the quality of an individual to have a social interaction is severely limited. The same is the case with playing games and navigating through scenes and mazes in the game, where players have a limited point of view, the spatial hearing can add a great deal to the experience.

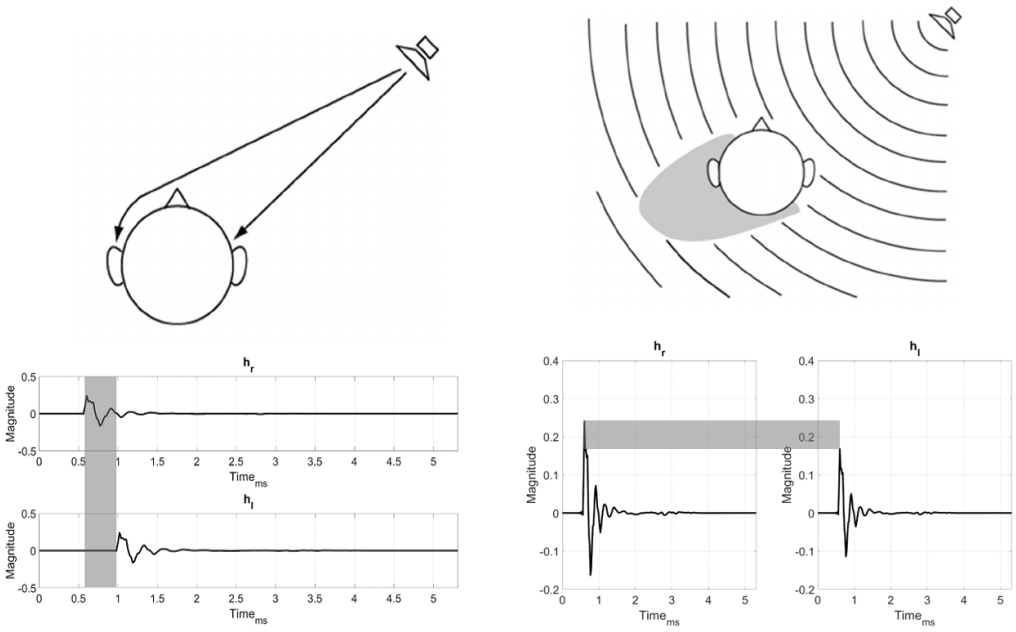
When it comes to the accuracy of sound localization in space, the past studies suggest, an average human listener can localize an audio source in space very accurately and precisely. The precision and resolution with which the sound sources can be localized are different for azimuthal and elevation planes. For example, in the azimuthal planes, a human listener can distinguish between the direction of arrival for a sound signal with a resolution of 1° to 3° on average [35]. On the other hand, localization of the sound source in elevation plane is more complicated and depending on the sound source properties and stimuli humans can only localize a sound source in elevation plane with an average resolution of 4° (for white noise) and 17° (for speech stimuli) [35,36]. These values for minimum resolution angles are called minimum audible angles (MAA). The ability of spatial hearing is the result of the spatial cues generated by the interaction of the sound field with listeners' anatomy and surroundings on its way to the eardrums of the listener. Following, we provide a brief introduction to these cues.

2.1.1 Sound Localization Cues

The studies conducted in the past generally put spatial audio cues into two main categories. The first set of cues are the cues that are inferred from the sound signal received at both ears, hence called binaural cues. While the other set of cues are monaural cues, the cues which are generated or can be interpreted by using just single ear data. These cues are also called spectral cues [37, 38].

Binaural Cues

The spatial cues which are inferred by utilizing the sound signals in both left and right ears are called the binaural cues. Binaural cues are one of the oldest spatial hearing cues known by humans, firstly introduced by Lord Rayleigh, more than a century ago in [37]. The underlying phenomenon for the generation of these cues is the physical separation between two ears by the head. This separation results in the difference in the time of arrival and received signal intensity in both ears, in a location depending way. For example, the sound coming from a sound source in the right direction from the head of the listener has to travel more distance to reach the left ear than to the right ear. Hence, the signal is delayed when it reacher to the left eardrum compared to the right eardrum. This difference in time of arrival is called inter-aural time difference (ITD). Also, the head casts a shadow to the signals coming from the right direction, which results in a smaller value of the sound intensity for the received signal in the left ear than the right ear. This difference between the received sound intensity or level is called inter-aural intensity difference (IID) or inter-aural level difference (ILD). Both of these cues are very important and useful for sound source localization on the horizontal plane. The ITD is used for localizing the low-frequency sounds when the wavelength of the signal is comparable to the size of the head. While the ILD is used for higher frequencies when the wavelength becomes smaller than the size of the head [39].



(a) *ITD* **(b)** *ILD*

Figure 2.1: *This figure shows the binaural cues in play. a) Inter-aural Time Differences (ITD), (the magnitudes of both HRTFs are normalized to only show the time delays) and b) Inter-aural Level Differences (ILD).*

Monaural Cues

The other type of spatial cues that are important for spatial hearing is the monaural cues. These cues are the results of the spectral coloration imprinted in the sound signal

due to the interaction between the sound field and the head, torso, and the ears of the listener. As these cues are different for left and right ears, they are called monaural cues [2, 40]. The previous studies report that the notches and peaks in the monaural cues in lower frequencies are contributed by head and torso. In contrast, the notches and peaks in the high frequencies are mainly contributed by ear shapes [2, 40].

An example ear is shown in Fig. 2.2. This figure shows that the outer ear shape is a complex surface with multiple ridges and cavities. Each of these cavities plays a role in the generation of the spectral coloration in a frequency and direction-dependent way [41]. This happens because upon getting reflected from these surfaces, the sound waves get delayed in comparison to the direct path sound and depending on the direction and frequency the cause either constructive or destructive interferences, creating peaks and notches.

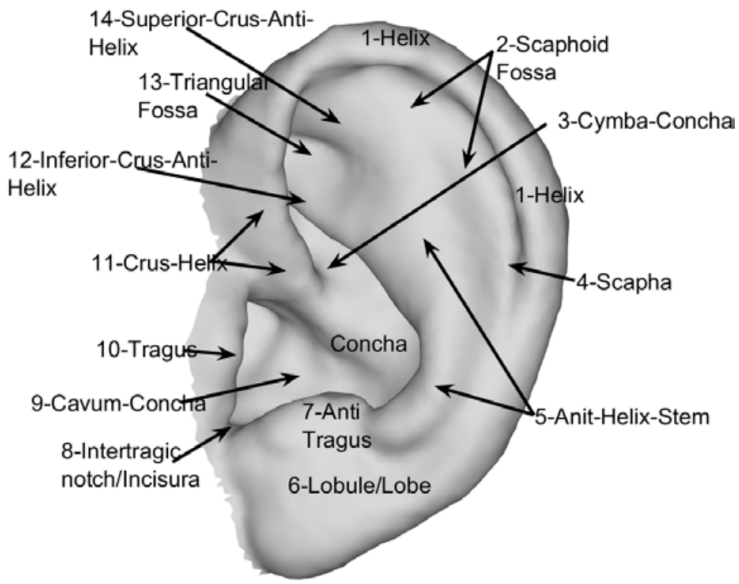


Figure 2.2: *This figure shows an image of the left ear of a human subject with annotations indicating different parts of the ear shape. Picture taken from [1].*

The previous studies have shown that although all torso, head, and ear shapes play role in the generation of monaural cues, when it comes to the perceptual relevance, the outer ear shapes are the main contributor [18, 42, 43].

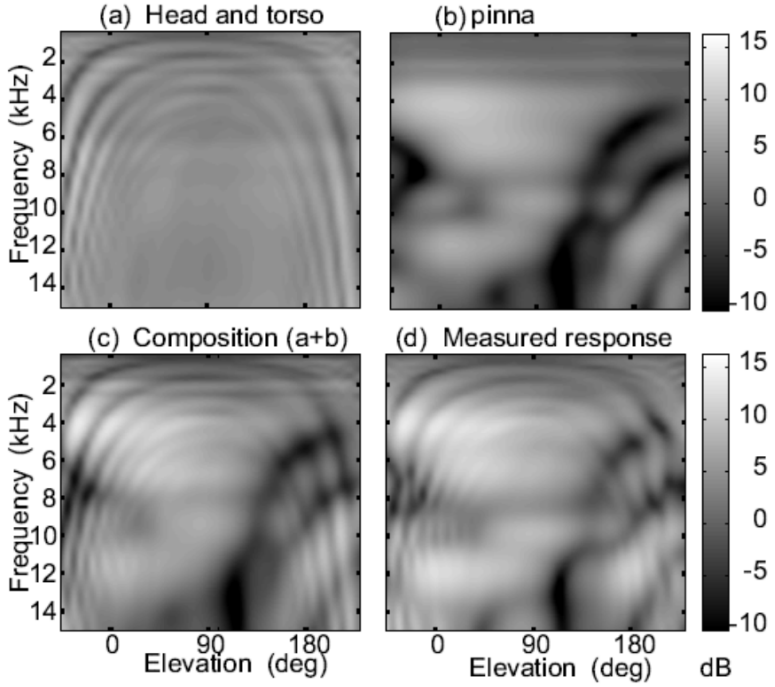


Figure 2.3: *The acoustic transfer functions corresponding to three different shapes are presented in this figure for a cone of confusion located at azimuth angle $\theta = 25°$. (a) pinna less KEMAR, (b) The acoustic response of the pinna shapes(PRTF), (c) The sum of (a) and (b). d) The measured acoustic response of KEMAR head, ear, and torso shape. Image reprinted from [2].*

This argument is further strengthened when we have a look at Fig. 2.3. In this figure, (a) shows the acoustic response recorded for KEMAR head and torso shape without an outer ear attached. While the response recorded for just the outer ear shape is reported in Fig. 2.3(b). Fig. 2.3(c) shows the sum of first two, while in Fig. 2.3(d) presents the response recorded on head, torso, and ear shape is provided. Having a look at these figures, it is evident that the outer ear shapes make the main contributions in the spectral coloration. This finding inspired some of the scientists working in this field to work only on the acoustic responses of the outer ear shape, ignoring the effects of head and torso shapes [3, 23, 40, 44]. The same is the case of this study. In this study, we mainly focus on the morphoacoustics of the ear shapes, paying little to no attention to head and torso shapes.

It is generally believed the cues for localization in the median plane are the first peak and first two notches. To further verify that these cues are generated by the ear and not the head and torso, authors in [3], numerically calculated the acoustic responses for some human subjects, by running finite-difference time-domain method on the ear shape only and, head and ear shape attached. The results for four subjects, two males M_1 and M_2 and two females F_1 and F_2 are reprinted in Fig. 2.4. The first columns show the acoustic responses calculated for head and torso meshes, which are also called HRTFs. While the second column contains the figures showing the acoustic responses calculated only on the ear shapes, which are also called Pinna-Related Transfer Functions (PRTFs). It is quite evident from these figures that even in the absence of the

heads, PRTFs manage to capture most of the variations which are present in HRTFs, further affirming that the outer ear shape contributes the most of the acoustic features in the head and ear acoustic responses (HRTFs) [2].

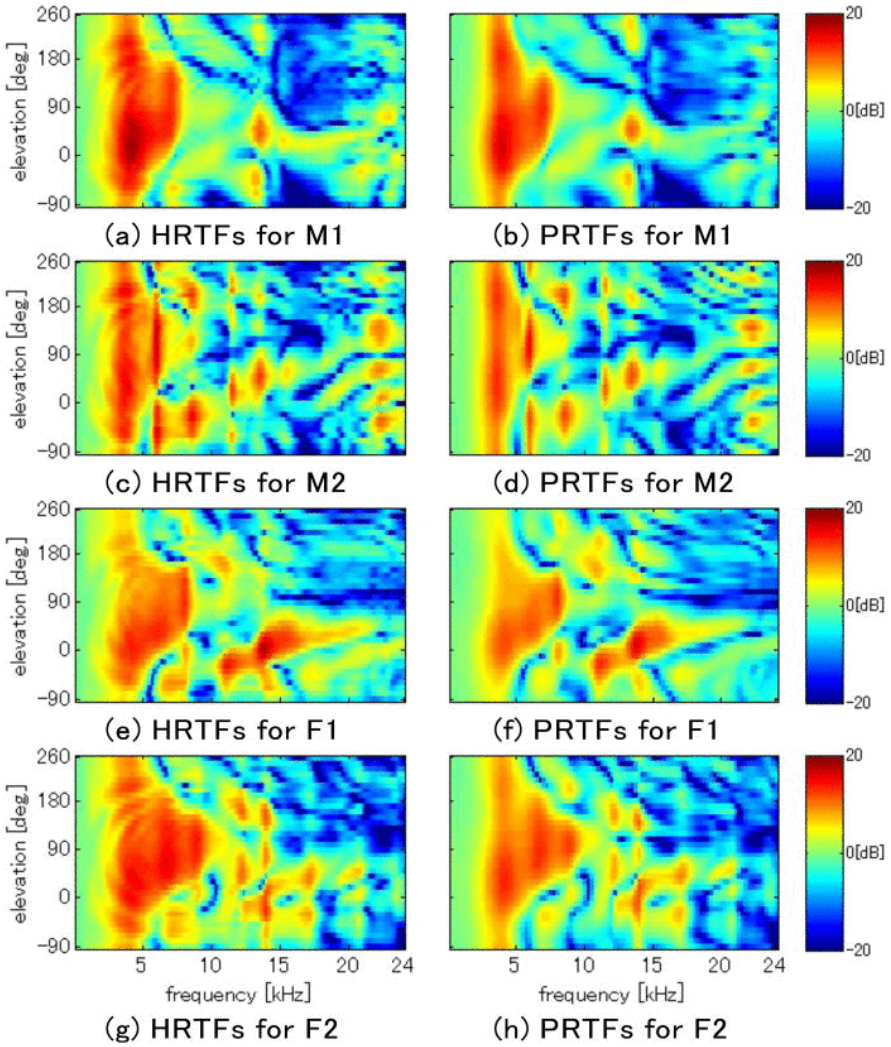


Figure 2.4: *The first column contains HRTFs for two male subjects M1 & M2 and two female subjects F1 & F2. The second column contains the PRTFs for these subjects. Image reprinted from [3]*

.

The notches and peaks shown in these figures are primary cues for elevation perception and are produced as a result of the interaction of the sound field with the complex geometry of the pinna. It would be almost impossible to have elevation perception om the absence of these notches. This was confirmed by the studies in [18], reporting that when the responses recorded on the occluded ears are used for the localization tests, the sound localization is almost impossible. Furthermore, psycho-perceptual experiments conducted in [45, 46] suggest that removing notches from the sound signal deteriorates the spatial hearing experience and considerably reduces the ability to localize audio sources. All these studies confirm the importance of the outer ear shape and its acoustic implications in the spatial cues. Considering these findings of these studies in this work, we focus our studies only on the ear shapes and their corresponding acoustics.

2.2 Head Related Impulse Responses (HRIRs) and Head-Related Transfer Functions(HRTFs)

The Head-related Impulse Response (HRIR) is a finite impulse response that mathematically explains the spatial acoustic filtering properties of the head, torso, and the ear shape for a given direction in space. These impulse responses describe how the sound signal transforms on its way from the sound source, sitting at an arbitrary position in space to the eardrum [47,48]. A pair of HRIRs for the left and right ears contain all the binaural and monaural cues required by a listener for spatial hearing. The frequency-domain counterpart of the HRIRs is obtained by performing the Fast Fourier Transform (FFT) and is called the Head-Related Transfer Function (HRTF). The HRTF represents gains and losses for the sound signals at a given frequency on its way from the sound source to the eardrum for a given direction. It is a complex-valued function and has a magnitude and phase at each frequency and direction. Directions in space are usually identified by the head centered spherical coordinate system. The directions have two coordinates which are donated by θ called the "azimuthal angle" and ϕ or the "elevation angle". The coordinate system for these angles is shown in Fig. 2.5.

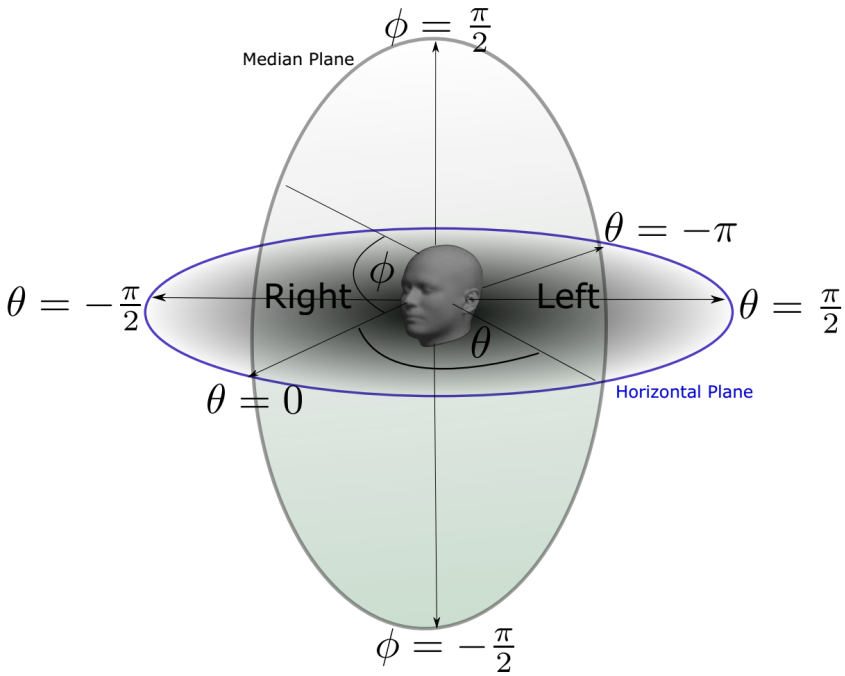


Figure 2.5: *A head centered coordinate system showing the auditory angles. The angle θ denotes the horizontal or azimuth angle, while ϕ denotes the vertical or elevation angle. Image taken from [1]*

HRTFs exist for each direction for both left and right ears and are unique. Fig. 2.6 shows a plot of the left ear HRIR and its corresponding HRTF for a given direction $(\theta = 0°, \phi = 0°)$ of the space for Subject 1 in SYMARE database.

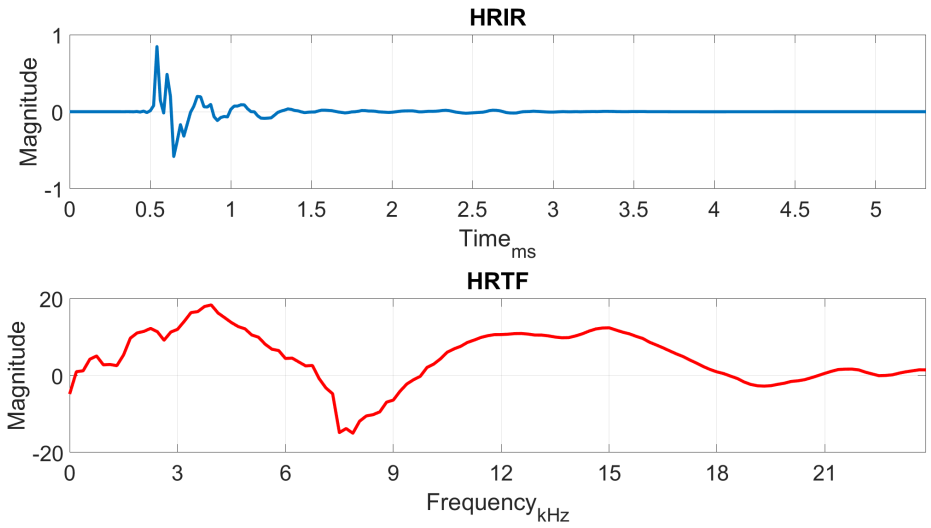


Figure 2.6: *The head-related impulse response (HRIR) and the head-related transfer function (HRTF) is shown in the above plots for azimuth angle* $0°$ *and elevation of* $0°$ *for Subject1 in SYMARE database.*

These two representations show the variation in the gains as a function of time and frequency for a given direction. However, we can also show the gain or directivity of the ear shape for all the directions in space for a given frequency, as shown in Fig. 2.7. This representation is known as the directivity pattern and is represented as a Spatial Frequency Response Surface (SFRS) [49]. Figure 2.7 shows the directivity pattern for subject 1 at $6kHz$ for left ear.

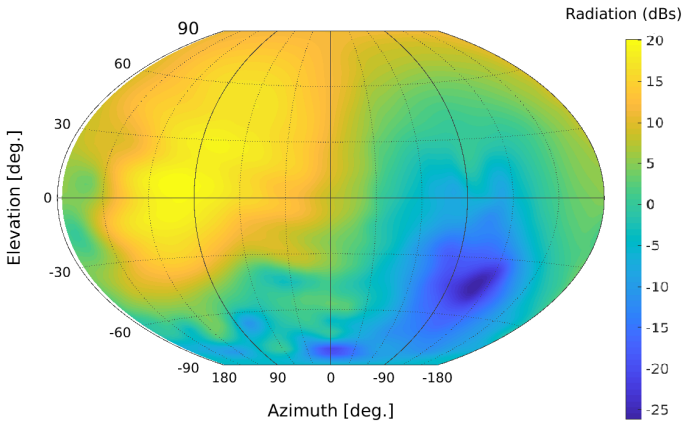


Figure 2.7: *The SFRS plot shows the directivity pattern for Subject 1 at 6 kHz for the left ear. Positive values of azimuth angles correspond to the ipsilateral side, and the contralateral side is represented by negative values. It can be seen from the SFRS plot that the left side shows higher gains than the right side. Also, the upper quadrant in the left side has higher gains than the lower side due to the shadowing casted by the torso of the subject.*

(a) *HRTF Measurement Setup 1* **(b)** *HRTF Measurement Setup 2*

Figure 2.8: *The above images show the HRTF recording setups. In (a) the individual sits in the chair and the loudspeaker arc is rotated around him and in the (b) the loudspeaker arc is fixed and the user is rotated with the help of a rotation table.*

2.3 Acquiring Individualized HRTFs

This section provides details on how to acquire the individualized HRTFs for a subject.

Traditionally and to date, the most accurate method to obtain the individualized HRTF is to measure it through acoustic measurements. In this method, the HRTFs of an individual are extracted from the acoustic recordings made using miniature microphones placed in the ear canals of the listener and playing known signals from loudspeaker by placing it at different locations [14, 50, 51]. The recordings are made in anechoic chambers to emulate the free space conditions. As mentioned in section 2.2, HRTFs represent the transfer characteristics for the sound as it travels from one particular location to the ear canal of the listener. The recordings are required to be made for every single location for which HRTFs is needed. These recordings are generally performed in one of the two most popular settings. In the first setup, the listener sits in the middle of the room, and the single or multiple loudspeakers are moved around the user with the help of a robotic arm. While in the second setup, multiple speakers are placed on a fixed vertical circular arc, and the user is rotated with the help of a rotation table. In both setups, the user is supposed to avoid any kind of movement and always keep his head still. Fig. 2.8a and Fig. 2.8b shows these two setups respectively.

For high-quality recordings, in both settings, not just room but also the apparatus involved has to be anechoic, and covered with the same material that covers the walls, roof and floor. The equipment includes the robotic arm, the loudspeakers, and the sitting chair. These two settings have a very strict requirement for user to keep his head still for the whole measurement process. Any involuntarily made movements will distort the acquisitions and the whole recording process is to be repeated all over again. Keeping ones head still for this long is almost impossible. Some works use a guidance system for

Figure 2.9: *HRTF recordings setup at Auditory Localization Facility at Wright-Patterson AFB, Dayton, OH. Image taken from [4]*

the head tracking while the measurement process to help user to keep his head still via a feedback system and head tracker [51]. While some other research groups have focused on making the measurement process faster by using multiple speakers and innovative measurement techniques [50]. An extreme solution to this problem is also used by US air force labs. In there setup they placed loudspeakers at every location, which shortens the measurement process time by removing the need of moving of speakers or user [4]. Fig. 2.9 shows this setup.

Although a combination of these two studies together can provide a very fast and accurate measurement technique, acoustically measuring the HRTFs is still relatively expensive and time-consuming method. Also, the requirement of anechoic chamber and such a cumbersome setup limits the acquisition to a single place or lab as it is almost impossible to move it. All these limitations make it impractical for commercial use and limit it to laboratory use only.

The second and more flexible method for getting individualized HRTFs is, to numerically calculate them by solving boundary integral equations, generated based on a high-resolution 3D model for the head, torso, and pinna of the listeners [51–53]. In this thesis most of the studies are performed on the data coming from this process, as well as, this process is also used to get the HRTFs for a synthetic dataset created in this thesis. So, to provide readers a brief review following provides details on this method.

2.3.1 Numerical Simulations for HRTFs

Numerical solutions for obtaining the HRIRs involve numerically solving and computing the solution to the Partial Differential Equation (PDE) that governs the scattering of sound waves around objects in three-dimensional space. The PDE for the scattering of sound waves is known as the Helmholtz PDE [54, 55]. The use of numerical simulations to get the HRTFs is a very interesting topic for many scientists for multiple reasons. First and foremost, it lets one, to avoid the cumbersome measurement process to get the individualized HRTFs. The second reason is the recent advent of more powerful and fast computing machines, and the advancements of the algorithms such as the fast multiple method (FMM) [52] make this reasonably fast. The third reason is that in these days, obtaining the high-quality 3D model has become feasible due to the availability of various forms of scanning, such as magnetic resonance imaging (MRI), laser scanning, and computed tomography (CT Scan). Some labs have also started to use photogrammetry for this purpose.

There are different kinds of numerical simulations which can be used to calculated the HRTFs. For example, authors in [22] used the Finite-Difference Time-Domain Method (FD-TDM) to simulate the HRIRs of pinna after applying micro-perturbation in a different position on it. The problem with this kind of studies is that they require volumetric data for the 3D models to voxelate the model and surface data only is not enough.

Another very famous method is the Finite Element Method (FEM). The studies presented in [53, 56, 57] used FEM to calculate the HRTFs numerically. However, the problem with FEM is that it not only requires one to model the 3D morphology of the listener but also the space around it needs to be sampled to include the loudspeaker positions in the analysis. The third most widely used method for the HRTF calculations is the Boundary Element Method (BEM). Many studies has used this method included [51, 52, 58]. In this work and another thesis [1], which is of same nature, Fast Multipole BEM or FM-BEM, a faster and more accurate version of traditional BEM is used. Following we provide details on getting the HRTFs from FM-BEM simulations. Parts of this sectioin are borrowed from [1].

Preparing meshes for BEM simulations

The quality of obtained HRTFs by the numerical simulations is directly proportional to the resolution of the used meshes, i.e., for a mesh with more number of mesh elements, the result will be more accurate when compared to a coarser mesh [1]. An example of this is shown in Fig. 2.13, showing results of BEM simulations for same ear in high and low resolution settings. However, a denser and high-resolution mesh will require more computational resources and longer simulation time. This demands to find a trade-off between the number of mesh elements to get reasonably accurate results, and the computation time required to perform numerical simulations. This section details on an iterative mesh coarsening process to prepare the meshes for BEM simulations, which provide a good and accurate acoustic response to a given maximum or critical frequency. After passing through this procedure, each mesh is optimally coarsened and converted to a uniformly sampled mesh to reduce the number of mesh elements and simulation time, without compromising the accuracy of the results.

The study presented in [59] reports the criteria for preparing such a mesh in great detail. The findings of this study suggest that to obtain an accurate set of HRTFs from

FM-BEM simulations, the mesh must have a certain minimum resolution for a given maximum or critical frequency up to which the simulations have to be accurate. Furthermore, all edges in the mesh should be close to of the same size, and there should not be any triangles that are very long and pointy, with one of the edges being a lot smaller than the other two edges. Finally, it emphasizes the need for having a smooth and uniform sampling of the mesh. In this work, an iterative process of remeshing is performed using a remeshing software called ACVD [60]. After passing through this software, each of the meshes must satisfy the following four criteria to be BEM ready mesh.

1. The smallest edge should not be smaller than one-fifth of the largest edge.
2. The minimum angle between any two edges should be greater than 15°.
3. The maximum angle between any two edges should be less than 150°.
4. The length of the longest edge should be smaller than one-sixth of the wavelength corresponding to the largest frequency.

Although, ACVD is generally used for coarsening and down-sampling the high-resolution meshes, it can also be used for the opposite, i.e., for remeshing or subdividing triangular elements for the upsampling purpose. The main parameter used as an input to the ACVD software is the number of vertices N_v in the resulting mesh. Using this value, ACVD tries to create a uniform mesh such that the output mesh is an approximate to the input mesh to a given threshold and have the number of vertices equal to or very close to N_v. If the number of vertices in the mesh is more than N_v, the mesh is coarsened, and if the number of vertices is smaller than N_v, the new triangles are created by applying the surface subdivision first resulting in more number of vertices.

This process is performed using a MATLAB script that iteratively checks the quality of the output mesh against the requirements one through four. This adjusts the number of vertices allowed N_v iteratively by checking if the resulted mesh can be used for the given critical frequency of f_c. In this work we used the same scripts with little to no modifications as were used to create SYMARE database [1, 51]. For every iteration, it prepares the mesh, which has $N_V(I)$ vertices in it. This mesh is supposed to provide us with the accurate results up to f_c or critical frequency of the mesh. The critical frequency is calculated using Eq. 2.1. This frequency is used as the input to the MATLAB script. The script adjusts the value for $N_v(I+1)$, at every iteration, and input it to the ACVD program in the next iteration to provide a mesh that has a critical frequency, which is equal or higher than the target critical frequency f_c^{targ}. The number of vertices for the next iteration is calculated using the expression given in Eq. 2.2.

$$f_c = \frac{c}{6e_{max}} \tag{2.1}$$

$$N_V(I+1) = N_V(I)(\frac{f_c^{targ}}{f_c})^2 \tag{2.2}$$

In Eq. 2.1, e_{max} denotes the length of the longest edge in the mesh. For our simulations we set the critical frequency $f_c = 26kHz$. It is observed that for this critical frequency, the average edge length (AEL) for the meshes is below $1.5mm$. HRTFs obtained by running the FM-BEM simulations on these meshes are shown to be in good agreement

with the acoustically measured HRTFs [51,61]. It is to be noted that due to the iterative nature of the MATLAB program and also because the output mesh shapes have to satisfy criteria $1-4$ specified above, the resultant meshes can have a critical frequency value that is higher and not equal to $26kHz$.

Finally, the created meshes are passed through a final cleaning process. This cleaning is performed using Meshlab [62]. The cleaning process does three things:

1. It removes all the duplicated vertices.
2. It removes all the unreferenced vertices.
3. It removes all the non-manifold edges.

After passing through this final cleaning process, the meshes are now ready to be used for FM-BEM simulations.

Performing BEM simulations

There are various ways to perform the BEM simulations, i.e. someone can write there own scripts or can use some commercial tools for it. In this work all the BEM simulations were performed using Coustyx [63], a commercial software for advanced numerical solutions provided by Ansol. The same software was used to obtain the SYMARE database originally in [51]. Coustyx lets one to use reciprocity principle based simulations in which during numerical calculations of the HRTFs, the virtual loudspeaker (a vibrating source) is placed in the ear canal, and virtual sensors are placed at the desired directions in space. The benefit of using this technique while running the BEM simulations is that the HRTFs for all the directions can be computed simultaneously, reducing the simulation times by multiple orders. The use of the reciprocity principle was first validated by [53]. The virtual sound source is simulated by cutting an element of the triangular mesh loose and vibrating it to generate the signal for a given frequency. The position of the vibrating element is chosen by finding the intersection of the inter-aural axis with the gear mesh. In Fig. 2.10 the position of a vibrating element is shown for the left ear. Fig. 2.10 shows a section of head mesh prepared for BEM simulations, showing the vibrating element in red color, and the inter-aural axis by a green line. The intersection between the inter-aural axis and the mesh is presented with a blue point. The element containing this point is cut loose and treated as a vibration element, as shown in the figure. Keeping the center of the head as origin, an imaginary sphere is created of a radius equal to $1m$. This imaginary sphere is sampled uniformly using Icosahedral subdivision [64] to have 2562 uniformly spaced points on the sphere. These points are used as the positions for the virtual sensors, and HRTFs for all these positions are calculated. Fig. 2.11 shows an imaginary sphere of such kind.

The HRTFs are then calculated from the results of the numerical simulations, using the Burton-Miller BIE formulation using the Galerkin implementation. The benefit of using this method is that it provides a better accuracy compared to other formulations while solving the Helmholtz PDE using the BEM [1]. This work uses a special kind of BEM simulations called Fast Multipole Multi-level BEM method. To compute the number of FM levels N_{FM}, first, the dimension r_{HE} and average edge length e_{avg} of the object (head shape) is calculated. These values are then used in the Algo. 1 to find the value of levels required N_{FM}. The FM-BEM implementation in Coustyx uses the Generalized minimal residual method (GMRES) iterative solver for the FM

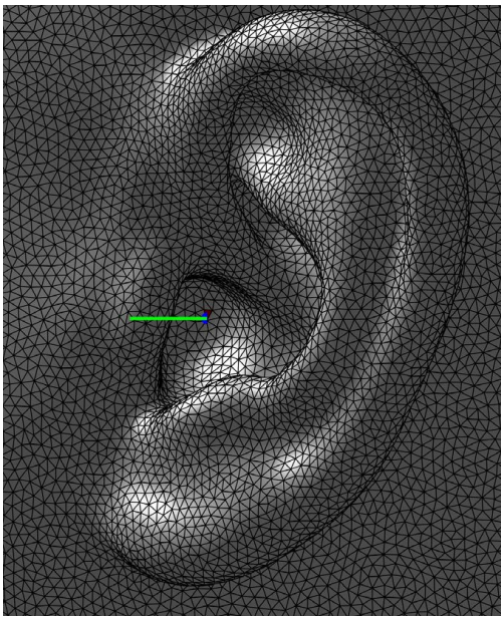


Figure 2.10: *This figure shows a screen shot of the a portion of head mesh in which the vibrating element is shown in red and the inter-aural axis is represented with a green line. The blue dot shows the point of intersection between the inter-aural and the 3D model.(Picture taken from [1]*

computation on the BEM system of equations, which accelerates the convergence to a solution.

Algorithm 1 Calculating the required number of levels for FM-BEM simulations.

inputs: r_{HE}, e_{avg}.
outputs: N_{FM}.
$L \leftarrow 20$.
for $l = 2$ to L **do**
  if $2 \times r_{HE}^{l} - e_{avg} \leq 0$ **then**
    $N_{FM} = l$
    $break$;
  end if
end for

Extracting acoustic data from BEM simulations

Coustyx provides the results of FM-BEM simulations on the head, torso and ear surfaces of the affine models as complex pressure values for given frequencies f at given directions represented by angles (θ, ϕ). These results are denoted as $\Phi(f, \theta, \phi)$ and needs to be interpreted and processed to find the HRIRs/HRTFs. This section describes how the HRTFs can be extracted from the obtained pressure levels. Fig. 2.12 shows the convention used for angles θ and ϕ in 3D space. The angle θ is in the range $\pi \leq \theta \leq \pi$ and the angle ϕ is in the range of $-\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2}$. The obtained values for the pressures are normalized to obtain the raw HRTFs using free field Green function, given in Eq. 2.3.

$$G(f) = \frac{-i1.21f}{2\pi} e^{\frac{2\pi f}{c}}, \tag{2.3}$$

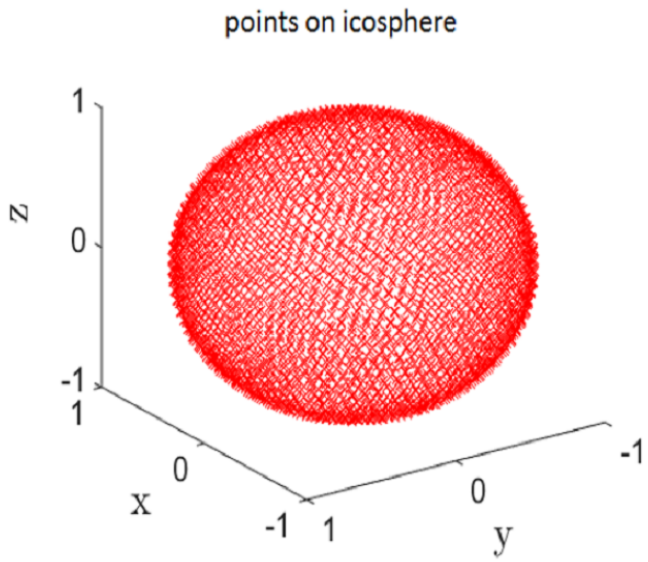


Figure 2.11: *This figure shows the spatial grid on which the positions for which the HRTFs are measured using FM-BEM simulations using reciprocity principal.(picture taken from [1]).*

here f denotes the sound wave frequency in Hz and c denotes the speed of the sound in the medium (air at room temperature) and is 343 m/s. The HRTF value for a given frequency, azimuth and elevation angle is then obtained from these pressure values as below:

$$HRTF(f, \theta, \phi) = \frac{\phi(f, \theta, \phi)}{G(f)}. \tag{2.4}$$

As the HRTFs are considered to be minimum phase filters, their phase information can be reconstructed from the Hilbert transform of the magnitude responses. However, as the phase information is not used in this thesis, it is not relevant to discuss the process of the phase calculation.

2.4 Virtual Auditory Space (VAS)

The spatial sound field reproduced through rendering the audio scenes over loudspeaker arrays [65]or headphones [33, 66, 67] is called virtual auditory space. Fig. 2.14 shows the loudspeaker arrangements of 5.1 and 7.1 surround sound systems for VAS. Where "X" in $X.1$ denotes the number of loudspeakers, and $.1$ denotes the one sub-woofer. The sub-woofer is usually placed outside of these arrangements for its inherent non-directive nature. The deriving signals for these loudspeakers are calculated by using complex wave field synthesis techniques. The focus of this study, however, is the reproduction of VAS over headphones, which is also called binaural VAS.

A simplified binaural VAS setup is presented in Fig. 2.15. Here S_n denotes the n^{th} sound source, and $h_{L/R}(\theta_n, \phi_n)$ denotes the left or right ear time-domain acoustic responses or HRIRs for the direction of sound source S_n, in a head-centered spherical coordinate system presented in Fig.2.5. It can be seen in Fig. 2.15 that the knowledge of HRIRs is crucial for binaural VAS reproduction. As it is not possible to measure

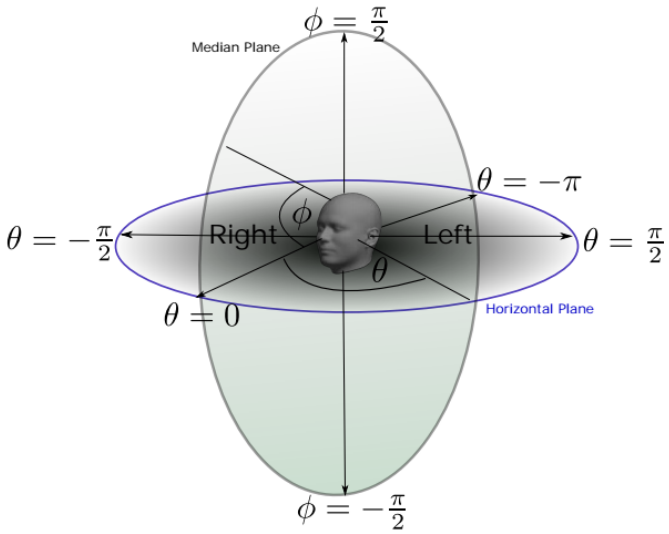


Figure 2.12: *Figure shows the conventional coordinate system used for for the angles θ and ϕ when running the HRTF simulations. It also shows the horizontal and vertical or median planes. (Image taken from [1]*

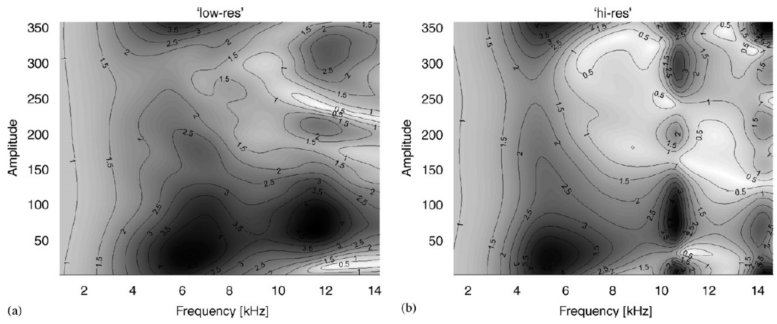


Figure 2.13: *The results of BEM simulations obtained by solving BIE for low resolution and high resolution mesh. The "low-res" and "low-res" meshes had and 12000 and 13488 triangular faces, respectively. Image taken from [1].*

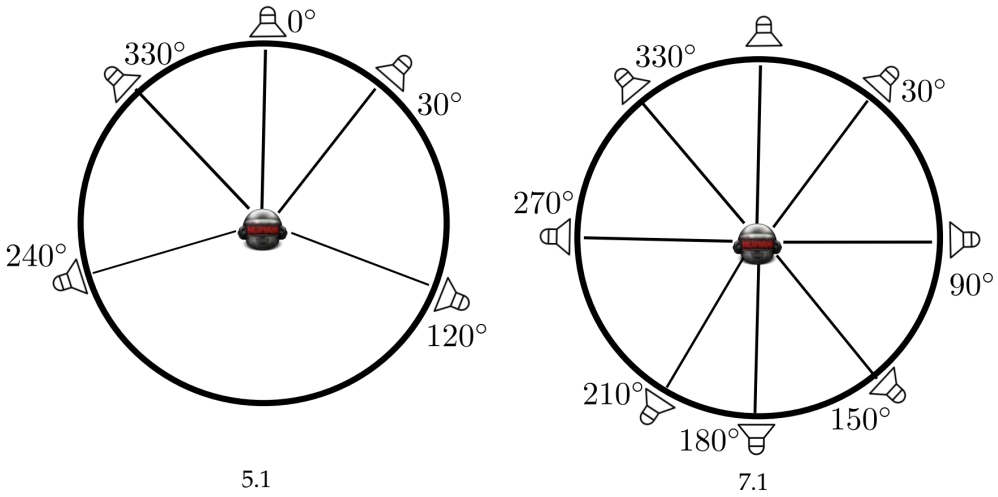


Figure 2.14: *A typical 5.1 and 7.1 surround sound speaker configuration. Image taken from [1].*

the HRIRs for every individual, the HRIRs measured for standard mannequins (dummies with average anthropometric properties for a given population), such as KEMAR, Bruel, and Kjaer or Samurai are used in such systems. However, the past psychoperceptual evaluations of these systems using average HRIRs suggest these systems generally result in poor localization and externalization experience, making the virtual spatial hearing experience unnatural and somewhat dry [18, 20]. This nullifies the whole purpose of VAS, which is to provide realistic and immersive, virtual spatial auditory experiences. The past studies suggest that in order to provide high quality and immersive experience, individualized HRTFs must be used. Despite this, there are not any comprehensive frameworks to provide individualized HRTFs/HRIRs for listeners for the mass market, which limits the use of binaural VAS to laboratory settings only. This demands a comprehensive and computation and inexpensive cast method for HRTF personalization. Some of the studies in the past have proposed some basic personalization methods for HRTF personalization. Details on a few of these methods are provided in Sec. 3.4.

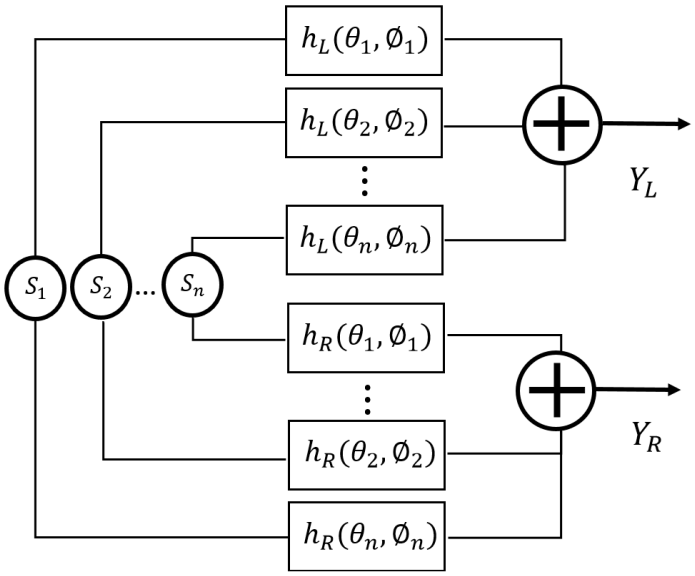


Figure 2.15: *Binaural VAS over headphones*

Having the HRIRs in hand, one can generate a very basic binaural VAS following the steps in Fig. 2.15. While reproducing the binaural VAS audio, we can emulate different cases, such as a) single source single direction, b) multiple sources with a single direction, c) multiple sources multiple directions. The basic mechanism is the same for all these setups. For the production of VAS over headphones involves emulating a sound signal, S_1, coming from a source located in a particular direction in space given by the auditory spherical coordinate angles, θ_1 and, ϕ_1. It is worth noting that we are only discussing the direction, not the position, which also requires the value of r. The reason for that is usually for a point source at a distance greater than 1 m is considered to be in the far-field making it easy to emulate the distance by simply changing the intensities of the sound signal (higher intensities for a closer source and lower intensities for a far located source). However, this is not the scope of this work; hence the details on this are not provided here.

To virtually put a sound source S_1 in a given direction (θ_1, ϕ_1), the sound signal is simply convolved with the left and right ear HRIRs for that direction providing the deriving signals for left and right channels as Eq. 2.5.

$$\begin{aligned} L_1 &= S_1 * h_L(\theta_1, \phi_1) \\ R_1 &= S_1 * h_R(\theta_1, \phi_1) \end{aligned} \tag{2.5}$$

Here symbol $*$ denotes $1D$ convolution operation. Now lets consider we have n sources, S_1, S_2, ... ,S_n, in n directions (θ_1, ϕ_1), (θ_2, ϕ_2), ... , (θ_n, ϕ_n). To virtually put these sources in these directions the steps in Eq. 2.5 has to be repeated to generate the driving signal for each sound source. At the end these individual signals for left and right ears are combined by using simple addition operation as shown in Eq. 2.6 and in the block diagram presented in Fig. 2.15.

$$\begin{aligned} L &= S_1 * h_L(\theta_1, \phi_1) + S_2 * h_L(\theta_2, \phi_2) + \cdots + S_n * h_L(\theta_n, \phi_n), \\ R &= S_1 * h_R(\theta_1, \phi_1) + S_2 * h_R(\theta_2, \phi_2) + \cdots + S_n * h_R(\theta_n, \phi_n). \end{aligned} \tag{2.6}$$

Another thing to be noted here is that the VAS system presented in Eq. 2.6, and Fig. 2.15, considers that the sound reproduction is happening in the "free-field" using the HRIRs recorded in the free field. Which means that, the other effects such as the reverberations are not accounted for. If someone requires to reproduce the spatial audio with a particular room environment over the headphones including the reverberations in the room etc. a more comprehensive impulse response called binaural-room impulse responses (BRIR) is to be used [67]. BRIRs are the transfer functions from the sound sources placed at any arbitrary point in a particular room to the ear drum. If the room and equipment in the room is anechoic BRIRs will be equalled to the HRIRs and the VAS produced using these BRIRs will be same as the "free-field VAS".

Furthermore, to make it more natural, we need to consider that when the human listeners move their heads while listening to a sound, the position of sound also changes depending on the movement of the head. So in order to make it real, one needs to track the head and adjust the direction for the used HRIRs accordingly. The reason is that when an individual moves his head, the relative direction of the sound source with respect to the ears also moves with the movement of the head [67,68]. Such systems are called dynamic VAS systems. This feature is not present in the VAS system presented in figure 2.15.

Another thing to be noted for is that the VAS produced over the loudspeaker arrays can be converted to the binaural VAS by simply convolving the loudspeaker driving signals with the HRTFs of the loudspeaker directions and combining the signals using equations 2.6 to produce the driving signals for left and right channels.

The following subsection provides a set of possible applications of VAS in various fields.

2.4.1 Possible Application of VAS

The listening experiences with the virtual auditory displays are different from the other acoustic displays as the listener perceives if the audio signal is generated by a real source around him [19]. Especially when the dynamic binaural VAS is used, a very natural immersion and localization experience can be provided. However, when VAS over loudspeaker array is used, the performance is good only in a small region called the sweet spot. The size of this sweet spot changes depending on the rendering technique and the number of used loudspeakers.

The applications of VAS spans on various areas ranging from science and academics to entertainment, music and gaming industry, and from medical and clinical studies to defense and military training, and even in simple social life (enabling a patient with hearing aid to have a natural conversation interaction) [69]. For example, in the research sector, there have been many studies and clinical experiments to study how the spatial hearing abilities of subjects can be restored for the listeners having hearing impairments [70]. VAS has also been used in localization experiments in order to obtain a better understanding of the mechanisms and underlying psycho-physical phenomenons of spatial hearing [71]. In clinical setup tests for spatial hearing have been suggested for evaluating everything from hearing aid functions to deficits in auditory brainstem [72–74]. In real social life, at the emotional health level, adding VAS capabilities to traditional hearing implants can assist many people who have hearing impairments to interact more effectively with their environment and other people when

having interactions and discussions.

From an entertainment perspective, broadcasting agencies such as BBC and Microsoft are investing a significant amount of resources into the production of VAS for listeners. Facebook has also built a whole devoted research lab for this. With the advent of graphics accelerators, games are becoming much more involved, through rendering 3D scenes not just on screens but also on head-mounted displays and special glasses. Although this engages and immerses the user more, without the availability of hi-fi VAS to complement it, the experience is still somewhat dry. Adding VAS headphones instead of standard stereo audio can make the experience with the virtual reality and augmented reality much more immersive and natural. The quality of movies and cinemas has also rapidly increased with the availability of 3D screens using special glasses, and 4K movies for home entertainment systems are becoming increasingly popular these days. VAS can complement these cinematic, and home entertainment technologies with personalized spatial audio experience, which can increase the quality of the immersion and naturalness of the multimedia experience by several folds. The same is also applicable to mobile phones and communication software on computers as well. VAS can add a great to the teleconferencing and online collaborative meeting tools by making them more engaging, sociable, and immersive [75].

VAS can also be of great aid in the applications where human operator analyzes and respond to the spatial information. For example, in aviation, an air controller must keep track of multiple planes under their control tower. Pilots must land, take off and fly their crafts, tracking the enemy crafts, hostile targets, and objects and buildings at the ground. Operators operating the remote rescue robots can be assisted with the help of reproducing the sound field in the remote target area on VAS to make the rescue more effective. All these applications make the production of binaural VAS a hot and important area.

2.5 Principal Component Analysis (PCA

Principal component analysis (PCA) is a statistical technique that employees an orthogonal transformation to convert a set of values on correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables, called principal components. The aim is usually to study the variations in the data. This section provides a brief overview of the standard principal component analysis (PCA) with the prospect of its application to analyze the variations of the HRTF magnitude responses for SYMARE users. In this section, using HRTF data, we show how the PCA is applied and what we can expect from it. The readers are advised to familiarize themselves with the notations and procedures developed in this section, as these will be referred and used in different parts of this thesis.

To begin with, let us suppose a single HRTF vector is denoted as H. H is a function of frequency f, azimuth angle θ, and elevation angle ϕ, hence $H(f, \theta_1, \phi_1)$ denotes the value for the frequency bin f of a single HRTF vector for direction (θ_1, ϕ_1). Each HRTF has K elements, which are the complex gains for each of the K frequency bins of the HRTF spectrum. The discrete K, frequencies are denoted as f_i in the HRTF spectrum. Considering that the HRTFs can be modeled as minimum phase filters and the constant delay and minimum phase can be retrieved using Hilbert transform from the magnitude response, for the sake of simplicity, we use only the magnitudes of the HRTF spectrum

by ignoring the phase. Let us consider we have a matrix X containing L HRTFs for L different directions in space given as:

$$\mathbf{X} = \begin{bmatrix} H(f_1, \theta_1, \phi_1) & H(f_2, \theta_1, \phi_1) & \cdots & H(f_K, \theta_1, \phi_1) \\ H(f_1, \theta_2, \phi_2) & H(f_2, \theta_2, \phi_2) & \cdots & H(f_K, \theta_1, \phi_1) \\ \vdots & \vdots & \cdots & \vdots \\ H(f_1, \theta_L, \phi_L) & H(f_2, \theta_L, \phi_L) & \cdots & H(f_K, \theta_L, \phi_L) \end{bmatrix} \tag{2.7}$$

To perform PCA on the dataset X following steps are performed.

Step 1

The first step is to compute the mean of the data frequency by frequency. As we have L data points the mean for k^{th} is given as:

$$\bar{\mathbf{H}}_k = \frac{1}{L} \sum_{i=1}^{L} H(f_k, \theta_i, \phi_i) \tag{2.8}$$

Step 2

The zero mean data $\hat{\mathbf{X}}$ is computed by subtracting the mean of the feature from each of the features for all HRTFs:

$$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{1}^T \bar{\mathbf{H}}, \tag{2.9}$$

where $\mathbf{1}^T$ denotes a column vector of ones of length L.

The singular value decomposition on this zero mean data matrix $\hat{X}$ can be performed as:

$$\hat{\mathbf{X}} = \mathbf{G}_{\hat{\mathbf{X}}} \mathbf{O}_{\hat{\mathbf{X}}} \mathbf{F}_{\hat{\mathbf{X}}}^T. \tag{2.10}$$

This will be used later in this section.

Step 3

The next step is to compute the covariance matrix $\mathbf{C}$ as follows:

$$\mathbf{C} = \hat{\mathbf{X}}^T \hat{\mathbf{X}}, \tag{2.11}$$

and each entry $\mathbf{C}_{k,p}$ can be computed as:

$$\mathbf{C}_{k,p} = \sum_{j=1}^{L} H(f_k, \theta_j, \phi_j) H(f_p, \theta_j, \phi_j) \tag{2.12}$$

Step 4

The singular value decomposition on this covariance matrix is then computed as:

$$\mathbf{C} = \mathbf{F}\mathbf{O}\mathbf{F}^T. \tag{2.13}$$

However, using the Eq.2.10, the singular value decomposition can also be written as:

$$\begin{aligned} \mathbf{C} &= \mathbf{F}_{\hat{\mathbf{X}}} \mathbf{O}_{\hat{\mathbf{X}}}^T \mathbf{G}_{\hat{\mathbf{X}}}^T \mathbf{G}_{\hat{\mathbf{X}}} \mathbf{O}_{\hat{\mathbf{X}}} \mathbf{F}_{\hat{\mathbf{X}}} \\ &= \mathbf{F}_{\hat{\mathbf{X}}} \mathbf{O}_{\hat{\mathbf{X}}}^T \mathbf{O} \mathbf{F}_{\hat{\mathbf{X}}}^T \end{aligned} \tag{2.14}$$

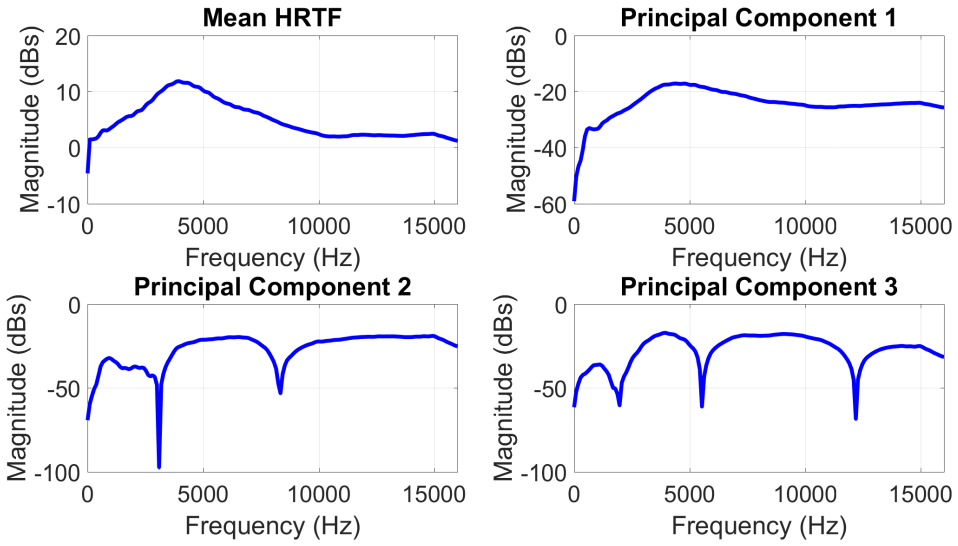


Figure 2.16: *The mean HRTF $\bar{H}$ and first three principal components f_1, f_2, and f_3*

From Eq. 2.14 and Eq. 2.13 it is deduced that $\mathbf{F}_{\hat{\mathbf{X}}} = \mathbf{F}$ and $\mathbf{O} = \mathbf{O}_{\hat{\mathbf{X}}}^T \mathbf{O}_{\hat{\mathbf{X}}}$. The dimension for the matrices $\mathbf{F}, \mathbf{O}, \in \mathbb{R}^{K \times K}$. The matrix $\mathbf{F}$ constitutes the new orthogonal bases for our data, where each column f_i in $\mathbf{F}$ is a principal component, which means each column is orthogonal to other columns:

$$\langle \mathbf{f}_i, \mathbf{f}_j \rangle = 0, \tag{2.15}$$

where $i \neq j$, and $\langle ., . \rangle$ denotes the inner product in $\mathcal{R}^3$ Furthermore, the matrix $\mathbf{O}$ is a diagonal matrix containing the eigen values from the SVD.

Step 5

All of the HRTFs can be represented as the weighted sum of these principal components f_i, where **weights** for these principal components for $H(f, \theta_x, \phi_x)$ can be computed as:

$$\mathbf{w} = (H(f, \theta_x, \phi_x) - \bar{\mathbf{H}})^T \mathbf{F} \tag{2.16}$$

the vector $w \in R^k$. Furthermore, the complete weights in the form of a matrix $\mathbf{W}$ for our data matrix $\hat{\mathbf{X}}$ are obtained as:

$$\mathbf{W} = \hat{\mathbf{X}}\mathbf{F} \tag{2.17}$$

To see the PCA in action, we now show an example of its application to real-life HRTFs. The data matrix $\mathbf{X}$ is generated by combining the HRTFs for left ears available in the SYMARE database. We have 61 subjects, and HRTFs are available for 393 directions, hence $X \in \mathbb{R}^{23973 \times 172}$, where 172 is the number of frequency bin corresponding to 16 kHz. Fig. 2.16 shows the mean ($\bar{H}$) and the first three principal components f_1, f_2 and f_3 for the example DTF data.

A full and accurate reconstruction for any of the DTFs in our data set can be obtained by linearly combining all of the principal components f_i with the appropriate

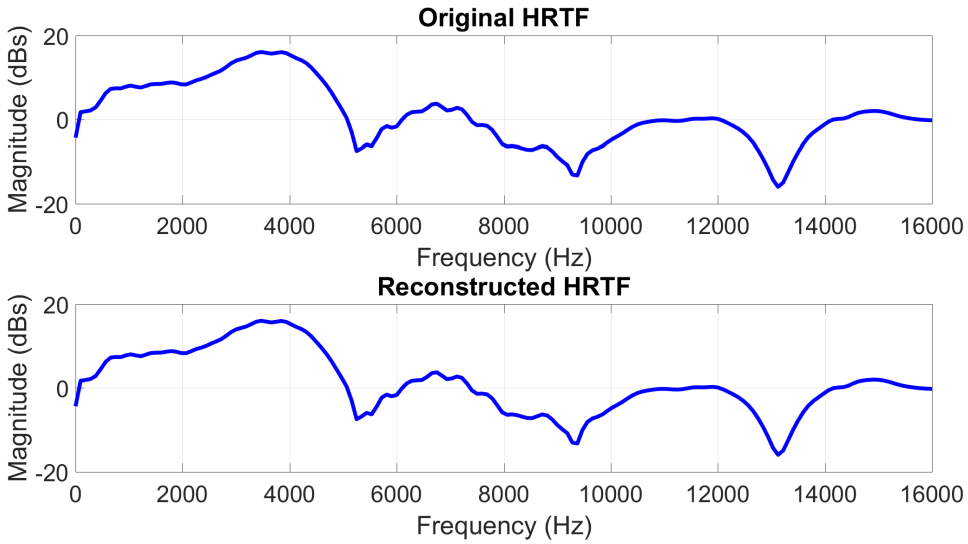


Figure 2.17: *The HRTF for a given direction and the same HRTF reconstructed using all the acoustical principal components*

w_i. Fig. 2.17 shows a specific HRTF and the reconstructed HRTF when all principal components are used.

When using PCA, we typically do not require the full set of L weights to represent the data with reasonable accuracy. In this sense, PCA provides a lossy compression of the data. Fig. 2.18 shows the reconstruction of the PCAs using a subset of the weights.

2.6 Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework

Large Deformation Diffeomorphic Metric Mapping (LDDMM) is another framework which was extensively used in this work to model the variations of the ear shapes across the database population. The section is borrowed from [1], another Ph.D. thesis which extensively used LDDMM to for morphoacoustics and provides readers with the necessary background and concepts of LDDMM in ear shape modeling viewpoint. The LDDMM framework was originally presented in [76, 77] and then further developed and used for surface matching by [78]. It includes knowledge of functional analysis, variational analysis, and reproducible kernel Hilbert spaces. In this study, we use LDDMM to match the triangular meshed surfaces, so in this section, a brief overview of LDDMM is provided in this context. Let us consider we have two surfaces given as $S_1(X)$ and $S_2(Y)$, containing the vertices and triangular connectivity information. LDDMM models the matching or morphing of $S_1(X)$ to $S_2(Y)$ as a dynamic flow of diffeomorphism of the ambient space, $\mathbb{R}^3$, in which the surfaces are embedded. This flow of diffeomorphism, $\phi^v(t,.)$, is defined via the partial differential equation:

$$\frac{\partial \phi^v(t,\mathbf{X})}{\partial t} = v(t) \circ \phi^v(t,\mathbf{X}) \tag{2.18}$$

where $v(t)$ is a time-dependent vector field with a vector defined for each point in space, for $t \in [0,1]$. This vector field models the infinitesimal efforts of the flow, and $\circ$ denotes

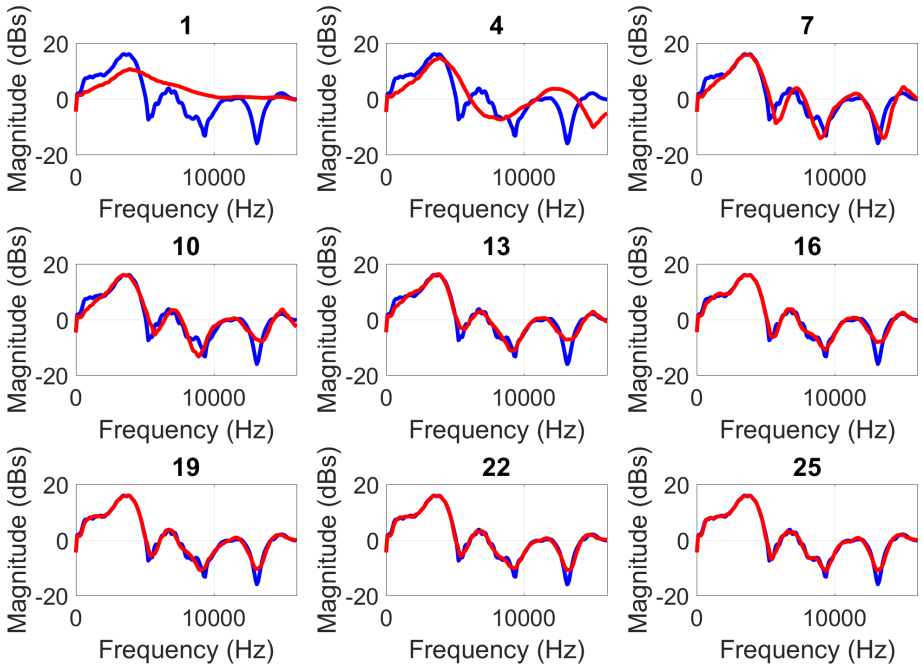


Figure 2.18: *The above plot shows how the the same HRTF (blue curve) is reconstructed using different number of weights (red curve). The number of principal components and weights used in the reconstruction of the HRTF is shown above the plot [1].*

function composition function.

At this stage, the vector field represented by $v(t)$ can be any vector field belonging to a Hilbert space of regular vector fields denoted by V which is equipped with a kernel, k_V, and a norm $\|.\|_V$ that models the infinitesimal cost of the flow.

The superscript v in Eq. 2.18 simply denotes that, this diffeomorphic flow is defined for a particular time-dependent vector field $v(t)$. The $v(t)$ can be determined by minimizing the cost function J_{S_1,S_2}:

$$J_{S_1,S_2}(v(t)) = \gamma \int_0^t \|\mathbf{v}(t)\|_V^2 dt + E(S_1(\phi^{\mathbf{v}}(t,\mathbf{X}), S_2(\mathbf{Y})). \tag{2.19}$$

As shown in Eq. 2.19, the cost function is composed of two terms. The first term $\|\mathbf{v(t)}\|_V^2 dt$, is called the energy term and is the measure of the energy required to transform the shapes $S_1(\mathbf{X})$ to match to the target shape $S_2(\mathbf{Y})$. The solution to term $v(t)$ can be expressed as the convolution of the momentum vectors, $\alpha_n(t)$ with the kernel k_V, with one momentum vector defined for each of the N vertices in $\mathbf{X}$ [79], as:

$$\mathbf{v}(t) = \frac{d\mathbf{x}(t)}{dt} = \sum_{n=1}^{N} k_{\mathbf{v}}(\mathbf{x}_n(t), x_n)\alpha_n(t), \tag{2.20}$$

where term k_V is called the deformation kernel, and is given by:

$$k_V(x,y) = \frac{1}{1 + \frac{||x-y||^2}{\sigma_V^2}}. \tag{2.21}$$

σ_V is called the deformation scalar and controls the smoothness of the deformation.

While the second term in the cost function, $E(S_1(\phi^{\mathbf{v}}(t,\mathbf{X})), S_2(\mathbf{Y}))$ is called the the shape comparison term and quantifies the differences between the matched shape $S_1(\phi^{\mathbf{v}}(t,\mathbf{X}))$ and the target shape $S_2(\mathbf{Y})$. This term is computed as:

$$\begin{aligned} E(S_t^1, S_2) = &\sum_{f,g} \langle n_{S_1^t}(f), k_W(c_{S_1^t(g)}, c_{S_1^t(f)}) n_{S_1^t(g)} \rangle \\ &+ \sum_{p,q} \langle n_{S_2}(p), k_W(c_{S_2(q)}, c_{S_2(p)}) n_{S_2(q)} \rangle \\ &-2 \sum_{f,q} \langle n_{S_1^t}(f), k_W(c_{S_2(q)}, c_{S_1^t(f)}) n_{S_2(q)} \rangle, \end{aligned} \tag{2.22}$$

where $n_{S_1^t}(f)$, $n_{S_2}(p)$ and $c_{S_1^t}(f)$, $c_{S_2}(p)$ denote the normal vectors and centers for faces f and p for shape S_1^t and S_2 respectively. The lengths of the normal vectors for every face are equal to the area of the face. The terms k_W is called the shape matching kernel and controls how the matching of a given vertex effects the neighbouring vertices through variable σ_W. k_W is given by:

$$k_W(x,y) = \frac{1}{1 + \frac{||x-y||^2}{\sigma_W^2}}. \tag{2.23}$$

Hence minimizing the cost functions in Eq. 2.19 ensures that the shape error is minimum, which means the shape best matches the target shape, as well as the energy required to morph the given shape to the target shape, is also minimized. The parameter γ in the cost function is a scaling parameter for the energy term, higher the value for γ, more the energy term is penalized when doing the cost minimization. The energy term also ensures that the deformations are diffeomorphic; hence this is called the regularization term.

One can notice that the solution presented in Eq. (2.20) is a continuous-time differential equation; however, when solving this differential equation numerically, the time axis is discretized. The differential equations are then solved using the Euler scheme, i.e., the flow operation is performed on the discrete meshes with discrete time steps. Using the Euler method, this diffeomorphic flow is characterized by a sequence of deformations, which are uniformly ordered in time with a single deformation occurring at every time step.

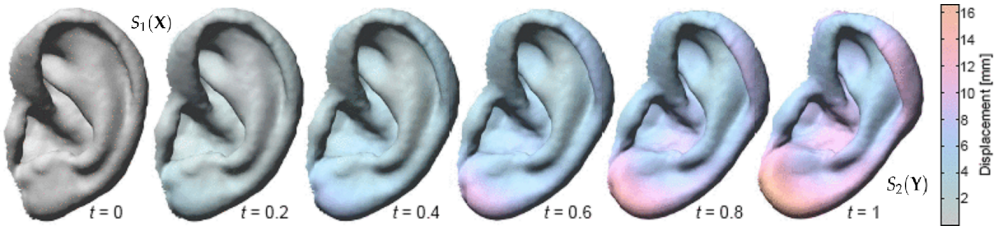


Figure 2.19: *The results of the flow of diffeomorphisms for several time steps are shown for the matching of S_1 to S_2. The colour indicates the displacement. A a constant luminance color map is used for clarity. (picture must be seen in color) Picture taken from [5].*

Fig. 2.19 shows the evolution of ear shape $S_1(\mathbf{X})$ to ear shape $S_2(\mathbf{Y})$ using the LDDMM flow operation with a vector field that has been computed specifically for map-

ping on ear to other. Further, Fig. 2.20 shows an example of a single point in space and its displacement over seven contiguous time steps under LDDMM. The displacement vectors, $v(t)$, are obtained by applying the convolution given in Eq. 2.20.

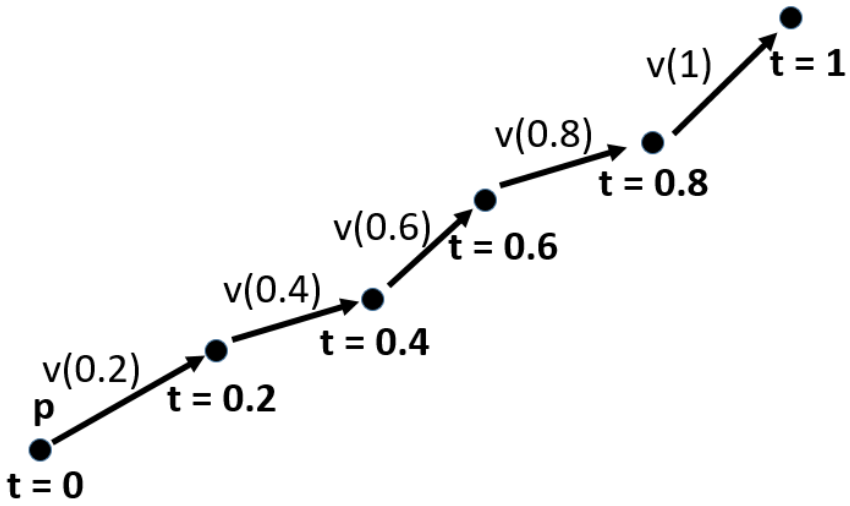


Figure 2.20: *The flow of a point or particle p is shown in space. The velocity vectors v(t) signify the direction and magnitude of the displacement of the particle at each time step starting from time t=0 and ending at time t=1 [1].*

[80] suggests that instead of using normal Euler scheme, the use of centered Euler scheme provides a more accurate solution when used to solve Ordinary Differential Equations (ODEs) in Eq. (2.20). Furthermore, [5] suggests that while mapping one ear shape to another target ear the whole matching process can be modeled using 11 steps.

Once the matchings are done from the template to the target shape, the target shape can be computed using the geodesic shooting from the initial momentum vectors. To provide an analogy for understanding the geodesic shooting between two shapes, one can consider the path two points on the sphere surface. The surface of the sphere is a non-linear Riemannian space. Fig. 2.21 shows two points on a sphere denoted by green stars. The shortest path, known as the geodesic path, connecting these two points is colored in black and lies along the great circle containing the two points. Other paths on the sphere can also be constructed that join these two points, but these paths are longer.

2.6.1 LDDMM Induced Distances in Shape

This section provides the details on the cost function used in the LDDMM framework. As described before the first term in the cost function described in Eq. (2.19) is called the energy term and provides the measure of the energy required to deform a given shape $S_1(\mathbf{X})$ in to a target shape $S_2(\mathbf{Y})$, and is given by:

$$d_{Def}(S_1, S_2) = \int_0^1 \|\mathbf{v}(t)\|_V^2 dt. \tag{2.24}$$

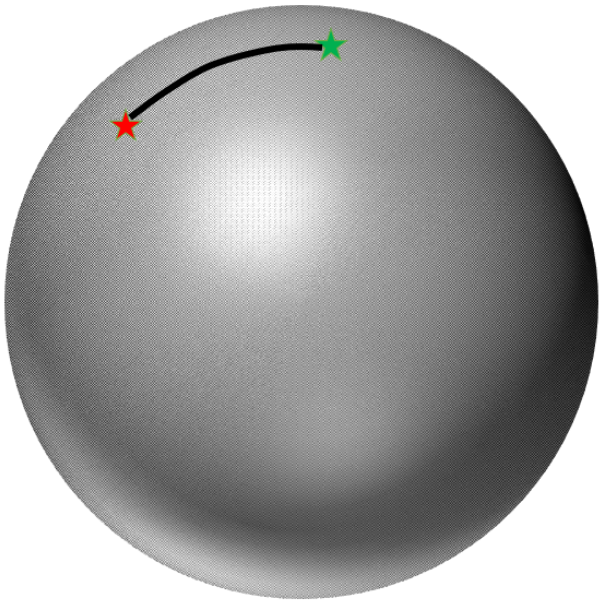


Figure 2.21: *The above figure illustrates the concept of the geodesic path using the surface of a sphere. The surface of a sphere is non-linear Riemannian space the two points on a sphere are shown using green and red stars, the optimal geodesic path between the points is shown as a black curve [1].*

When the optimal momentum vectors are obtained by minimizing the cost function, the energy term shown in Eq. (2.25) will become a measure of the geodesic distance between the source and the target shapes. In order to obtain an optimal solution or set of momentum vectors and to minimize the LDDMM cost function, an optimization algorithm is run. If we have unlimited time, there will be a possibility to find the most accurate momentum vectors resulting in a zero or minimal value for the second term. Still, in reality, due to time constraints, the optimization is conducted for only a finite number of iterations, which yields a set of momentum vectors that are close to optimal ones. The term $\|v(t)\|_V^2$ is a normed squared value that can be expanded using the discrete flow equations Eq. (2.20):

$$\|\mathbf{v}(t)\|_V^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\langle\alpha_i(0), k_V(\mathbf{x}_i(t), \mathbf{x}_j(t))\alpha_j(t)\rangle. \tag{2.25}$$

There is a wide range of kernels k_V that can be used for the purposes of LDDMM mappings. The kernels that are selected in LDDMM mappings have to adhere to a set of mathematical conditions. Details on the selection of the deformation kernel can be seen in [81]. Radially decaying kernels such as the Cauchy or gauss kernels are widely used for LDDMM surface mappings and are also used for mapping shapes in this study. The Cauchy kernel is defined as:

$$k_V(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma_V^2}} \tag{2.26}$$

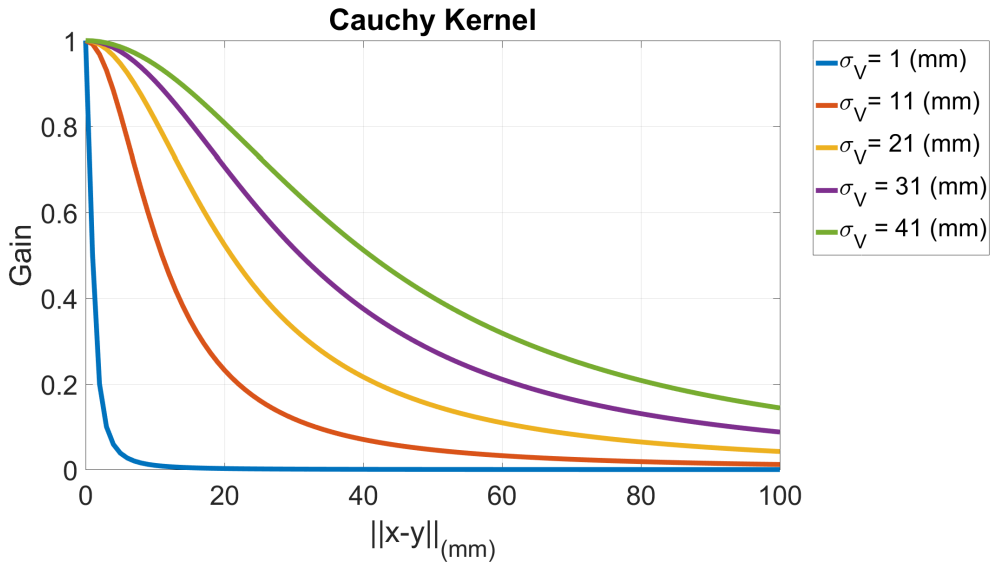


Figure 2.22: *The gain of the Cauchy kernel verses the distance between x and y for different values of σ_v [1].*

where the σ_V parameter is the deformation scale parameter. This parameters determines the range of the influence of a given momentum vector through kernel function k_V. The deformation scale parameter σ_V is an important parameter and plays a vital role when mapping shapes using LDDMM. The gain for Cauchy kernel as a function of the distance between x and y for different values of σ_v is show in Fig. 2.22.

The Role of σ_V Parameter

The σ_V parameter defines the coupling between the vertices of the source shape as they move along the deformation path from the source shape to the target shape. The value of σ_V greatly impacts the quality and characteristic of the mapping between the two shapes. Consider the discrete mesh-flow equation shown in Eq. (2.20) and a point on the source mesh denoted by $x(t)$. As the σ_V parameter is made larger, the momentum vectors surrounding this point will have a greater impact on the movement at $x(t)$. Consequently, the vertices will move more coherently along the deformation path. The σ_V parameter greatly influences the energy for deforming a shape in the LDDMM framework.

2.6.2 Measuring shape differences using currents

The second term of the cost function, $E(S_1(\phi^v(1,\mathbf{X})), S_2(\mathbf{Y}))$, provides the difference in the surface geometry of the matched surface $S_1(\phi(1,\mathbf{X}))$ and the target surface $S_2(\mathbf{Y})$ and is calculated based on the theory of currents [78]. Current can be used to represent the surfaces and are linear functionals on the space of differentials. The intuition behind using currents to represent surfaces is that they can be integrated over a surface to give a real value. When two surfaces are similar the difference in the value of the integrals of the surfaces is close to zero. Let $[S]$ to be denoting the current representing the surface S. $E(S_1, S_2)$ is defined as $E(S_1, S_2) = \|[S_1] - [S_2]\|^2_{W^\star}$. $E(S_1, S_2)$ is the squared norm

of the difference of the two surfaces for a dual norm (i.e. $\|, \|_W^\star\|$) in a Hilbert space W of differential forms [78].

In the discrete settings, a surface, S, is approximated by a triangular mesh in $\mathcal{R}^3$. Given a face f of S, let $c_S(f)$ denote the centre of the face and $n_S(f)$ denote the normal vector to the face with a length equal to the area of the face. We can then express $E(S_1, S_2)$ using the mesh elements as:

$$\begin{aligned} E(S_1, S_2) = & \sum_{f,g} \langle n_{S_1}(f), k_W(c_{S_1}(g), c_{S_1}(f)), n_{S_1}(g)\rangle \\ & - 2\sum_{f,q} \langle n_{S_2}(f), k_W(c_{S_2}(q), c_{S_1}(f)), n_{S_1}(q)\rangle \\ & + \sum_{p,q} \langle n_{S_1}(p), k_W(c_{S_1}(p), c_{S_1}(q)), n_{S_1}(q)\rangle \end{aligned} \tag{2.27}$$

where in the above $\langle, \rangle$ represents a vector dot product and the kernel k_W is typically chosen as the Gauss or Cauchy kernel. In this work, we use the Cauchy kernel for measuring shape mismatches:

$$k_W(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma_W^2}} \tag{2.28}$$

The shape comparison scale parameter σ_W determines the physical scale at which the shapes are compared. Larger values for σ_W result in a comparison of shapes at a coarse level of detail and small values of σ_W result in a comparison of shapes at a fine detail. The effect of the σ_W value on the quality and characteristic of the matching is explained and shown in Sec. 4.4.2. Please note that because the shape difference measure E is a metric it can also be defined using the notion of the inner or scalar products:

$$\begin{aligned} E(S_1, S_2) &= \|[S_1] - [S_2]\|_{W^\star}^2 \\ &= \langle [S_1] - [S_2], [S1], -[S_2]\rangle_{W^\star} \\ &= \|[S_1]\|_{W^*}^2 + \|[S_2]\|_{W^*}^2 - 2\langle [S_1], [S_2]\rangle_{W^*} \end{aligned} \tag{2.29}$$

2.6.3 Geodesic Shooting

Once the initial momentum vectors, $\alpha_n(0)$ are computed by minimizing the cost function $J(S_1, S_2)$ the diffeomorphic mapping between S_1 to S_2 can entirely be determined [82]. This follows from the fundamental principle in the LDDMM framework known as the conservation of momentums, which was proved in a seminal work by [83]. In other words, S_2 can be generated (modelled) as a deformation of S_1 through the diffeomorphic flow defined by the only the initial momentum vectors $\{\alpha_n(0)\}_{1\leq n\leq N}$.

Geodesic shooting consists in using a set of initial momentum vectors, $\{\alpha_n(0)\}_{1\leq n\leq N}$, to morph a shape S_1 into another shape, S_3. The shooting is found by solving the shooting equations, which couple the momentum vectors to the vertex positions across time

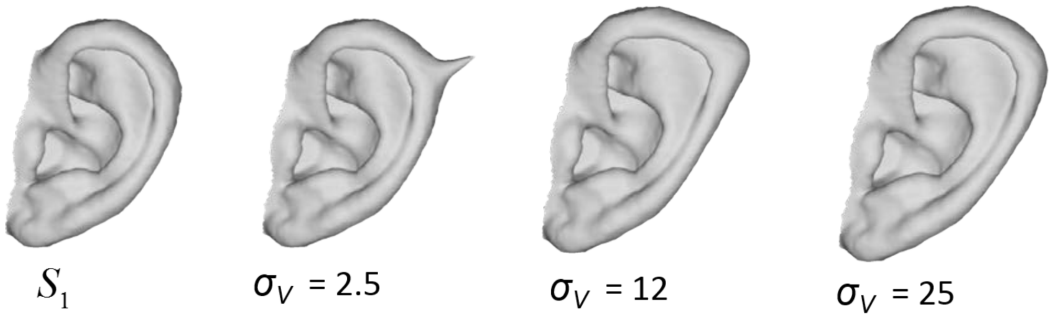


Figure 2.23: *The above figure shows the importance of σ_v. The plots show three ear shapes generated using geodesic shooting with a single non-zero initial momentum vector and varying σ_V from 2.5 to 25. The deformations are very local when a small value for σ_V is used resulting in abnormal ear shapes having sharp features, while when the large value of σ_V is used a single momentum vector does not change the shape too much and the results are very natural looking. Image taken from [1].*

and are given by:

$$\begin{aligned}\frac{d\alpha_r(t)}{dt} &= -\sum_{n=1}^{N}\langle\alpha_r(t),\alpha_n(t)\rangle \bigtriangledown_{x_r(t)}\left(k_V(x_r(t),x_n(t))\right)\\ \frac{dx_r(t)}{dt} &= \sum_{n=1}^{N}k_V(x_n(t),x_r(t))\alpha_n(t)\end{aligned} \tag{2.30}$$

where $\bigtriangledown_x(t)(.)$ denotes the gradient operator and $1 \leq r \leq N$. Note that the initial conditions for Eq. 2.34 are given by the initial positions of the vertices and the corresponding momentum vectors. Following the findings of this section Fig. 2.23 shows three ear shapes which are generated by applying geodesic shooting through a single non-zero momentum vector for different σ_V values varying from 2.5 to 25. When generating the three ear shapes, the same initial momentum vector located on the source shape S_1 was used.

2.7 Kernal Principal Component Analysis

This section provides an introduction to Kernel-based Principal Component Analysis (KPCA) and its use in a particular setting to statistically analyze the deformations of the multi-scale template ear $\bar{E}$ to any other ear in SYMARE database. The template ear was calculated using a multi-scale template calculation method proposed in [5]. The difference between KPCA and PCA is the calculation of the covariance matrix. Instead of using normal inner product KPCA uses inner product through the same kernel used by LDDMM operations. This process ensures the calculation of the correlation considers not just the momentum vector for a given vertex but also for the neighboring vertices through σ_V.

The first step is to calculate the momentum vectors $\alpha_n^{(l)}(t)$, where $t = [0,1]$, and $n = 1,2,\cdots,N$ denotes the id for the vertex, for every ear l in the SYMARE database which captures the deformation of the $\bar{E}$ to the ear shape S_l. This is done by solving the minimization problem for cost function $J(\bar{E},S_l)$ provided in Eq. 2.19. As high-

lighted in the previous section, once the momentum vectors for deformations are found the whole flow of deformation can be defined using geodesic shooting on just initial momentum vectors $\alpha_n^{(l)}(0)$. For simplification reasons $\alpha_n^{(l)}(0)$ is going to be denoted as α_l from now one, where l denotes the subject id. All the initial momentum vectors can be arranged in a matrix shape as:

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_L \end{bmatrix}_{3N \times L}, \tag{2.31}$$

where N is the number of vertices in $\bar{E}$ and L is the number of ear shapes in SYMARE. Note that the dimensions of the matrix are $3N \times L$ because we are taking all the coefficients of the initial momentum vectors for x, y and z axis. The next step is to find the centered initial momentum vector matrix $\hat{\mathbf{A}}$ by subtracting mean initial momentum vectors from each column. The mean initial momentum vectors are computed as:

$$\bar{\alpha} = \frac{1}{L} \sum_{l=1}^{L} \alpha_l. \tag{2.32}$$

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 & \cdots & \hat{\alpha}_L \end{bmatrix}_{3N \times L}, \quad where, \hat{\alpha}_l = \alpha_l - \bar{\alpha} \tag{2.33}$$

In the next step we form a kernel function $\mathbf{K}$, which contains the kernel values for the kernel function provided in equation 2.26 for every pair of the vertices for $n = 1, 2, \cdots, N$ for both shapes, given by:

$$\begin{aligned} \mathbf{K} &= \begin{bmatrix} K_{11} & K_{21} & \cdots & K_{1N} \\ K_{21} & K_{22} & \ddots & K_{2N} \\ \vdots & & & \vdots \\ K_{N1} & K_{N1} & \cdots & K_{NN} \end{bmatrix}, \\ K_{mn} &= k_V(x_m, x_n) I_{3\times 3}, \end{aligned} \tag{2.34}$$

where $I_{3\times3}$ denotes a 3×3 identity matrix. The next step is to calculate the covariance matrix $\mathbf{C}$, calculated as:

$$\begin{aligned} \mathbf{C} &= \frac{1}{1-L} \hat{\mathbf{A}}^T \mathbf{K} \hat{\mathbf{A}} \\ c_{i,j} &= \frac{1}{1-L} \langle \{\alpha_n^{(i)}(0)\}, \{\alpha_n^{(j)}(0)\} \rangle_V = \frac{1}{1-L} \hat{\boldsymbol{\alpha}}_i^T K \hat{\boldsymbol{\alpha}}_j. \end{aligned} \tag{2.35}$$

Having the covariance matrix in hand the next step is to perform the Singular Value Decomposition(SVD) on it as was performed in the case of normal PCA.

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T. \tag{2.36}$$

The matrix of the principal components $\mathbf{U}$ is then calculated as:

$$\mathbf{U} = \hat{\mathbf{A}}\mathbf{V}\mathbf{D}^{-\frac{1}{2}}. \tag{2.37}$$

The matrix $\mathbf{U}$ contains the principal components for the covariance matrix with each column representing the corresponding principal component. It is to be noted that the

principal component matrix is orthogonal in the Hilbert space of deformations, i.e., $\mathbf{U}^T\mathbf{K}\mathbf{U} = \mathbf{I}$. The matrix for the intimal momentum vectors can be reconstructed as:

$$\hat{\mathbf{A}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T + \hat{\alpha}, \tag{2.38}$$

where $\mathbf{D}^{\frac{1}{2}}\mathbf{V}^T$ provides the weights or coefficients for the principal components. Having this in hand we can perform the KPCA on the morphology domain and reconstruct the ears using any m principal components, where $1 \leq m \leq L$.

2.8 Affine Matching two Shapes

This section provides detail on how one shape can be rigid transformed to match the other shape, in scale, rotation and position. The applied transformation includes scaling, rotation, and translation. In this process the scaling is performed uniformly across all axis (rigid scaling), i.e. a single scale is used for all the axes. The same matching process was used in [1] to create the morphable model of the ear shapes details on creating the morphable model are provided in Sec. 3.2.2. This study uses the matching performed using the methods described in the study [84]. This study provides a set of methods to match the 3D distributions and measures. A brief detail of the affine matching process are provided following in this section. The reader interested to read the full method is invited to read the paper [84].

The process of rigid matching in [84], works by first defining the measure for the template and given target ear shapes. Let us denote the triangulated mesh for template ear shape as T, with the vertices denoted as x_i, and the target ear shape for a given subject as S_l, with vertices denoted by y_j^l. Here, the superscript l indicates the subject number, and the subscripts i and j are the indices for the vertices of T and S_l, respectively.

Subsequently, the measures μ and v^l for template and subject ear shapes are defined as:

$$\begin{aligned} \mu &= \frac{1}{m}\sum_{i=1}^{m}\delta_{x_i} \\ v^l &= \frac{1}{n_l}\sum_{j=1}^{n}\delta_{y_j^l} \end{aligned} \tag{2.39}$$

To perform the affine matching between these two measures, two dual spaces I and $I\star$ with Hilbert norm $||.||_I$ and $||.||_I^\star$ are defined. The detail of these two spaces and norms can be found in [84]. The affine matching between the two spaces is then obtained by minimizing the cost function $J(M, b)$. In the cost function given in Eq. 2.40, M_l provides the rotation and scaling matrix from template to the ear shape while b_l provides the translation vector which moves the template ear shape to the same position as the subject ear shape.

$$J(M_l, b_l) = ||\psi_{M,b}(v^l) - \mu||_{l^\star}^2 \tag{2.40}$$

This cost function minimizes the norm of the difference between the affine transformed measure v^l and the measure μ. $\psi_{M,b}(v^l)$ denotes the affine transformations for the measure μ and is performed as:

$$\psi_{M,b}(v^l) = My_j^l + b, \quad y_j^l \in R^3. \tag{2.41}$$

One thing to be noted here is that before performing these affine matchings, the center of the mass for every ear shape, including the template ear shape, needs to be moved to the origin. For this the center of mass for all shapes is calculated and a vector from center of the mass to the origin is added in the vertices to move the ear shapes to the origin. These center of masses are denoted as t_l (for shape l) and are computed as:

$$t_l = -\frac{1}{V(l)} \sum_{j=1}^{V(l)} y_j^l \tag{2.42}$$

where, y_j^l denotes the j^{th} vertex of the l^{th} shape, and $V(l)$ denotes the number of vertices in shape l. Once the shapes are all centered at the origin the process of affine matching was performed to minimization of the cost function $J(M_l, b_l)$ for every shape l. In Fig. 2.24 shows a set of ear shapes before and after the are affine matched with the template ear shape. The subscript $_{RTS}$ denotes the order of the transformations applied, scaling, translation and then rotation to perform the matching.

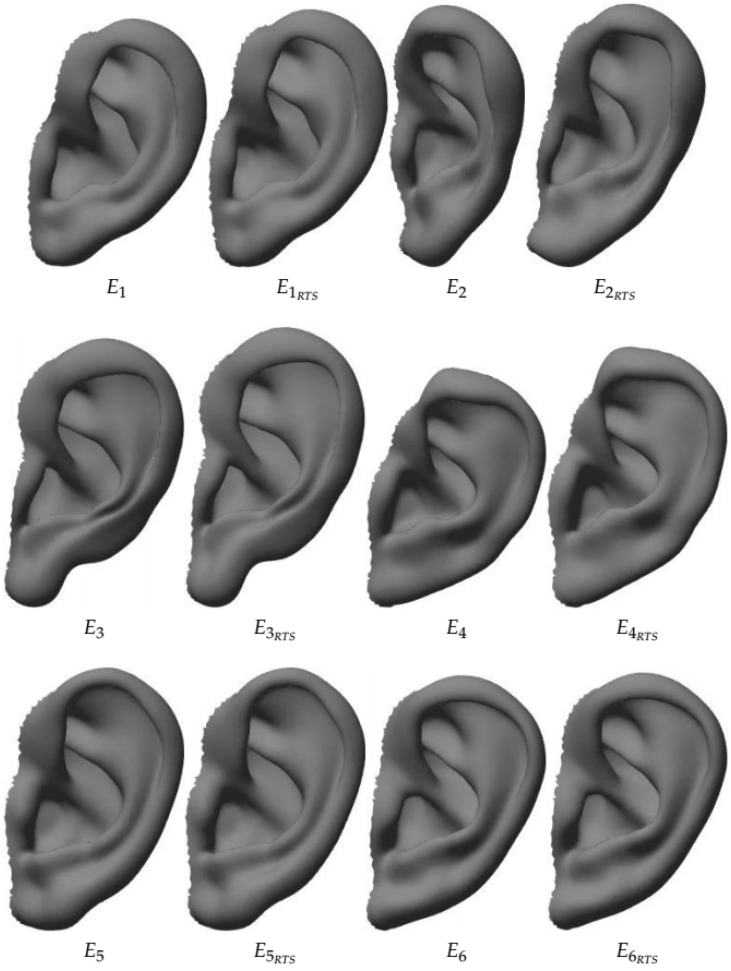


Figure 2.24: *Original and affine matched ear shapes for the first six subjects in SYMARE. The subscript RTS signifies the fact that these ears have been scaled, translated, and rotated to match the template ear shape. (Image taken from [1])*

2.8.1 Extracting Scale and Rotation Information from Affine Matching Matrix

Once the two shapes are rigid matched, the relative scaling, rotation and translation information can be extracted from the matching matrix M_l and translation vector b_l. In the above section it is show how any ear can be matched to the template ear. This section provides details on the process of extracting the relative scale, rotation and translation information from the transformation matrix and translation vectors obtained in this process. One thing to be noted is that to keep the shapes the same, the scale on all three coordinates of the point clouds is the same (rigid matching). Hence following the basic property of the affine transformations with a single scale on all coordinates, we can extract the scale using one of the following expressions:

$$\begin{aligned} SF_l =& \frac{1}{K}, \\ \text{where } K \text{ comes from } & K^2 I = M_l M_l^T, \text{or} \\ SF_l =& (det(M_l)^{\frac{1}{3}})^{-1} \end{aligned} \tag{2.43}$$

Once the scale factor is computed the rotation matrix R_l is obtained by dividing the M_l matrix with scale factor:

$$R_l = (\frac{M_l}{SF_l})^{-1}. \tag{2.44}$$

The **Tait-Bryan angles** can then be calculated from the R_l matrix as follows:

$$\begin{aligned} \theta_x =& \arctan(\frac{R_l(3,2)}{R_l(3,3)}) \\ \theta_y =& \arctan(\frac{-R_l(3,1)}{\sqrt{R_l(3,2)^2 + R_l(3,3)^2}}) \\ \theta_z =& \arctan(\frac{R_l(2,1)}{R_l(1,1)}) \end{aligned} \tag{2.45}$$

This expression is valid only when the angles are calculated according to the z-axis, then the y-axis, and finally the x-axis, i.e., the order is created in a way that, $R_l = R_z \times R_y \times R_x$. Note that these angles can be computed in a different order as well, in which the values would be different.

CHAPTER 3

Literature Review

This chapter provides a brief overview of the literature related to the morphoacoustic study of the outer ear shapes. The ultimate goal for the morphoacoustic study is to obtain a comprehensive framework for rapid mapping between the ear morphology and the corresponding individualized acoustic transfer functions.

The morphoacoustic approach considers modeling the variations in both domain the morphology and the acoustics and then create a mapping between both of these models in a way that we have an end to end personalization method. In this chapter, we start by providing an overview of the literature related to the shape parameterizations using two techniques spherical harmonics and the elliptical Fourier transform in Sec. 3.1.2 & Sec. 3.1.3 respectively. Both these techniques can be used for both the acoustic analysis and the morphology of the head and ear shapes. Furthermore, we also provide a review of the literature on modeling the ear shapes using simple geometric objects such as cylinders and parabolic surfaces in Sec. 3.1.1. The motivation behind this is that despite the simplicity of such models, they still explain some important features seen in the spectrum of acoustic responses of ear shapes, namely the center frequencies of the notches and peaks in PRTFs.

Having reviewed these traditional approaches, we provide the literature review on the use of the LDDMM framework on the ear shapes, along with the methods to calculate the template ears using both single and multi-scale methods in Sec. 3.2. We also provide the readers with the details on the derivations of a parametric model for ear shapes in Sec. 3.2.2.

The other study that is described in this chapter is known as morphoacoustic perturbation analysis (Sec. 3.3). In this method, a small portion of the reference ear shape is perturbed, and the acoustic response is computed using numerical simulations on this new modified ear. The changes in the acoustics of the perturbed shape are compared to the acoustic response of the reference shape. These kinds of studies help us to identify

the regions in the ear shape, which are more sensitive to these perturbations and play an essential role in the generation of notches and peaks in the acoustic spectrum [22].

In Sec. 3.4 we provide an overview of the literature on the existing personalization methods for HRTFs. At the end of this chapter in Sec. 3.5 we provide well known methods to evaluate the performance of ear shape and HRTF modeling. This section provides the overview on the error metrics used for matching the ear shapes and HRTFs in Sec. 3.5 & Sec. 3.6 respectively.

3.1 Morphology modeling

This section provides a brief overview of the three of the most popular techniques used by the researchers to model the head and ear shapes of the human listeners for HRTF personalization. These include the spherical harmonics for modeling the head shape of the listener, the use of elliptical Fourier transform to model the head and ear of the listeners, and the use of simple geometric shapes to approximate the human morphology and to model the of HRTFs. Following, we provide details on each of these individually.

3.1.1 Modeling the Acoustics of The Human Morphology Using Simple Geometrical Objects

One of the pioneer studies to understand the underlying phenomenons generating spatial hearing cues was performed by Lord Rayleigh more than a century ago. In the early model, he proposed to approximate the acoustic response of the head and ear of the listener using a sphere [85]. This estimation is very ambitious and is based on a simple model. Still, it provides important insights into the generation of binaural cues. Furthermore, it enables one to personalize these approximated cues by changing the diameter of the sphere.

This model was further studied and improved by authors in [6], where they modeled not just the head but also the torso of the listener and proposed a snowman model to approximate the listeners morphoacoustics. The image of the model proposed in this study is provided in Fig. 3.1. The size of the spheres can be chosen as one of the two most used methods, i.e., IAS (Inter-aural Sphere) sphere or EqVol (equal volume) spheres. IAS sphere has the diameter equal to the distance between both left and right pinna, while in the EqVol sphere, the volume of the sphere is chosen to be equal to the volume of the actual head shape of the listener. The same case follows for the torso. The model proposed in [6] showed that the torso plays an important role in the generation of elevation cues. Furthermore, the findings in this work report that the spherical head and spherical torso models provide a good approximation to the HRTF's of a pinna less KEMAR mannequin. The results are presented in the Fig. 3.2. Following the same lines, authors in [7], modeled the concha shape of the pinna. They considered the concha and region around the ear canal is the most relevant region. This study first proposed to model the concha shape using an inclined cylindrical shape with radius R and depth L, which are equal to the radius and depth of the concha. Later they improved this model by adding a metallic rectangular flang to the cylindrical concha. This enhancement improved the directivity of the acoustic spectrum compared to just a cylindrical concha. Fig. 3.3 shows the increased response of a cylindrical concha

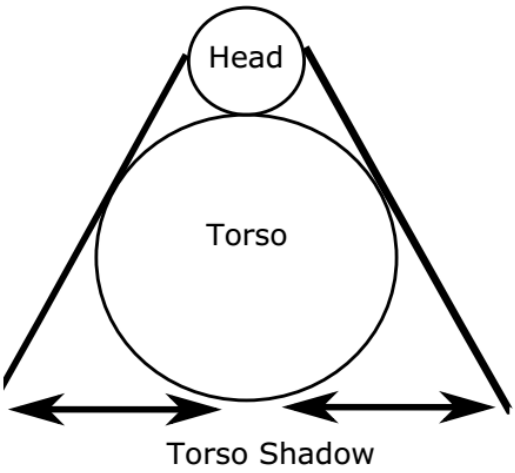


Figure 3.1: *The figure shows the snowman model consisting of two spheres. Image taken from [6]. The top (smaller) sphere is used to approximate the head of the listener while the bigger (bottom) sphere is used to approximate the torso.*

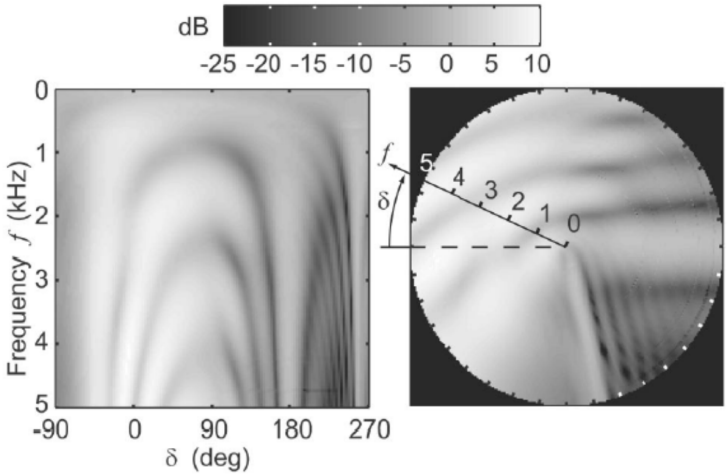


Figure 3.2: *This figure shows the acoustic response for frontal elevation angles δ and frequency ranging from 0-5 kHz. The results show that the notch center frequencies are symmetric about the elevation angle $\delta = 90^\circ$ and are due to the reflections from the torso region. Image taken from [6].*

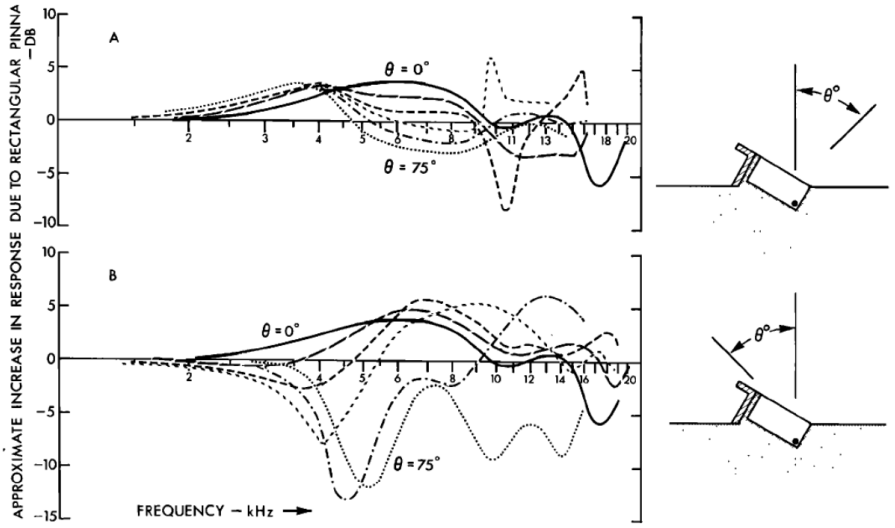


Figure 3.3: *Increase in response when a rectangular flang is added to cylindrical Concha for various angles of incidence θ . Plot (A) is for sources originating in front and plot (B) is for sources originating at the back. Image taken from [7]*

with the added rectangular flang compared to just a cylindrical concha. Fig. 3.4 shows a comparison of modes and resonances between the average responses of real ears to that of replicated ear shapes using various geometrical objects, including a cylindrical concha and cylindrical concha with the addition of a rectangular flang. [7] reports that in the lower frequencies i.e., frequencies up to 7 kHz, the acoustic response of the real and modeled ears looks very similar. More recently, authors in [8] explored the use of parabolic surfaces for the approximation of the pinna models. The experiments in these studies show that the important features, such as the first and third notches in the acoustic response of the pinna (PRTF), can be accurately modeled using these models. Fig. 3.5 shows that the HRTFs obtained from the KEMAR using the DB-61 pinna shape and the HRTFs calculated using the diffraction and reflection model in [8] are in good agreement.

3.1.2 Spherical Harmonics

The next technique discussed here for modeling the morphology of humans to study morphoacoustic problems is the use of spherical harmonics for modeling the head shapes of the listeners. [86] provides a review of the studies applying the spherical harmonics for shape modeling. If the surfaces to be modeled are given in the form of a function, the spherical harmonics framework, like any other transformation framework, can transform the given shapes into orthogonal components called spherical harmonics. Like, cartesian coordinate systems where any point in space is represented by x, y, and z components, in the spherical harmonic system, the surface functions are represented using spherical harmonics of different degrees and orders. Some of the sample basis functions for spherical harmonics are given in Fig. 3.6. However spherical harmonics can only be used to model a certain type of the surfaces which can be presented as a

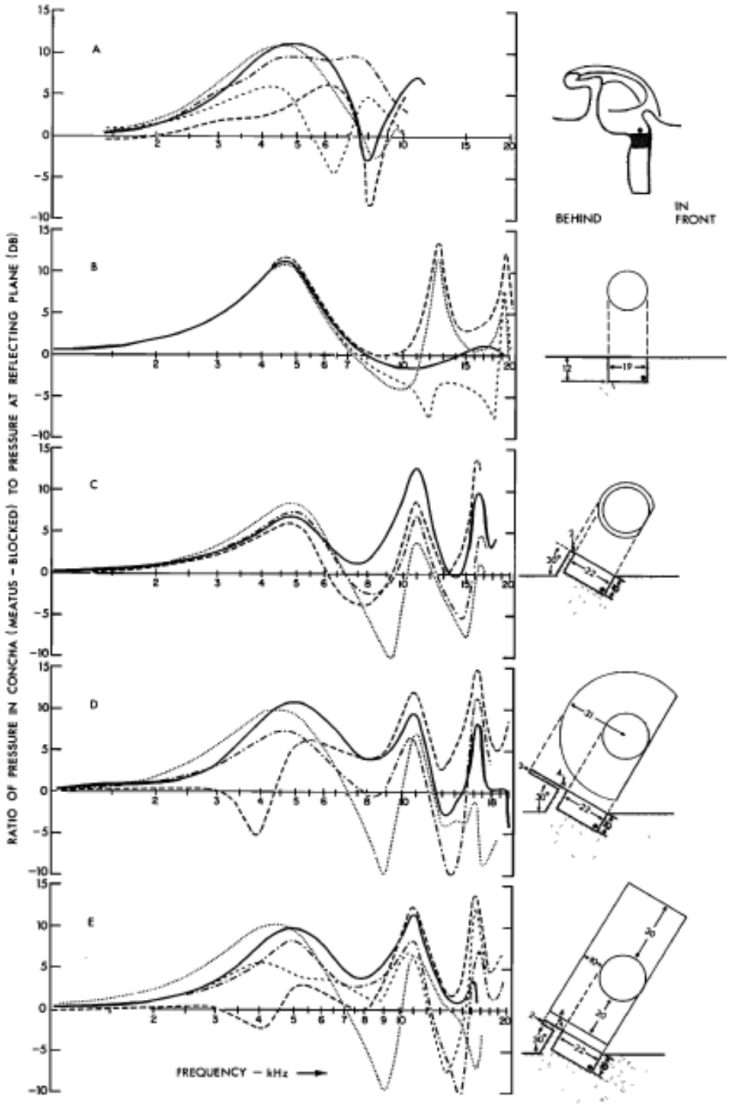


Figure 3.4: *The above shows (A) the average frequency response of the real ear shapes, averaged over six subjects. (B) cylindrical Concha, (C) tilted cylindrical Concha (D) cylindrical Concha with tilted segmented pinna (E) tilted cylindrical Concha with rectangular flang. It can be observed that adding the flang adds to the directivity of the cylindrical Concha . Reprint from [7]*

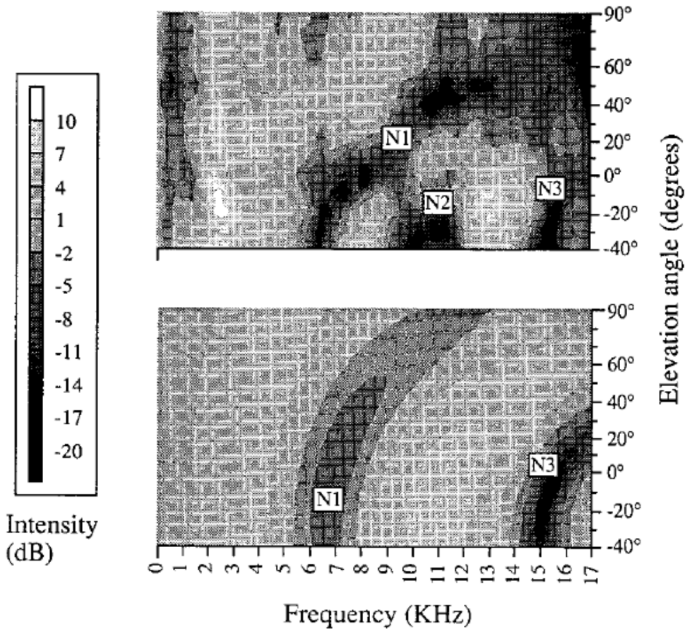


Figure 3.5: *This figure depicts the acoustic response of the KEMAR with DB61 pinna attached (top plot) and the acoustic response of the modeled pinna using the diffraction and reflection model of a parabolic sheet. The results presented in the figure show that the first and third nothces N_1 and N_3 can be modeled reasonably well. Image taken from [8]*

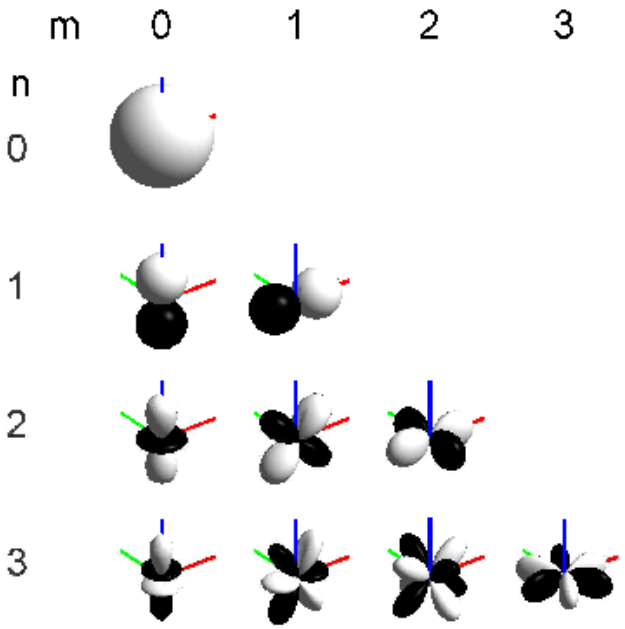


Figure 3.6: *First 10 spherical harmonic functions. n and m represents the order and degree of the spherical harmonics respectively.*

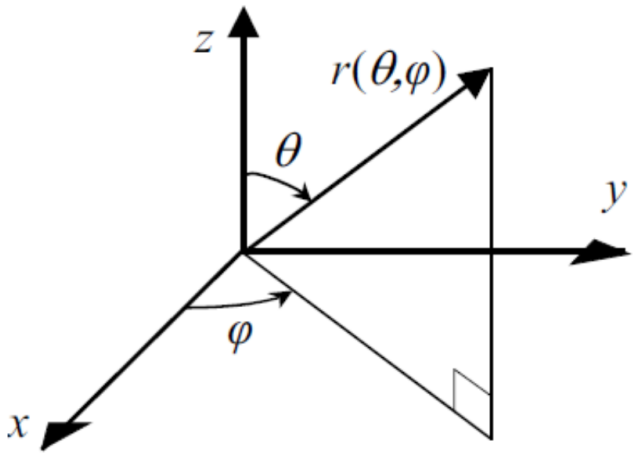


Figure 3.7: *Spherical coordinate systems. Taken from [9]*

function of θ and ϕ, and have only one value for ever (θ, ϕ) pair, i.e. any surface which is a function of θ and ϕ.

In other words any surface function $r(\theta, \phi)$, can be represented using spherical harmonics [9, 87, 88] as:

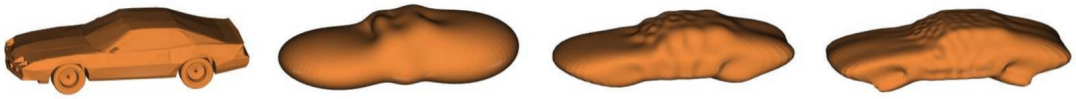
$$r(\theta,\phi) = \sum_{n=0}^{\infty} a_n P_n(\cos\theta) + \sum_{n=0}^{\infty}\sum_{m=1}^{n} P_n^m(\cos\theta) \times [a_{nm}\cos(m\phi) + b_{nm}\sin(m\phi)], \quad (3.1)$$

where in the above P_n^m represent the Legendre Polynomials of degree n and order m. Also, one thing to be noted is that we will be using only the non-negative values for m, such that $0 \leq m \leq n$. To model any shape function $r(\theta, \phi)$, perfectly all the spherical harmonics of degrees n are to be used where $0 \leq n \leq \infty$. However, in most cases this is neither efficient nor required. Hence, only a limited degrees of Legendre polynomials are used by truncating the higher order polynomials. This truncation results into a low-pass filtered shape, i.e. some of the details of the shape are lost during this process. Eq. 3.2 shows the truncated version of the representation of the function $r(\theta, \phi)$ using spherical harmonics when the Legendre polynomials only to the order N are used.

$$r(\theta,\phi) = a_{00} + \sum_{n=0}^{N}\sum_{m=1}^{n} P_n^m(\cos\theta) \times [a_{nm}\cos(m\phi) + b_{nm}\sin(m\phi)] \quad (3.2)$$

The coefficients a_{nm} and b_{nm} are called the spherical harmonic coefficients and are computed as:

$$\begin{aligned} a_{nm} &= \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} r(\theta,\phi)\bar{Y}^1(\theta,\phi)\sin(\theta,\phi)d\theta d\phi \\ b_{nm} &= \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} r(\theta,\phi)\bar{Y}^0(\theta,\phi)\sin(\theta,\phi)d\theta d\phi \end{aligned} \quad . \quad (3.3)$$



Original 8^2 harmonics 16^2 harmonics 24^2 harmonics

Figure 3.8: *Multi-resolution representation of the function* $r(u) = maxr \geq 0 | ru \in I_U 0$ *used to derive feature vectors from Fourier coefficients for spherical harmonics.*

Here, the functions $\bar{Y}^1$ and $\bar{Y}^0$ are the real spherical harmonic representations and can be written as follows:

$$\begin{aligned} \bar{Y}^1 &= \frac{1}{N_{nm}} \cos(m\phi) P_n^m(\cos(\theta)) \\ \bar{Y}^0 &= \frac{1}{N_{nm}} \sin(m\phi) P_n^m(\cos(\theta)) \end{aligned} . \tag{3.4}$$

Where, N_{nm} is a normalization factor for the real spherical harmonics [89], and is given by:

$$N_{nm} = \sqrt{\frac{4\pi}{\epsilon_m} \frac{1}{2n+1} \frac{(n+m)!}{(n-m)!}} \qquad \epsilon = \begin{cases} 1, & \text{if } m = 0 \\ 2, & \text{otherwise.} \end{cases} \tag{3.5}$$

The value of truncation order N is usually a trade off between the computation and memory required against the accuracy required, and is usually adjusted depending on the application. Some of the examples for the truncation to get the low-pass filtered shapes are given in Fig. 3.8. In [9], authors used spherical harmonics to study the shape of the head for morphoacoustic studies. A low-pass version of the KEMAR head modeled using truncated spherical harmonic representation is presented in Fig. 3.9. The truncation order used for this figure is $N = 17$. To do the morphoacoustic studies on any shape, the shape is first modeled using spherical harmonics and the coefficients a_{nm} and b_{nm} are calculated. Then different deformations of the shape are obtained by doing small perturbations to these coefficients. However, the perturbations that can be applied to the coefficients are limited and restricted. The details on this will be provided in the section 3.3. The study [9], also investigated the accuracy of the representations of spherical harmonic using $n = 0$ to $n = N = 34$ degree Legendre polynomials in the spherical harmonic expansions for KEMAR head. This study reports both the shape errors, as well as the errors in the corresponding acoustic pressure in the acoustic response calculated using numerical simulations for the modeled low-pass filtered head shape. The Root Mean Square shape error (RMS) between the reference shape and the shape reconstructed with degree n Legendre polynomials is given by:

$$\epsilon_{r_{rms}}(n) = \sqrt{\frac{\sum_{i=1}^{2N^2} [r_n(i) - r_N(i)]}{2N^2}} \tag{3.6}$$

The domain function described in Eq. 3.6 is representing the shapes and consists of a 3D grid. This grid has N distinct points for the vertical angles ϕ and $2N$ points for horizontal angles θ with a total resolution of $2N^2$ mesh grid of points. The percentage

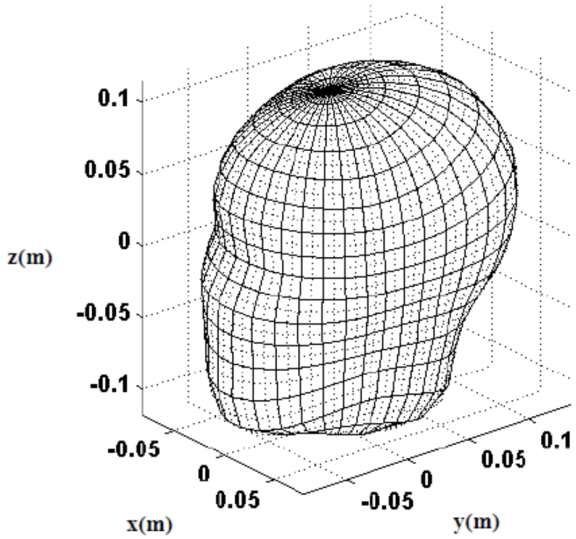


Figure 3.9: *Simplified head model by low pass filtering using spherical harmonics for KEMAR head shape. the truncation order used for reconstructing the shape is N = 17. Image taken from [9]*

error for the reconstruction of the KEMAR head is shown in Fig. 3.10 as a function of various values of n. This graph shows that head shape can be reconstructed quite accurately by using a truncation order $N = 15$ for all the sections except for the nose region, which requires higher degree Legendre polynomials for accurate reconstruction. The plots for the generated pressure field error are discussed in Sec. 3.3.1. Although the results of the head shape modeling are promising when the spherical harmonics are used, a big disadvantage in using the spherical harmonics is that it can not be used for modeling the head shape with pinna on it. The reason for this is that this nullifies the following conditions for the suitability of the surface function to be modeled using spherical harmonics, i.e., the surface function $r(\theta, \phi)$ has to be a one to one function of (θ, ϕ). Hence it can not be modeled with the equation 3.2.

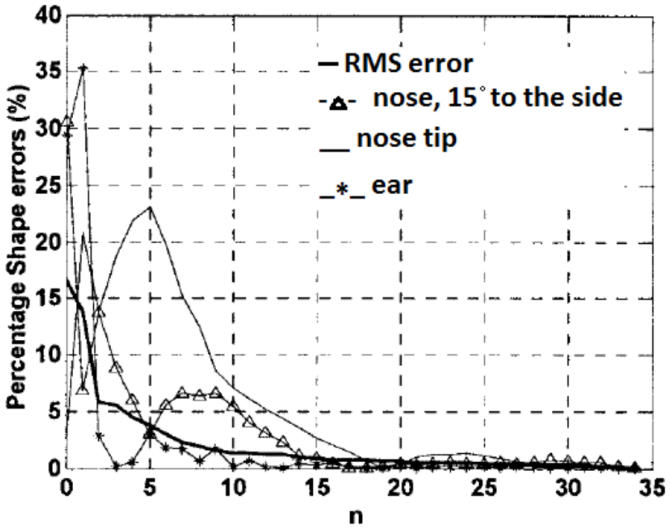


Figure 3.10: *The above graph plots the RMS error between the reconstructed KEMAR head shape using Legendre polynomials of degree n and a reference head shape. Image taken from [9]*

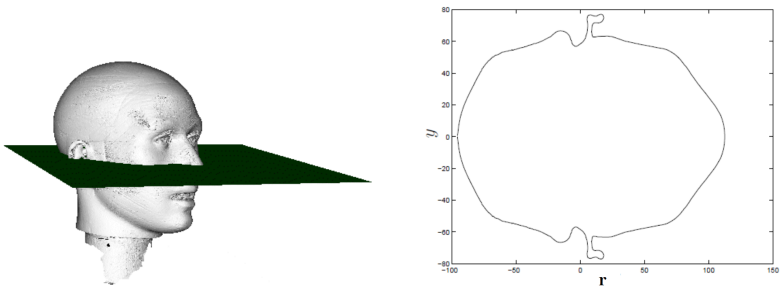
.

3.1.3 Elliptical Fourier Transform

The last popular technique reviewed here to model the 3D shapes and then deform these for morphoacoustic studies is the Elliptical Fourier Transform (EFT). 3D Elliptic Fourier descriptors EFD3D were originally introduced in [90] for a parametric representation and reconstruction of 3D shapes. EFT is a double Fourier transform of serial cross-sectional contours of shape in three dimensions (3D). This transform retains all the necessary information about the shape and provides a compact and invariant representation of 3D shapes. Furthermore, this also quantifies the volume enclosed in the 3D surface.

This modeling method was further used by authors in [10,91] to model the head and pinna shapes of the human subjects to perform the morphoacoustic studies and study the HRTF estimations. One major advantage of using EFT for morphoacoustic studies of humans is its ability to model the head shape with pinna. However, as highlighted in the previous section, spherical harmonics failed to do so because of their limitation to model only the surfaces which can be presented as functions.

The EFT model used by [10,91] is a modification of the EFT method originally proposed in [90]. These two studies suggested that to use the EFT for parametric modeling of the head and ears, the head and ear mesh is needed to be aligned in a way that the y-axis passes through the ear canals, i.e., the inter-aural axis lies on the y-axis. Then by rotating the head and pinna shapes with a step angle of $2\pi/S$, multiple intersections of the shape are created with the XY plane. In other words, the head and pinna (or only pinna shape) are rotated for an angle θ, where θ takes the values $0, \alpha, 2\alpha, \cdots, 2\pi - \alpha$. Fig. 3.11 shows this process of rotation and creation of cross-sections. In the Fig. 3.11a the rotation of the head and pinna surface is shown, while in Fig. 3.11b, the resulting cross-section with the XY plane is provided. Once all the $S + 1$ slices are obtained they are then regularised using the linear interpolation function and then the parametric



(a) (b)

Figure 3.11: *(a) shows the intersecting plane with the head shape while,(b) shows an example contour (i.e slice) of the head and ear shape. Image taken from [10]*

form of these slices is computed based on the parameter t. Where the parameter t, can take values from, $t = 0, 1, 2, 3, 4, \cdots, T$. For each of the x and y components two function $f_s^x[t]$ and $f_s^y[t]$, are constructed which depend on the parameter t. The EFT on these functions is computed separately as shown in Eq. 3.7 using sequential elliptic fourier transform.

$$\begin{aligned} A_x[s,n] &= \sum_{t=0}^{T-1} f_s^x[t] e^{\frac{-jnt}{T}} \\ B_x[m,n] &= \sum_{s=0}^{S-1} A_x[s,n] e^{\frac{-jms}{s}} \end{aligned} \tag{3.7}$$

This results in a set of complex-valued coefficients $A_x[s,n]$ and $B_x[m,n]$. These coefficients provide the parametric representation of the head and pinna shapes. The perturbations introduced in any of these coefficients $A_x[s,n]$ or $B_x[m,n]$, will result in deformation in the head and pinna shape. This method can be used to study the morphoacoustic perturbation analysis for HRTFs. However, the problem with using EFT for morphoacoustic studies of humans is that perturbations in the coefficients of EFT fail to provide smooth and evenly distributed spatial deformations, like spherical harmonics. To be more specific, the deformations in the head and pinna produced using EFT perturbation result in non-linear changes in the acoustics and made the mapping difficult, which is not desirable for morphoacoustic analysis.

In [11], authors proposed a further adaptation to the EFT proposed in [10, 91]. The technique proposed in [11], suggest to apply the deformations perpendicular to the contour of the slices (Fig. 3.11b). The generated coefficients $A_{u,v}$ and $B_{u,v}$ are called surface harmonic amplitudes. Similar to the $A_x[s,n]$ and $B_x[m,n]$, perturbing $A_{u,v}$ and $B_{u,v}$ also results in the deformations of the shape. Some of the examples for the perturbation of the coefficients $A_{u,v}$ and $B_{u,v}$ for two values, u and v are given in Fig. 3.12.

3.2 LDDMM on SYMARE Ear Shapes

This section provides an overview of the past studies which used the LDDMM framework to model the outer ear shapes. We have already seen how the LDDMM is used

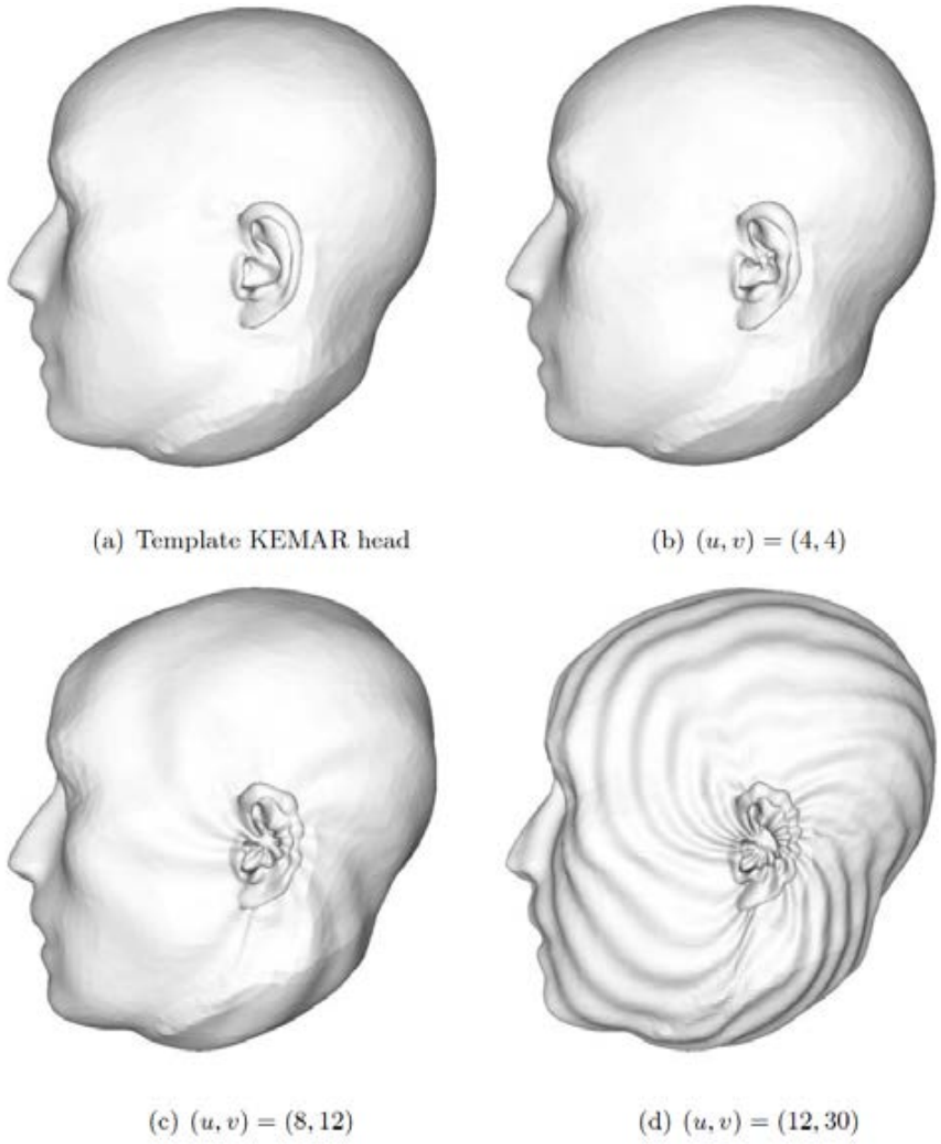


Figure 3.12: *The above figure shows the effect of perturbing the surface harmonic amplitudes which is detailed in [11] for a range of u and v. u is cross harmonic and v is the slice harmonic. Image taken from [11]*

to quantify the deformations and shape differences in the outer ear shapes in Sec. 2.6. In this section, we particularly look at the template calculation using the LDDMM framework for both single and multi-scale approaches presented in [17]. Furthermore, we provide a brief overview of how the morphable parametric model for the outer ear shapes was generated.

3.2.1 Template Calculation

This section discusses the method used to calculate the template or average shape for the head, ear, and torso for the SYMARE population. We will see both single and sequential multi-scale methods for calculation of template ear for the SYMARE user population using the LDDMM framework discussed in 2.6. In this work, the importance of template or average shapes is very high. A lot of studies have been conducted to calculate the template shape for a set of shapes using LDDMM framework. Almost all of these approaches use a single and fixed LDDMM scale, which is chosen to be smaller than the smallest feature to be modeled [92–94]. In [5], an improved version of this template creation was proposed, which uses a sequential multi-scale calculation approach. The calculation of the template is a very lengthy and computation hungry process. The calculation of the template using both single and multi-scale calculations can be sped up by first finding a barycenter or rough average of the shape population, which then can be a seed to the process. In [94], author proposed a quick method to estimate the barycenter for the shape population. The proposed barycenter calculation algorithm is presented in Algo. 2. This algorithm works like a moving average filter

Algorithm 2 Barycentre Calculation

inputs: $\{S_1, S_2, \cdots, S_R\}, \sigma_V, \sigma_W$
outputs: B
for $r = 2$ to R **do**
  $\{\alpha\} \leftarrow \mathcal{M}(B, S_r, \sigma_V, \sigma_W)$
  $B \leftarrow \mathcal{F}(S_r, \{\alpha^r(t)\}, \sigma_V, 0, \frac{1}{r})$
end for

and performs a moving average on the shapes using LDDMM framework. The algorithm begins by initializing the barycenter B by the first shape in the population S_1. Then it calculates the momentum vectors for matching the shape S_1 to shape S_2, using matching operation $\mathcal{M}$, described in section 2.6. Once done with this process, it uses the geodesic shooting and initial momentum vectors. It maps the barycenter B using the flow function $\mathcal{F}$ in the LDDMM framework, by setting the ending point as halfway or $T = \frac{1}{2}$. The B is updated and set to the newly obtained shape. The process continues for all iterations r by finding the mappings from the current barycenter to shape S_r and then using the flow function to the time point $\frac{1}{r}$, until $r = R$. Where R denotes the total number of ear shapes in the whole population. At this point, the barycenter variable B contains the barycenter for the shape population. This barycenter shape is then used as a seed to the template calculation function as a rough estimate of the template shape. Following, we provide details for both single and multi-scale template calculation methods using the calculated barycenter.

Single-scale Template Calculation

The template calculation method exploits the geodesic shooting and flow functions in the LDDMM framework 2.6 originally proposed in [83]. The template ear denoted by T also called the Fréchet mean, for a shape population $\{S_1, ..., S_R\}$ is a shape for which the sum of the geodesic LDDMM distances to all other shapes in the population is minimum:

$$T = \underset{U}{\operatorname{argmin}} \sum_{r=1}^{R} \int_0^1 \|\mathbf{v}_r(t)\|_V^2 \, dt \tag{3.8}$$

The assumption here is that $v_r(t)$ provides an exact mapping from the source shape U to a target shape S_r, for every r, $1 \leq r \leq R$. Considering the complexity of the ear shapes and limited computational resources, obtaining a template for a big population can be very hard, particularly when the gradient descent algorithm is used, the number of the iterations are only limited due to time constraint.

Here we explain the template shape estimation algorithm originally proposed in [82] and further improved in [5]. In [5], instead of using the landmarks for computing the shape differences, authors used current based distances for the LDDMM cost function J (revisit to Sec. 2.6) for more details. Furthermore, unlike the template calculation procedure in the [82], the starting point is the barycenter calculated using the algorithm 2. Again, the template shape must be a shape for which the squared norm of the initial momentum vectors from the template shape to all other shapes is zero, which means the template shape is the real center of the shape population.

$$\begin{aligned} & T = 0 \\ where \quad & \sum_{r=1}^{R} \alpha^r(0) = 0 \\ and \quad & \{\alpha^r(t)\} = \mathcal{M}(U, S_r, \sigma_V, \sigma_W) \end{aligned} \tag{3.9}$$

In practice the value of $\sum_{r=1}^{R} \alpha^r(0) \approx 0$ and not exactly equal to zero due to the reason that only a limited number of iterations are performed. The steps to compute the single scale template are given below in Algo. 3. At each iteration, the momentum vectors that map the template shape to each of the ears are computed, and the average initial momentum vectors are calculated. These average initial momentum vectors are then used to transform the current template shape to the new template shape using geodesic shooting. This process continues until the convergence has reached. As highlighted before the aim is to keep repeating until there is no change between the current and old template shape or the average of the initial momentum vectors is zero, however as we have only limited time and resources, only a limited amount of iterations are performed with setting a small threshold of change to be the limit.

Multi-scale Template Calculation

The second method for template calculations discussed in this work is the sequential multi-scale method. The motivation behind using a multi-scale template calculation is that it is observed that for scale parameters σ_V and σ_W different values result in different results. Furthermore, the LDDMM scale parameters have a direct impact

Algorithm 3 Single Scale Template Estimation

inputs: $\{S_1, S_2, \cdots, S_R\}, \sigma_V, \sigma_W$, **Optional:** B, I.
outputs: T
if B is not provided **then**
  $T \leftarrow S_1$
else
  $T \leftarrow S_1$
end if
if I is not provided **then**
  $I \leftarrow 20$
end if
for $i = 1$ to I **do**
  for $r = 1$ to R **do**
    $\{\alpha^r(t)\} \leftarrow \mathcal{M}(T, S_r, \sigma_V, \sigma_W)$
    $\{\hat{\alpha}\} \leftarrow \frac{r-1}{r}\{\hat{\alpha}\} + \frac{1}{r}\{\alpha^r(0)\}$
  end for
  $T = S(\hat{\alpha}\}, T, \sigma_V)$
end for

on the deformations obtained between the template shape and the ear shapes in the database, i.e., large values of σ_V imply a larger coupling between neighboring vertices in the source shape. In contrast, a smaller value will results in a smaller coupling at the time of the deformation. The same happens with the σ_W; for a larger value coarser differences are considered when performing LDDMM shape comparison, while a smaller value of σ_W computes the difference to finer detail and is more accurate.

The second benefit of using the multi-scale approach is that it results in a smoother and natural LDDMM deformations. Starting from a high-level matching, such that matching the rotation, translation, and scale operations, then it drills down to finer and finer details with each coming step, such as matching the contours and cavities of the ear shapes.

The detailed procedure to calculate the template using multi-scale LDDMM is described in Algo. 4. It uses the single scale template calculation in an iterative way using successively smaller scales. The aim of choosing the multi-scale template calculation method over the single scale is to have a better convergence of the Eq. 3.8. To this end, we use multi-scale and iteratively compute the template.

Algorithm 4 Sequential Multiscale Template Estimation

inputs: $\{S_1, S_2, \cdots, S_R\}, [\sigma_V(1), \sigma_V(2), \cdots, \sigma_V(L)], [\sigma_W(1), \sigma_W(2), \cdots, \sigma_W(L)]$ **Optional:** B, I.

outputs: T^L
$S_r^0 = S_r$ for $r = 1, 2, \cdots, R$
for $l = 1$ to L **do**
  $B^l \leftarrow barycenter(S_1^{l-1}, \cdots, S_R^{l-1}, \sigma_V(l), \sigma_W(l))$
  $T^l \leftarrow TempEstim(S_1^{l-1}, \cdots, S_R^{l-1}, \sigma_V(l), \sigma_W(l), T_o)$
  for $r = 1$ to R **do**
    $\{\alpha^r(t)\} \leftarrow \mathcal{M}(S_r^{l-1}, T^l, \sigma_V(l), \sigma_W(l))$
    $S_r^l \leftarrow \mathcal{F}(S_r^{l-1}, \{\alpha^r(t)\}, \sigma_V(l))$
  end for
end for

3.2.2 Driving a Morphable Ear Shape Model

This section provides the details on how a using LDDMM framework a morphable model for ear shapes cane be created [17]. The aim of creating a morphable model is to have a parametric model of the ear shape which can assist in defining a mapping function between the morphology of the ears and the corresponding acoustic responses. The morphable model generation approach proposed in [17] creates a parametric model that compactly models the 3D representation of the ear shapes using few parameters.

Modeling the ear shapes is a complicated and challenging task, however, as we have shown in Sec. 2.6 and Sec. 3.2.1, the LDDMM framework does a tremendous job to model the variations in the ear shapes as deformations. We also described in Sec. 3.2.1, how a template ear shape can be calculated for a given ear shape population using single or sequential multi-scale LDDMM approach. The calculated template shape is the centerpiece of the morphable model derivation. We have also provided the details on the KPCA in Sec. 2.7. We also have seen how using only the initial momentum vectors, the template shape can be deformed to other target shapes through geodesic shooting in Sec. 2.6. The readers are advised to review these aforementioned concepts before reading this section.

[17] proposed a three step procedure to compute the parameters for a new shape S_{new}.

Step1: The first step is to find the momentum vectors $\alpha_{new}(t)$ that map the multiscale template shape T to the new shape S_{new} using mapping function $\mathcal{M}(T, S_{new}, \sigma_V, \sigma_W)$ of LDDMM framework.

Step2: As highlighted before any shape can be represented having the reference of template shape and initial momentum vectors for the mapping in hand. So in this case also only the initial momentum vectors are used. In second step the initial momentum vectors $\alpha_{new}(0)$ or α_{new} are centered by subtracting the mean intimal momentum vectors $\bar{\alpha}$ (given in equation 2.32) from the new initial momentum vectors as:

$$\hat{\alpha}_{new} = \alpha_{new} - \bar{\alpha}. \tag{3.10}$$

Step3: In the third step the projections of the initial momentum vectors are computed over the KPCA components, to find the parameters $\tilde{v}_{new}$. The detailed algorithm for this procedure is given in the algorithm below.

Algorithm 5 Calculating the parameters for morphable model of a new shape

inputs: $U, \bar{\alpha}, S_{new}, \sigma_V, \sigma_W$
outputs: $\tilde{v}_{new}$
$\{\alpha_n^{(new)}\}_{1 \leq n \leq N} = \mathcal{M}(T, S_{new}, \sigma_V, \sigma_W)$
$\hat{\alpha}_{new} = \alpha_{new} - \bar{\alpha}$
$\tilde{v}_{new} = U^T K \hat{\alpha}_{new}$

As highlighted before in KPCA and PCA we can use any number of principal components n in range of $1 \leq n \leq N$. More the number of principal components used the accurate the result is, with using all the principal components we can construct the original ears. This model provide one with a capability of modeling the complex pinna shapes with only few numbers and aid a lot in terms of studying the morphoacoustics of the ear shapes.

3.3 Morphoacoustic Perturbation Analysis(MPA)

Anthony Tew initially coined the term Morphoacoustic Perturbation Analysis or (MPA) for short in his paper [11]. He proposed a method which studies the changes in the acoustic domain by adding small perturbations to the morphology of the subject and study the corresponding changes in the acoustic transfer functions. Mostly, the perturbations applied to the template shape are tiny to keep the effect of the corresponding changes in the acoustic domain linear. Although this term is new and came in 2012 for the first time, there have been many studies before which use the same method to study the effects of small perturbations on a reference pinna shape on the acoustic features of the corresponding HRTFs, like [3, 22, 95]. All these studies used the KEMAR [96] shape as the reference.

3.3.1 Differential Pressure Synthesis

As discussed before, morphoacoustic is a study which performs some small perturbations in the morphology and studies the corresponding modifications in the acoustics of the shape to unveil the underlying physical phenomenons that create certain acoustic features. In Sec. 3.1, provides an overview of various techniques to model the morphology of humans with interest to study the morphoacoustics. In this section, we discuss Differential Pressure Synthesis, a technique which is particularly used to study the morphoacoustics when the morphology is modeled using spherical harmonics [95]. This technique aims to rapidly compute the acoustic pressure field around a deformed template shape. The pressure field around the deformed shape can be computed using DPS without running the computationally expensive BEM simulations again and again after each deformation. DPS technique relies on a lookup table in which the acoustic responses for different spherical harmonic modes saved already. In DPS, the pressure field around the deformed template shape is calculated using scaling and summing these pre-computed pressure fields corresponding to orthogonal transformations, which are obtained by modeling the template shape through spherical harmonics.

Frequency	250 Hz	500 Hz	1 kHz	2 kHz	3 kHz
% Error	0.92	0.22	0.69	18.9	34.6

Table 3.1: *Percentage errors between the pressure field computed using BEM simulations, and the one approximated using the DPS when a sphere is deformed to match the pinna-less KEMAR mannequin. The image is taken from [9].*

These deformations are constrained as the modifications in the acoustic responses are dependent on the wavelength of the incident sound waves. Keeping these constraints in view, first order mappings between shape changes and the corresponding changes in the acoustic pressure fields are formulated. Considering the original pressure field around the template shape is Φ_o, after the deformation applied to the template the new pressure field Φ can be calculated as :

$$\Phi = \Phi_0 + \Delta\Phi, \tag{3.11}$$

where $\Delta\Phi$ denotes the changes in the pressure field. In brief, the method for calculating the $\Delta\Phi$ is obtained by a first-order Taylor series expansion that relates spherical harmonic coefficients representing the shape changes to the pressure changes [95]. In [9],

authors conducted an experiment in which they deformed a spherical template shape into pinna-less KEMAR head. Tab. 3.1 reports the percentage error between the pressure field computed through BEM simulations and the one approximated using DPS. The table shows that the accuracy of using DPS is very high up-to a frequency of $1kHz$, and it starts to drop greatly as the frequency increases, exhibiting large errors at higher frequencies. The table also shows the errors at $2kHz$ and $3kHz$. The reason for this performance degradation is that at these frequencies, the wavelengths of the incident waves become comparable to the size of the deformations. As the frequency increases, the wavelength becomes finer and finer. In summary, DPS requires the deformations from the template shapes to be moderate, depending on the sound wavelength under consideration, and not satisfying this criterion results in large errors.

3.3.2 Morphoacoustic Perturbation Analysis Frequency Domain (MPA-FD)

The inherent limitations of the DPS which says that the deformations from the template shape need to be linear with respect to the acoustical changes, it fails to study large variations in the ear shape.

To address this shortcoming, the authors in [11] introduced a frequency domain Morphoacoustic Perturbation Analysis (MPA-FD). Just like the DPS, the MPA-FD can be used to analyze the changes in the acoustics when a small perturbation is applied to a template shape. Additionally, it enables one to study the perturbations in a particular view to understand which perturbations create significant changes in a particular feature of the HRTF (i.e., in the central frequencies of notches in HRTF spectrum). Furthermore, MPA-FD also overcomes one of the key limitations of the DPS not to able to model the pinna shapes in the head and pinna mesh [95]. Just like DPS, the MPA-FD can rapidly identify interesting mappings between morphological features and corresponding acoustics by constructing a database of orthogonal deformations and the corresponding acoustic changes. However, the orthogonal deformations in the template shape are conducted using Elliptical Fourier Transforms (EFT) (refer to Sec. 3.1.3 for more details on EFT). The acoustic responses for the deformed template shapes are computed using BEM simulations. Fig. 3.13 shows the identified regions in the template ear shape that contribute significantly towards the generation of the first notch in the HRTF spectrum. In summary, MPA-FD is quite useful when it comes to identifying and relating the regions in the ear shape that influence a particular feature in HRTF and vice versa. However, similar to the DPS technique, the perturbations should be tiny, and the formulation requires a linear relationship between the morphological and acoustic changes. For this reason, the shape changes and their corresponding acoustic responses to create the database are conducted with tiny perturbations only. Consequently, this method can not be efficiently used to perform the morphoacoustic studies on ear shapes of multiple humans as the morphology of the ear shapes changes considerably across the population of humans.

3.3.3 Acoustic Sensitivity to Micro-perturbations of KEMAR's Pinna Surface Geometry

Following the same lines, authors in [12] studied the effects of small perturbation in voxelated KEMAR ear shape along with a small head patch. Studying these effects, they examined the acoustic sensitivity of different regions of the ear shape on

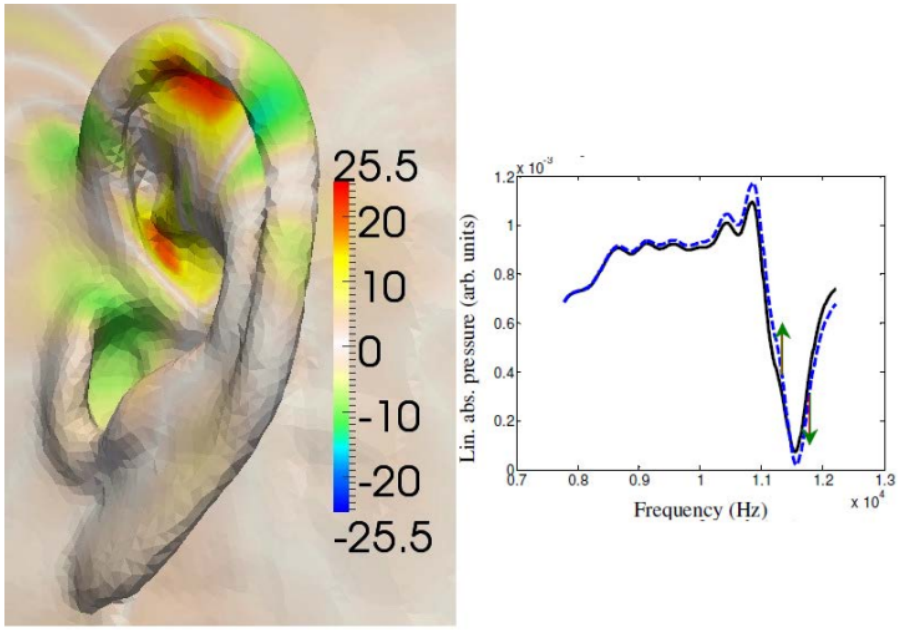


Figure 3.13: *The plot on the right shows an HRTF spectrum (solid line) with a notch in the frequency seen between 11 kHz and 12 kHz. The dotted blue line shows an HRTF spectrum in which the notch has been shifted to higher frequencies. The green arrows show the movement of the spectrum as the notch moves to a higher frequency. The ear on the left is the template ear shape with the regions that contribute towards the formation of the notch colored with warm (red) and cold (blue) colors. Image taken from [11]*

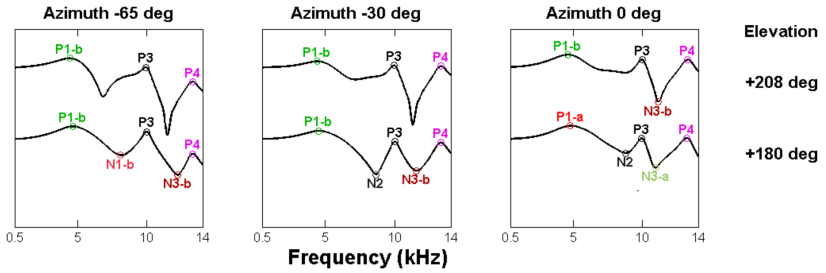


Figure 3.14: *Peaks and notch patterns for a series of PRTFs of the KEMAR pinna with a small head patch. Image taken from [12]*

the notches and peaks in the corresponding HRTFs. Using a Finite-Difference Time-Domain (FDTD) method, the HRTF for the reference model and the perturbed ear was computed. The voxelization happened in a way that the voxel grid had a uniform resolution of $2mm$. The resulting 3D model had a total of 1784 unique voxels. In this study, each of these voxels is inflated and deflated, creating 1785 FDTD simulations in total (one for original and 1784 for different voxels). Once the acoustic responses (Pinna related transfer functions PRTFs) for these shapes are calculated, the features (peaks and notches) in the resulting PRTFs are then compared with the features in the reference shape. Fig. 3.14 highlights the identification of a series of peak and notch patterns occurring for KEMAR PRTFs. Some of the key findings of the work [12] are:

1. The first peak generally has a center frequency around $5kHz$ and has three patterns denoted as $P1-a, P1-b, P1-c$. The $P1-a$ and $P1-b$ are mainly affected when the perturbations occur in the concha region. The sensitivity is maximum towards the back wall and decreases going towards the ear canal. $P1-a$ occurs in almost all directions. While $P1-b$ and $P1-c$ occur for some directions. $P1-b$ and $P1-c$ are also sensitive to the perturbations in the ear rim.

2. The center frequency for the third peak $P3$ is around $9.8kHz$. This peak appears consistently across all directions in space. This peak is sensitive to the perturbations in different regions of the Cymba concha, cavum concha, and triangular Fossa in both positive and negative way (i.e they shift the peak $P3$ to a lower or higher frequency value).

3. The first notch $N1$, also occurs in three different patterns, denoted as $N1-a$, $N1-b$ and $N1-c$. These patterns are direction-dependent. The pattern $N1-a$ appeared for sources from the lower front regions and had a center frequency around 7.1 kHz. It had positive sensitivity to a region that covers the Cymba concha and triangular Fossa. The $N1-b$ pattern occurs for sources in the front hemisphere near the horizontal plane, and it had a sensitivity to regions corresponding to the Cymba and the upper back wall of the Concha. While the $N1-c$ happened for twelve locations at high elevations and had a mean center frequency of 8.8 kHz. This pattern is affected by almost all the regions in triangular Fossa, Cymba, and Cavum Concha.

Figure 3.15 and figure 3.16 show the regions of the ear shape that had positive and negative sensitivity to the peaks and notches in the PRTF spectrum.

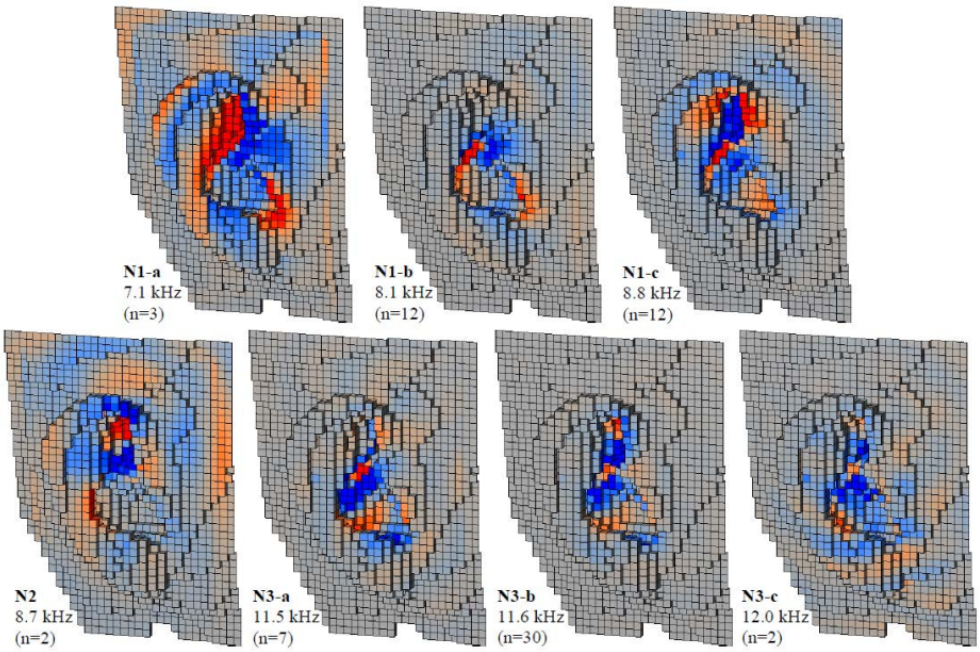


Figure 3.16: *Pinna sensitivity map for notches $N1 - N3$. The positive sensitivity is shown with warm (red) colors and negative sensitivity is shown with cold (blue) colors. Image taken from [12]*

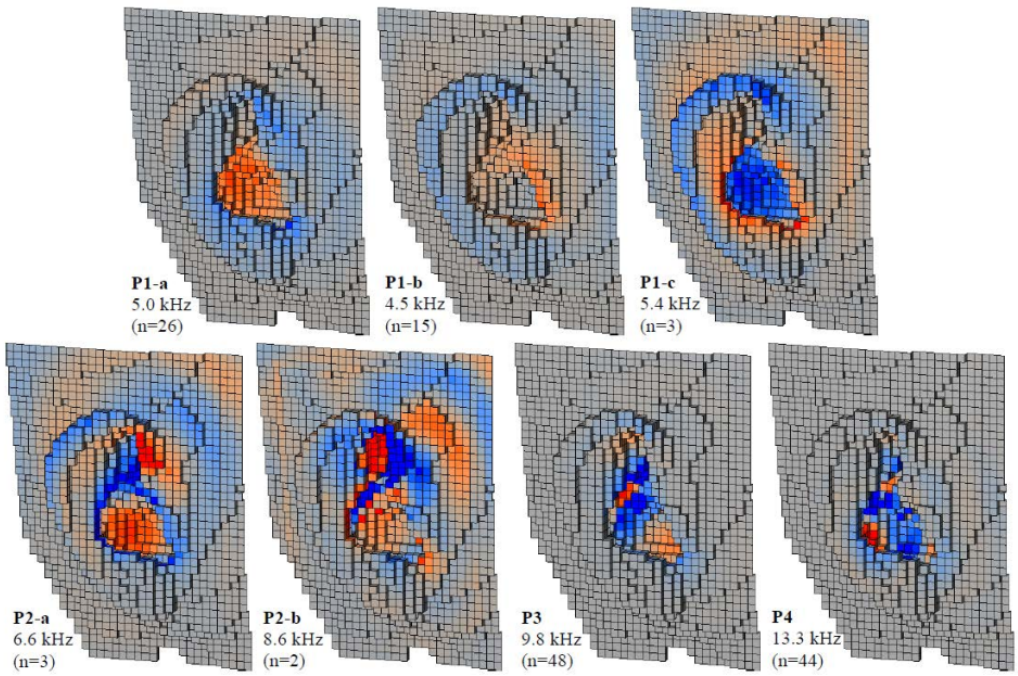


Figure 3.15: *Pinna sensitivity map for peaks $P1 - P4$. The positive sensitivity is shown with warm (red) colors and negative sensitivity is shown with cold (blue) colors. Image taken from [12]*

3.4 HRTF Individualization Methods

As highlighted in the section 2.2, personalized HRTFs are essential to provide VAS over headphones. At the same time, we have reviewed in section 2.3, that the acquisition of the individualized HRTFs is a costly and specialized task, requiring expensive, specialized equipment and operators. Hence it is limited to the laboratory environment only. To solve this problem, a compromise has to be made, and instead of using individualized HRTFs personalized HRTFs are to be used, which provide a listening experience very close to the individualized HRTFs. Although, there have been plenty of studies trying to solve this problem, to the best of author's knowledge there is not a single comprehensive framework available that can provide personalized HRTFs for a listener in a quick, efficient and easy manner which can be used for commercial use for the mass market. A high-level survey on the state of the art methods for HRTF personalization was recently published by Guezenoc et al. [97], which provides a taxonomy and review of HRTF personalization methods. This study categorizes HRTF personalization methods and reviews the advantages and shortcomings of these methods. This section follows the same scheme and provides a brief overview of different HRTF personalization methods. The HRTF personalization methods can be categorized into two broad categories, namely: a) direct personalization, and b) indirect personalization. Some of the most famous studies in these two categories are discussed below. The direct HRTF individualization methods are the acoustic measurements and the numerical simulation based HRTF calculations. The details of these two methods, along with their limitations, have already been discussed in sections 2.3. Following, we discuss the details of the indirect HRTF individualization of HRTF personalization methods.

Indirect HRTF individualization methods can be divided into two main categories. Following, we provide details of these two main categories.

3.4.1 HRTF Personalization based on Anthropometric Data

HRTF of an individual is strongly dependent on the morphology of the listener. Considering this strong dependence, many studies have tried to find a mapping between the anthropometry and the corresponding HRTFs of an individual. The aim of all these studies is to provide a low-cost HRTF individualization method that removes the need for performing acoustic measurements or running long and computation hungry numerical simulations. These methods can be further divided into three subcategories.

a) Adaptation : The methods in the first category take the existing publicly available HRTF sets and adjust them to make them more suitable for an individual listener. Because different people have different sizes of head and ear shapes, authors in [98] proposed that the differences in the HRTFs can be reduced by simply scaling the frequency axis of the HRTFs of one listener to better match the HRTFs of another listener. A year later, through simple studies, the same scientists found that the scaling factor can be estimated using linear regression on the ratio of scales of head and ear shapes of two individuals. Through performing listening tests on 9 and 11 subjects in both studies respectively, they reported that the spatial hearing experience for the listeners has improved compared to the non-individualized HRTFs, while still worse than their HRTFs. Later two studies combined the scaling corrections with the spatial rotation corrections to account for the head tilts [99, 100]. These studies report a further improvement in

HRTF matching. However, they did not perform any perceptual experiments.
b) Selection : The second subcategory of anthropometric based HRTF personalization methods is the selection methods. In this category, an HRTF set for an individual user is selected from a publicly available HRTF dataset, which has both anthropometry and acoustics of the listeners. For instance, a study reported in [101] implements a coarse nearest neighbor based selection method on the anthropometric feature vectors to find a user from the CIPIC database [14]. Considering the strong dependence of the RHTFs on the anthropometry, the HRTF of the selected user will match the best to the user in question. As making the anthropometric measurements is a difficult process, this work uses only seven anthropometric measurements, which they measure from a picture of the pinna. The results of the listening experiments show an average gain of 15% in elevation score compared to non-individualized HRTFs. Authors in [102] used a rather intuitive and hybrid approach to find the closest matching HRTF. In their earlier work, they reported that the three main notches in pinna-related transfer functions are the results of the three main contours on the pinna. Furthermore, the center frequencies for these notches can be estimated with reasonable accuracy from a single scaled image of the pinna [23]. By choosing an HRTF from the database which has notches with notch frequencies closely matching the estimated notch frequencies of the subject in question, one can find a closely matching HRTF for the subject. The listening tests showed an improvement of 17% better elevation perception compared to the use of generic HRTFs.
c) Regressioin : The final subcategory in this group is regression-based HRTF personalization methods. In this method, the HRTFs of a given dataset are modeled using different dimension reduction methods such as principal component analysis (PCA) [103] and independent component analysis (ICA) [104]. Using multiple linear regression, a mapping between the anthropometry and HRTF parameters is created. Some of the studies have also used neural networks [105] and High-Order SVD [104]. However, considering the availability of limited data, this can very easily cause over-fitting. Some other groups have used an even more, simpler method and tried to predict the HRTFs of the listeners by considering this as a sparse representation based problem. Relying on a strong assumption that the anthropometric parameters of any individual can be represented by the linear combination of anthropometric features of the users in the dataset, and the HRTF of a subject can be represented by the same linear combination, they modeled the HRTF of a given individual using linear combination of the HRTFs in the dataset [28].

Although all these methods are low-cost in terms of effort and computational cost, the performance of these methods greatly depends on the selection and measurement accuracy of the anthropometric parameters. Any in-accuracy and a small error in the measurement can result in errors in the resulting HRTFs.

3.4.2 HRTF Personalization based on Perceptual Feedback

This group also has two major categories. These subcategories and their details are given as follows:
a) Selection : This type of methods are being used since late 1990s. In this personalization method, a listener is presented with the spatial audio rendered using different HRTFs available in the dataset, and the listener chooses the one which seems to provide

more natural experience. This scheme was used by the researchers in [106, 107]. The results of these works provide a good listening experience with a better localization performance compared to generic HRTFs. However, these studies use only the azimuth plane HRTFs for both selection training as well as for the evaluation. In azimuth plane, interaural cues are more important than spectral cues; hence these methods need to be tested for the vertical localization. Furthermore, these tests can take a different amount of time required for different listeners, and for these two studies, the time taken by subjects to choose a personal HRTFs was anywhere between 15 to 35 minutes. This time can be further reduced by first clustering the HRTF sets a priori based on objective or subjective measures as proposed by authors in [107, 108]. Furthermore, this also depends on the ability of a listener to localize audio sources, as well as being able to concentrate as well as being able to conduct listening experiments.
b) Adaptation : Another way to get an individualized HRTF set for listeners is to perform the adaptation to the existing HRTFs guided through the listening experiments. The past studies which use this approach can be divided into three different kinds [97]. The first type relies on the findings of frequency scaling studies proposed in [25, 26]. An example of this is proposed in [98], which showed the gain in terms of localization performance when the right scaling is used compared to an unscaled HRTF set. The biggest benefit and also the limitation of this method is that it lets us control the amount of personalization available through a single number, i.e., the scaling factor. The studies reported that by spending only 20 minutes, users were able to personalize the HRTFs well enough to provide almost the same performance as was provided by using the optimal scaling factor.

The other way of tuning the HRTFs is to model the HRTFs as different filters and adapt to the right parameters for these filters through listening feedback. Two studies of this kind were reported in [109,110]. The studies presented in [109] do not consider the direction dependency and try to solve the problem by following a simple filter equalization scheme. In contrast, the studies showed in [110] used a direction-dependent tuning. This requires to tune the filters for each direction separately, making the time needed to do this very large.

In order to address this problem of tuning a large number of parameters the third group of studies can be very useful. This group of works try to reduce the number of parameters to model by relying on the statistical modeling. Using a direction dependent PCA can reduce the number of parameters for a given direction as low as between three to five making required number of tunings small. Adapting these parameters direction by direction can provide a personalized HRTF [19, 111, 112]. The results of these studies have shown localization improvements compared to non-individualized HRTFs, however the number of listeners included in the tests were very few in [19, 111], while elevation perceptions were not evaluated in [112]. Although these studies reduce the number of tunings to be performed for a single direction reasonably, one still need to perform the tunings for every direction rendering this solution very impractical (only 9 to 10 directions were tuned in these studies). A later study provided a global parametric model by using spherical harmonics to model the PCA weights of the directional PCA in [113], reducing the total number to model the HRTFs to 45.

3.5 Evaluation Metrics for Shape Matching

In section 3.1, a detailed review of some popular shape modeling methods has been provided. This section provides a review on some of the evaluation methods to evaluate the performance of modeling of shape by measuring the mismatch between the original and the modeled shapes.

3.5.1 Vertex Distance

The first and most trivial distance metric for the shape matching to use the Euclidean distance between the point clouds of the original and matching shapes. This believes that both the original and mapped shapes have an equal number of points. Given the original shape S_{orig} and mapped shape S_{mod}, as defined in section 2.6.1, and given as $S_{orig}(X)$ and the mapped shape is given as $S_{mod}(Y)$, the vertex distance is given as:

$$\mathbf{d}_{(S,M)}(n) = \sqrt{\|x(n) - y(n)\|^2}, \tag{3.12}$$

where $\mathbf{d}_{(S,M)}(n)$ denotes the distance between the vertices n of shape S and M. This distance is presented as the color map on the ear shapes, where color for a point denotes the distance between that particular point index between two shapes.

3.5.2 Hausdorff Distance

The second most used distance metric in ear shape modeling is Hausdorff distance used in [61, 114]. Let us consider the original shape is denoted as $S_{orig}(X)$ and the modeled shape is denoted as $S_{mod}(Y)$, where X and Y denote the vertices for S_{orig} and S_{mod}, and $X, Y \in \mathbb{R}^3$. The Hausdorff distance between two shapes denoted by $d^H(S_{orig}, S_{mod})$ is computed as:

$$d^H(S_{orig}, S_{mod}) = \max\{d_H^Y = \sup_{y \in Y} \inf_{x \in X} \sqrt{\|x - y\|^2}, d_x^H = \sup_{x \in X} \inf_{y \in Y} \sqrt{\|x - y\|^2}\}. \tag{3.13}$$

The Hausdorff distance is originally proposed as a scalar quantity that measures the distance between two population of the points, and it provides only a single value. In shapes however, one is usually interested to have a look on the areas which are closely matched and areas which are not matched well. In order to do so a distance map Q is defined. The distance map contains one distance for each of the vertices. Let us denoted the value of Q at the point x of shape S_{orig} as $Q(S_{orig}, S_{mod}, x)$. This value is given as:

$$Q(S_{orig}, S_{mod}, x) = \inf_{y \in Y} \sqrt{\|x - y\|^2}. \tag{3.14}$$

This equation means that we will measure the distance from point x to all the points y in shape S_{mod}, and the minimum value is chosen as a value of Q. This is repeated for every point in S_{orig}. One thing worth noting is that in this distance, the number of points in both shapes can be different. Hence, the distances from one shape to other are not necessarily same as the other way around, e.g. $Q(S_{orig}, S_{mod}, x) \neq Q(S_{mod}, S_{orig}, y)$ i

3.5.3 Face Distance

Although Hausdorff distance provides a good representation of the matching, it only accounts for the proximity between vertices in the shapes through the Euclidean distance metric. In [17] authors proposed a new shape difference analysis technique based on currents introduced in [78]. They called it Current based Shape Difference Analysis or (CSDA). The CSDA works considering two main matching ways, **1)** It takes into account the proximity of points on two surfaces using kernel function described in equation 2.28, as well as, it considers the orientation of the normal vectors or currents on the mesh faces in a selected region. Another motivation of using this is the success of the metric $E(S_1, S_2)$ given in equation 2.25, capturing the shape difference while performing mapping for LDDMM.

Given the surface S_{orig} and S_{mod}, the current based distance $d(S_1, S_2, f)$ on a face f of the S_{orig} is defined as

$$\begin{aligned}
\beta_1(f) &= \sum_g k_R(c_{orig}(f)c_{orig}(g))\langle n_{orig}(f), n_{orig}(g)\rangle \\
\beta_2(f) &= \sum_h k_R(c_{orig}(f)c_{mod}(h))\langle n_{orig}(f), n_{mod}(h)\rangle \\
d(S_{orig},S_{mod}, f) &= |\beta_2 - \beta_1| \\
\hat{d}(S_{orig},S_{mod}, f) &= \min\left(\frac{d(S_{orig}, S_{mod}, f)}{|\beta_1(f)|}, 1\right)
\end{aligned} \tag{3.15}$$

where k_R is cauchy kernel with σ_R, $n_{orig}(f)$, is the normal vector on the face f, this points outwards, located at the center of the face with the length proportional to the area of the face f. $c_{orig}(f)$ is the center of the face f in shape S_{orig}. $\beta_1(f)$ provides the sum of the convolution of the normal vector at f with every other face g for the S_{orig}, while $\beta_2(f)$ provides the sum of the convolution between the normal vector at face f of the original shape with all the other faces h on the modeled shape S_{mod}. In case both surfaces are similar the distance is very small and for exact same surfaces it is zero. On the other hand when two surfaces are very different the value of β_2 will be very small giving us a large value for the distance. In order to obtain a meaningful representation [17] created a normalized similarity measure representation as $\hat{d}(S_{orig}, S_{mod}, f)$. The overall similarity measure is computed by summing the values of $\hat{d}(S_{orig}, S_{mod}, f)$ on all mesh faces as:

$$\overline{d}(S_{orig}, S_{mod}) = \frac{1}{F}\sum_{f=1}^{F} \hat{d}(S_{orig}, S_{mod}, f). \tag{3.16}$$

here F denotes the total number of vertices. One thing to be noted is that in general $\bar{d}(S_{orig}, S_{mod}) \neq \bar{d}(S_{mod}, S_{orig})$.

3.6 HRTF Evaluation Metrics

In this section, we provide popular evaluation methods to evaluate the performance of the personalization of HRTFs. These methods can be divided into two wider categories, namely: a) Objective, and b) Subjective or pyscho-perceptual metrics. Following, we provide some of the most famous methods for each of these categories.

3.6.1 Objective metrics

This section provides a review of the objective methods for evaluating the personalization of the HRTFs. These methods are purely mathematical and only consider how well two sets of HRTFs match without considering the perceptual relevance of certain features or content of the HRTFs.

Root Mean Square Error : A root means square error (RMSE) or root mean square deviation (RMSD) are the measures widely used to measure the difference between the values predicted by a model or estimator and real values. This has also been used for the HRTFs. The RMSE provides the square root of the second sample moment of the differences between the original and predicted values or the quadratic mean. These deviations are also called residuals when the calculations are performed over the data sample that was used for estimation and are called errors (or prediction errors) when computed out-of-sample. Given the an actual HRTF $\mathbf{H}(f_i, \theta, \phi)$, and approximated one given $\hat{\mathbf{H}}(f_i, \theta, \phi)$, for N frequency bins in total, the RMSE is computed as:

$$RMSE(\theta, \phi) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{H}(f_i, \theta, \phi)^2 - \hat{\mathbf{H}}(f_i, \theta, \phi)^2\right)} \tag{3.17}$$

For getting a global metric called global RMSE or GRMSE the error from all directions is combined using following equation:

$$GRMSE = \sqrt{\frac{1}{D}\sum_{(\theta,\phi)\in D}\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{H}(f_i, \theta, \phi)^2 - \hat{\mathbf{H}}(f_i, \theta, \phi)^2\right)} \tag{3.18}$$

Spectral Distortion : Spectral distortion is another measure for objective evaluation of the HRTFs. This is the mean square difference in log domain given as:

$$SD(\theta, \phi) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(20\log_{10}\frac{|H(f_i, \theta, \phi)|}{|\hat{H}(f_i, \theta, \phi)|}\right)^2} \tag{3.19}$$

where, $H(f_i, \theta, \phi)$ denotes the original HRTF and $\hat{H}(f_i, \theta, \phi)$ denotes the estimated or approximated HRTF, for direction (θ, ϕ).

Inter-subject Spectral Difference (ISSD) : Another famous method to evaluate the matching between two HRTFs is inter-subject spectral difference (ISSD). Authors in [24] calculated the ISSD for the HRTFs for frequencies between 3.7 kHz to 12.9 kHz using 64 Equivalent Rectangular Bandwidth (ERB) filter banks. The calculation of the ISSD between two HRTFs belonging to subjects S_1 and S_2, is calculated in three steps. In the first stage the HRTFs are converted into ERB filter banks and differences between each of the 64 bands was computed as:

$$\Delta DTF(\theta, \phi, f_i) = 20\log_{10}|DTF_{S_1}(\theta, \phi, f_i)| - 20\log_{10}|DTF_{S_2}(\theta, \phi, f_i)| \tag{3.20}$$

Using the $\Delta DTF(q, f, fi)$ computed in equation 3.20, the means for all the 64 frequency bands f_i are computed:

$$\bar{\Delta DTF}(\theta, \phi, f_i) = \frac{1}{64}\sum_{i=1}^{64}\Delta DTF(\theta, \phi, f_i) \tag{3.21}$$

In the second stage the variance in $\Delta DTF(\theta, \phi, f_i)$ for every direction in space (θ, ϕ) are computed as:

$$\sigma^2(\theta, \phi) = \frac{1}{64} \sum_{i}^{64} \left\| \Delta DTF(\theta, \phi, f_i) - \bar{\Delta DTF}(\theta, \phi, f_i) \right\|^2 \tag{3.22}$$

Finally, in the third stage the ISSD between subject S_1 and S_2 is calculated as:

$$ISSD_{S_1,S_2} = \frac{1}{M} \sum_{\theta,\phi} \sigma^2(\theta, \phi) \tag{3.23}$$

Cross-Correlation : Cross-correlation is one of the most popular approaches to measure the similarity between two signals or sequences. In [51], authors employed the cross-correlation to validate the use of BEM simulations to identify if the acoustically measured HRTFs and BEM simulated HRTFs are matching to each other. In statistics, the Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables X and Y. Its denoted by r and lies in the range $-1 \leq r \leq 1$. A value equal to 1 means a totally positive relationship. A value equal to -1 means a negative relationship, while value equal to 0 means no correlation at all. Originally developed in 1880, it is widely used in different fields of science. Given X and Y the correlation coefficient is given as:

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{3.24}$$

where cov is the covariance function, and σ_X and σ_Y denote the standard deviation of X and Y respectively. The term $cov(X, Y)$ is given as:

$$\begin{aligned} \sigma_X &= \sqrt{E[X - E[X]^2]} \\ \sigma_Y &= \sqrt{E[Y - E[Y]^2]} \\ cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ cov(X, Y) &= E[XY] - E[X]E[Y] \end{aligned} \tag{3.25}$$

using this the value for the correlation ρ can be rewritten as:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X - E[X]^2]}\sqrt{E[Y - E[Y]^2]}} \tag{3.26}$$

Percentage Mean Square Error : One of the popular approaches to evaluate the HRTF modeling performance to use is the root-mean-square error used in [115]. In this method, the performance of the models for the magnitude modeling of HRTFs is evaluated by comparing the mean-square error of the disparity between the approximated and original HRTFs or HRIRs over the magnitude of the original HRTFs. Mathematically it is represented as:

$$e_j(\theta, \phi) = 100\% \times \frac{\left\| h_j(\theta, \phi) - \hat{h}(\theta, \phi) \right\|^2}{\left\| h_j(\theta, \phi) \right\|^2} \tag{3.27}$$

A small amount of this error will denote a greater matching, while a large error will mean that the performance of the model has deteriorated.

3.6.2 Subjective metrics

Although objective evaluation methods provide a quick way to test and evaluate the performance of the HRTF personalization methods by provide difference and matching measures, these methods do not provide perceptual ratings for the personalized HRTFs. For example, they fail to tell if these HRTFs are used for the listener in question, what kind of perception they will provide in terms of localization and externalization capabilities. The obvious way to test the HRTFs in this context is to perform the listening tests. Listening experiments are used to evaluate the performance of HRTFs in terms of localization in horizontal and sagittal plans for more than a few decades [97]. These tests can be of multiple types where listeners are provided with the virtual spatial audio, binaurally rendered through the HRTFs in question. The listeners are also presented with a reference sound either via actual speaker sources in different positions or virtual sources rendered through listeners' own HRTFs. The listeners are then asked to evaluate the performance of the synthesized or selected HRTFs vs. their HRTFs or vs. the anchor sound provided through the speaker sounds [30, 116]. Some test includes a graphical user interface (GUI) based reported system for listeners to report the results [102], while others use virtual reality and pointer-based reporting systems [117]. Furthermore, the questionnaires can include different kinds of questions such as evaluation of externalization, which can be either binary, i.e., if the sound is perceived inside the head or outside or can be a grading based score from one to five. Listeners report the perceived location of the sound, and the errors are calculated in terms of localization error or perceiving error (LE) or (PE), respectively. Furthermore, up-down and front-back reversals are also calculated [97, 117, 118] and reported as quadrant error. Although listening tests provide an interesting and perceptual way to evaluate the HRTFs, they are very difficult to conduct for multiple reasons. First of all, it requires human participants with some experience to perform the tests. It is reported in past studies the participants with less experience either required long training or their tests will have more uncertainties making the localization tests untrustworthy [119]. Furthermore, we need to have either the measured HRTFs of the listeners or sophisticated setup with an anechoic chamber where the anchor sounds can be presented to listeners like a reference.

To avoid all these problems, researchers came up with an interested computerized model that can provide a perceptual evaluation of the personalized HRTFs in sagittal planes for broadband sound signals [119]. This model works by creating a computerized peripheral system that processes sound in the band by band manner. Their model requires two inputs, namely, the HRTF to test and an uncertainty factor. The uncertainty factor counts for the contributions coming from non-acoustic factors, such as attention paid to the relevant cues, accuracy in responding, the ability to conduct scientific experiments, and the amount of training provided. In their later study [118], they suggested that these models require subject dependent uncertainty parameter, and if a right uncertainty parameter is provided, even a non-individualized HRTF can perform as good as the individualized one.

CHAPTER 4

HRTF Database Analysis and Personalization

This chapter provides details of two studies. The first study provides a simple statistical analysis on the notch frequencies of the median plane in two of the most famous public HRTF databases, namely the CIPIC [14] and SYMARE [51]. The deep notches are considered to be the most important cues for elevation perception in the median plane. There have been multiple studies to understand the underlying phenomenons which generate the notches, how these notches evolve and relate these notches to the morphology features in the ear shapes, [23, 44]. This study extracts the pinna contributions from the HRTFs of the median plane and using simple signal processing techniques to extract the notches from these transfer functions. It then uses the k-means clustering to statistically analyze how the notch frequencies of each of the databases evolve as a function of frequency. This study has two research questions. 1) Are the evolutions the same for both populations of users? 2) Is this evolution symmetric for the left and right ears? If not, are there some substantial binaural cues that can be used to localize the sound sources in the median plane? The results show that the average evolution of the notches for both databases is almost same, with first notch frequency starting from around 6 kHz and monotonically increasing to about 8.5 kHz, the second notch frequency starting from 10 kHz and monotonically increasing as a function of elevation to 12 kHz. At the same time, the third notch frequency starts from 13 kHz and keeps on increasing until 14 kHz around the horizontal plane as a function of elevation and then starts to decrease to reach a value of around 13 kHz at the plus 45°. These results are consistent with the previous exploratory studies performed on the CIPIC database [23, 102]. The results of the competitive analysis also happen to show that the binaural cues can be very useful for localization in the median plane. The difference between the notch frequencies is mapped to a psycho-perceptual unit. The results show that although the differences are not very high, there is a structure to these differences, and these are symmetric around the horizontal plane, suggesting that binaural cues can

help to localize the sound sources in median planes.

The second study presented in this work uses an existing HRTF personalization method based on the sparse representation [28] for the CIPIC database and modifies it to propose an improved version we call the weighted sparse representation. The research questions for this study are the following: 1) Knowing that different anthropometric parameters have different importance when it comes to generating the HRTF cues, can we calculate the important metric for these anthropometric parameters? In order to answer this question, we created a simple yet aggressive search algorithm that calculates relative importance relevance for the given set of anthropometric features in the CIPIC databases. 2) Can we use these important vectors as weights and propose a sparse representation based approach that uses these important factors as weights while calculating the sparse representation? To answer this question, we created a simple approach that uses these weights to calculate the sparse linear combination when modeling the anthropometric parameters of the query subject and represent. 3) The third question in this study is, can we reduce the required number of principal components and use only the easily gatherable parameters without compromising the performance? The results show that the HRTFs synthesized using our method are more accurate than the results obtained with the traditional sparse representation approach. 4) The final question posed in this work is that is there a dataset matching based personalization technique that can outperform our approach? The results show the performance of our systems is better than even from the case when the best HRTFs are matched in terms of smallest Spectral distortion, suggesting that no database matching scheme can outperform our approach [29].

Contribution: The main contributions of this chapter include:

1. A simple approach to understanding the evolution of notches in the median plane of PRTFs is provided in this chapter. The most prominent notches in all HRTFs for the left or right ears of all users in a database are grouped into three clusters using k-mean. These three groups present three main notches in rising due to three main contours in the ear shape, as reported by [23]. A comparative analysis for two databases and the results of [23], shows these findings are in line with each other.

2. Using a comparative analysis for both datasets between the left and right ear, it reports a novel concept that there are some binaural cues that can be used to localize the sound in the median plane. These cues are the differences in the notch frequencies for both left and right ear shapes.

3. A simple method to calculate the relevance of each of the anthropometric features in CIPIC features is proposed. This produces a relevance importance map suggesting which of the features are more useful when they are used for database matching based HRTF personalization.

4. The calculated feature importance vector is used as weights to propose a weighted sparse representation based HRTF personalization method. The results show that this method outperforms the pervious methods even when fewer anthropometric features are used.

5. A comparison of the performance between the popular database matching approaches, is performed suggesting that our weighted sparse representation approach outperforms all the available database matching approaches even when the best subject is chosen always (spectral distortion is used for evaluation).

The following sections provide details of these two studies describing the methodologies used and reporting the results. This chapter consists of two main sections. The Sec. 4.1 reports the details on the notch analysis, while in Sec. 4.2 the details on building of weighted sparse representation for HRTF personalization are provided. Finally Sec. 4.3 concludes the chapter.

4.1 Notch Analysis of HRTFs

One may recall from Ch. 2, that the spatial hearing is the result of the interaction between the sound waves and the listener's body before the audio signal reaches to the eardrums of the listeners. This interaction creates scattering, reflections, and diffractions of sound waves altering the spectral content of the sound waves in a direction and frequency-dependent way. The spectral coloration includes notches and peaks, providing cues to the listeners' brain enabling the listeners to have the ability of spatial hearing and sound localization in the median plane.

Although the shape of the head and torso also play a role; however, the deep spectral notches, which are the primary cues for providing the elevation sensation, are mainly contributed by the ears. These notches arise due to the interaction of the sound field with the intricate surface of outer ear shape and change as the elevation changes providing useful cues for elevation perception. This section contains the details on an analysis methodology that helps one to explore the relationship between notch frequencies and elevation angles in the median plane. In particular, the pinna contributions are extracted from the HRTFs and analyzed. The notches from the pinna related transfer functions (PRTFs) are extracted using a simple methodology for all the subjects and all the considered directions in the median plane $\frac{-pi}{4} \leq \phi \leq \frac{pi}{4}$. The extracted notch frequencies are then clustered using the k-means algorithm to study the statistics of the whole database, which reveals the relationship between notch frequencies and elevation angles. The results for SYMARE and CIPIC databases are provided. Furthermore, using comparative analysis, the values for left and right ear are compared in both databases suggesting the possible relevance of binaural cues for median plane localization. The details of the approach are provided below.

4.1.1 Analysis Methodology

This subsection provides the details on a statistical study conducted to understand the evolution of the notch frequencies with respect to elevation angles, for the median plane HRTFs of the user population in two of the most important HRTF databases. As it is believed that these notches are the results of the sound wave reflections from the pinna surface, the first step is to extract the pinna contributions from the HRTFs called pinna related transfer function (PRTF) by removing all the contributions due to the head and torso using the method described in [13]. Once we have the PRTFs in hand, the next step is to extract the notches using a simple method. These notches are then clustered using a K-means algorithm into three clusters. The motivation behind that

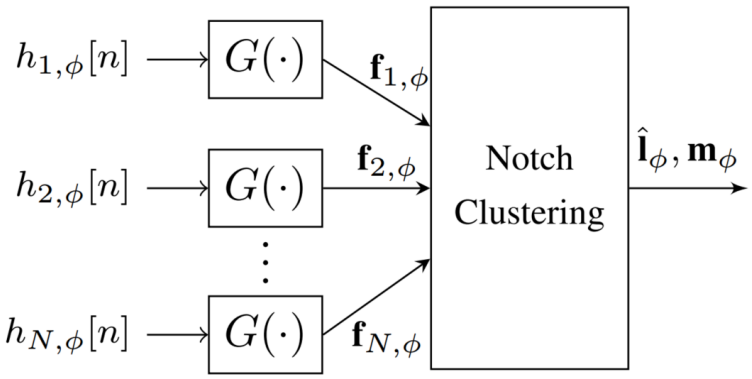


Figure 4.1: *Block diagram for the methodology. This block diagram shows the simple process of notch extraction happening in $G(.)$ and clustering of these notching using k-means to statistically analyse the notch frequencies.*

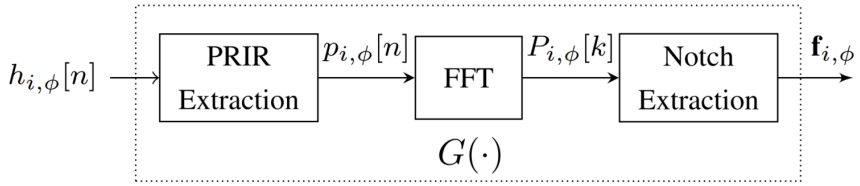


Figure 4.2: *An expanded version of $G(.)$ indicating how the PRTFs are extracted from the HRIRs and then used to extract the pinna notches $f_{i,\phi}$.*

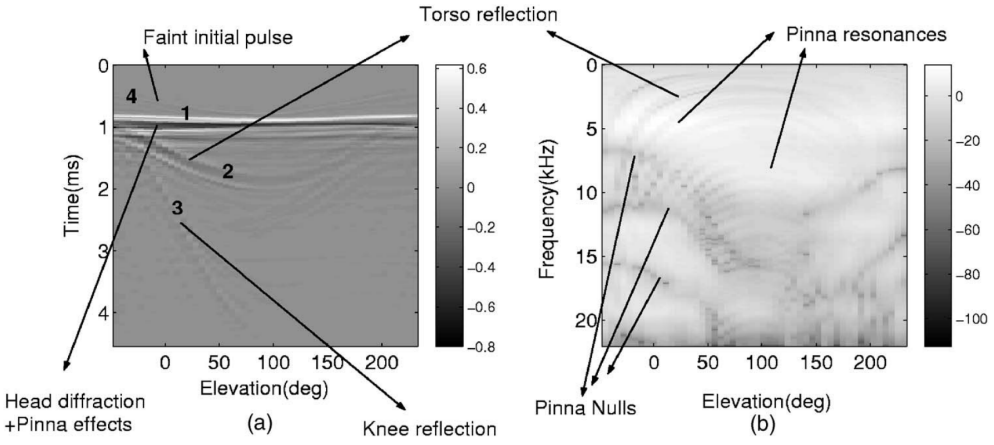


Figure 4.3: *Right ear median plan (a) HRIRs and (b) HRTFs are displayed as grey scale images for subject 10 in the CIPIC database. The directions are $\theta = 0°$ and $\phi \in [-45°, 230.625°]$. Different features in the HRIRs and HRTFs are marked as the contributions of different body parts. The scales for (a) and (b) are linear amplitude and dB scale log magnitude respectively. Image taken from [13].*

is that authors in [23] suggested that the three main notches in HRTFs are the results of the reflections of the sound waves from three main contours of the pinna shapes. Hence we created three clusters, only each corresponding to one of the contours. The centroids for these clusters are then analyzed to study the evolution of the notches for the median plane HRTFs as a function of elevation angles.

The overall methodology can be divided into three conceptual steps: including

- PRTF extraction,
- Notch frequency extraction, and
- Clustering and analysis of the notch frequencies.

The block diagram for the analysis methodology is provided in Fig. 4.1. Following the details for each of the steps is provided.

PRTF Extraction

The notches and peaks are one of the most important features in HRTFs. These notches are considered to be the most relevant cues for elevation perception [23]. It is long believed that these deep spectral notches are produced in HRTF due to reflections caused by different body parts, including pinna cavities, head, torso, and knees. The details of the contributions of different body parts in HRIRs and HRTFs are shown in Fig. 4.3. In this study we aim to analyze the spectral notches caused by pinna. So, the first logical step involved is of extracting the contributions of pinna from HRIRs by removing all unnecessary components which are not required for this study, namely the contributions of head, shoulders and knees. It was reported in [13] that the delays of pinna, torso and knee reflections are typically around 0.1, to 0.3, 1.6 and 3.2 ms [13, 120] respectively. Fig. 4.4 shows the simple process of PRIR extraction from HRIRs. To get rid of shoulders, torso, and knees reflection components, we shorten our HRIR by applying a half

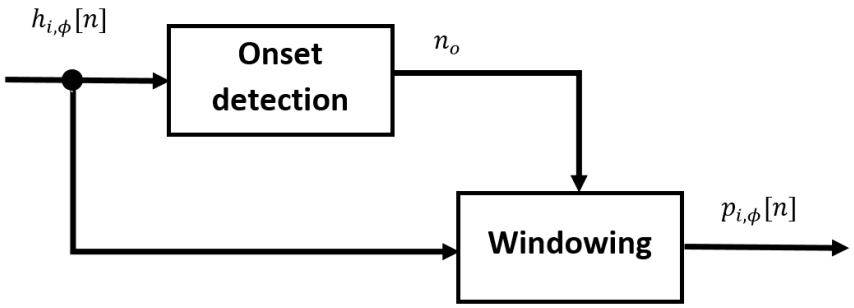


Figure 4.4: *Block diagram showing the process of PRIR extraction from HRIRs using the windowing process.*

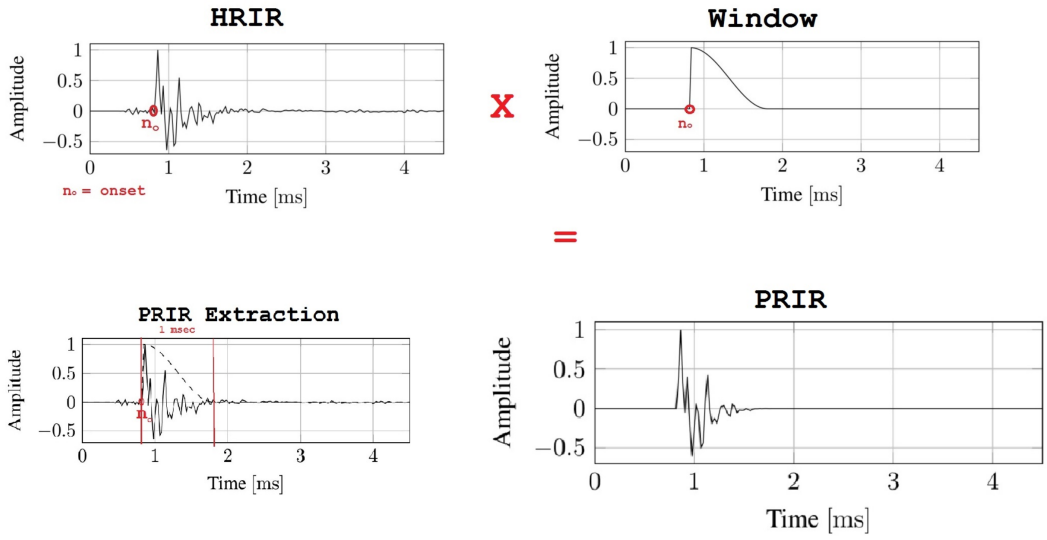


Figure 4.5: *Extracting the PRIRs from the HRIRs. The first step is to find the onset n_o. Knowing this a windowing operation is applied to remove the contributions of the shoulders, torso and knees.*

Hanning window. [13] suggested that a window of length 1 ms, starting from the onset of HRIR, is good for this task. This process removes the faint noise signal before the onset, along with the reflective components due to shoulders, torso, and knees, while preserving the reflection caused by pinna.

Given the HRIR $h_{i,\phi}[n]$ for the user i, and elevation angle ϕ, the PRIR $p_{i,\phi}[n]$ can be extracted by applying a half Hanning window $w[n]$ of length $1ms$, starting from the onset of HRIR n_o, i.e. $p_{i,\phi}[n] = h_{i,\phi}[n]w[n - n_o]$. Fig. 4.5 illustrates the windowing operation with the help of plots of the signal at each step. As HRIRs to be believed minimum phase filters with linear delay, the value of onset n_o can be found by taking the slope of unwrapped phase function of HRTF [13]. Once the PRIR $p_{i,\phi}[n]$ is obtained, the PRTF (Pinna related transfer function) $P_{i,\phi}[f]$ can be obtained by applying the Fourier transform, where f denotes the frequency of the signal. Next we describe the notch frequency extraction procedure from the PRTFs $P_{i,\phi}[f]$, $i = 1, 2, \cdots, N$ relative to all N users.

Notch Extraction

As highlighted before, the aim of this study to understand the importance of the notches in the median plane sound source localization and see how the notches evolve as the elevation angle changes. So we focused on the frequency ranges of the HRTFs, which are responsible for providing the elevation cues in the median plane. In [42] authors performed three experiments to understand which frequency contents in the HRTFs for median planes are responsible for source localization in the median plane, and reported that the frequency content in the range 4 kHz to 16 kHz provides the main cues for median plane localization. Knowing this, we restrict the frequency bandwidth of our analysis to this frequency range.

The notches of the PRIRs are extracted using a simple procedure. Having the PRIRs in hand, the PRTFs for all the subjects are computed by applying Fast Fourier Transform (FFT). We study these PRTFs in log or dB-scale. Instead of studying the positive log-scale HRTFs, this study works on the negative log-scale magnitude functions of the PRTFs, i.e.:

$$P_{i,\phi}[f] = -20\log_{10}(|P_{i,\phi}[f]|). \tag{4.1}$$

The purpose of studying the negative of the log magnitude functions instead of studying the positives is that this way, we turn the notches into peaks. Peaks can then be effectively extracted by finding the local maxima in $P_{i,\phi}[f]$. The steps are shown in Fig. 4.6. To get meaningful results, we also have to make sure that we are considering just the significant and prominent notches while discarding all those that are not relevant. Furthermore, we do not want to take two peaks which are very, very close to each other and have the same height. For this purpose, we consider the prominence of the local maxima as our deciding measure. The prominence describes how much the peak stands out from the neighboring peaks. For instance, a low isolated peak can be more prominent than one that is higher but is next to another higher peak and vice-versa.

In the following, we considered those peaks in $P_{i,\phi}[f]$ that have a prominence greater than $3dBs$. This results in a vector of notch frequencies for each subject i and elevation angle ϕ denoted as $f_{i,\phi}$ and has values as:

$$\mathbf{f}_{i,\phi} = [f_{i,\phi,1}, \cdots, f_{i,\phi,M}], \tag{4.2}$$

where $M_{i,\phi}$ denotes the number of relevant peaks (notches) in the negative log-scale PRTF of i^{th} user for elevation angle ϕ.

Once we have notch frequency vectors, $\mathbf{f}_{i,\phi} \in \mathbf{R}^{1\times M_{i,\phi}}$ for all the users and elevations, we arrange them into the vector f_ϕ, which contains the notch frequencies for all the users for a single elevation ϕ, i.e.:

$$\mathbf{f}_\phi = [\mathbf{f}_{1,\phi}, \mathbf{f}_{2,\phi}, \cdots, \mathbf{f}_{N,\phi}] \in \mathbf{R}^{1\times M_\phi}, \quad \text{with} \quad M_\phi = \sum_{i=1}^{N} M_{i,\phi} \tag{4.3}$$

Clustering the notches

The next step of the analysis is to drive the meaningful information from these frequency vectors $\mathbf{f}_\phi$. The findings of a recent study [23], reported that in each PRTF in CIPIC database up to three main spectral notches can be extracted, and mapped to three

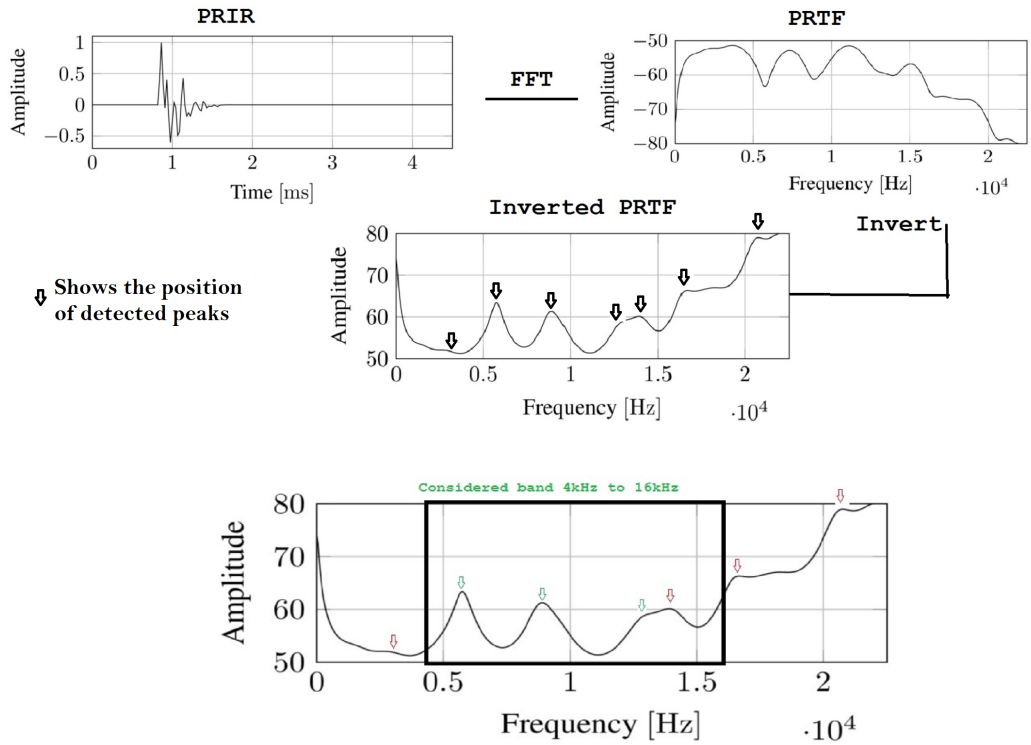


Figure 4.6: *Procedure for extracting notches and notch center frequencies from the PRIRs. The first step is to find the PRTFs using FFT. The notches are then found by taking the additive inverse of the log magnitude of the PRTFs and finding the peaks using simple local maxima finding functions. In the bottom plot the green pointers show the notches considered for this study while the red pointers show the ones ignored because they lie outside of the perceptually relevant frequency range of 4 to 16 kHz.*

distinctive and prominent pinna contours namely: the helix, anti-helix and outer wall of the concha or concha rim.

Based on these findings, this study clustered the notch frequencies in vector $\mathbf{f}\phi$ consisting of M_ϕ elements using K-means [121] into $K = 3$ clusters. Where, $M_\phi = \sum_{i=1}^{N} M_{i,\phi}$ and N denotes the number of users in the database. At the end of the process, each element in $\mathbf{f}_\phi$ will be assigned to a single cluster, whose centroid is the closest to the actual value of the element.

We evaluate the distance between each element $f_{i,\phi,j} \in \mathbf{f}_\phi$ and the corresponding centroid $m_{k,\phi}$ as the euclidean distance $D(f_{1,\phi,j}, m_{k,\phi}) = |f_{1,\phi,j} - m_{k,\phi}|$.

The K-means algorithm is initialized by assigning random values to the centroids $m_{k,\phi}$, $k = 1, 2, 3$. The algorithm is defined as an iterative two-step process. The details of each step are provided below.

Step1: The first step is the assignment of each notch frequency to a cluster having closest centroid and label it with the cluster id of that cluster e.g., $1, 2$, or 3. The cluster labels for the frequencies are found as:

$$\hat{l}_j = \underset{k}{\operatorname{argmin}}\{D(f_{i,\phi,j}, m_{k,\phi})\}, \tag{4.4}$$

where $j = 1, \cdots, M_\phi$ and $k = 1, 2, 3$. Moreover, a responsibility vector is defined for each cluster which tell which notches comes under which cluster as:

$$r_{k,j} = \begin{cases} 1, & \text{if } \hat{l}_j = k \\ o, & \text{otherwise.} \end{cases} \tag{4.5}$$

Step2: The second step is to update the centroid for all the clusters. Having the responsibility vectors in hand the new value for the k^{th} centroid is computed as:

$$m_{k,\phi} = \frac{\sum_{j=1}^{M} r_{k,j} f_{i,\phi,j}}{R_k}, \tag{4.6}$$

where R_k denotes the total responsibility of cluster k, and is equal to the number of data points falling into cluster k. This is mathematically given as:

$$R_k = \sum_{j=1}^{M} r_{k,j}. \tag{4.7}$$

These two steps keep on repeating until no further changes occur in the cluster centroids or the responsibility vectors do not change in consecutive iterations. After applying the K-means algorithm, we obtain the centroids $m_\phi = [m_{1,\phi}, m_{2,\phi}, m_{3,\phi}]$, corresponding to helix, anti-helix and outer wall of concha respectively. Moreover, to associate a relevance descriptor to the clustered data, we introduce the cluster spread as the standard deviation of their elements, i.e. :

$$\sigma_{k,\phi} = \sqrt{\frac{\sum_{j=1}^{M_\phi} (f_{i,\phi,j} - m_{k,\phi})}{R_k}} \tag{4.8}$$

4.1.2 Description of Databases

The analysis methodology described in Sec. 4.1.1 is applied to two of the most popular publicly available HRTF databases. The brief technical details about these databases

which are necessary to understand this work are given below.
CIPIC Database: CIPIC [14] is a public-domain database of acoustically measured HRIRs with a high spatial resolution. It contains HRIRs for 45 subjects (27 male, 16 female, and two KEMAR) measured at 1250 different directions around the head of the subjects. The measurements are done using Golay code as analysis signals, with a sampling frequency of $44.1kHz$. The loudspeakers are mounted on a circular arc of radius $1m$, which is rotated around a fixed listener. The length of each HRIR stored in the database is 200 samples. For this work, we consider all the HRIRs at azimuth $0°$ and elevations ϕ between $-45°$ and $45\circ$, with a uniform spacing equal to $5.625°$.
SYMARE Database: SYMARE [51] database was created by a collaborative team of Sydney University Australia and the University of York England. This database contains acoustically measured HRTFs for 61 users (45 males and 16 females) measured in 393 directions around the head at a distance of $1m$, with a non-uniform angular spacing in elevation for different azimuth angles. Impulse responses are recorded using Golay codes with a sampling frequency equal to $48kHz$. The length of each HRIR is 256 samples. For this work, we consider all the HRIRs at azimuth $0°$ and elevations ϕ between $-45°$ and $40°$.

4.1.3 Results

After the clustering has been performed on the notch frequencies of the median plane HRTFs in both databases, we performed two analysis studies. In the first study, we analyzed the notch frequencies and their evolutions for the HRTFs of both left and right ears separately and compared the results of both databases with each other. While in the second analysis, we compare the results of the left and right ear HRTFs. Following is a detailed discussion about the results.

Analysis 1
The HRIRs for the mentioned elevations were retrieved from both the databases, and PRIRs were extracted. The PRIRs were then transformed to PRTFs by performing a zero-padded 512-point FFT.

Notch vectors f_ϕ are estimated for each direction ϕ according to the angular grid adopted by the database, and notch frequencies are grouped into 3 clusters $m_{k,\phi}$, $k = 1, 2, 3$, along with their corresponding spread $\sigma_{k,\phi}$ using equation 4.8. The results are presented in figure 4.7. These graphs show the cluster centroids and spread as a function of the elevation angle ϕ for the left and right ears of all the subjects in CIPIC and SYMARE databases for three clusters. The x-axis has the elevation angle varying from $-45°$ to $45°$, while the y-axis shows the mean frequencies of the notch clusters, the range for the y-axis is set to $4 - 16kHz$, which is the frequency range under consideration for this study as is highlighted in the section while performing the notch extractions.

One thing to be noticed is that in general, all four plots (CIPIC and SYMARE databases, left and right ears) follow the same patterns. The other thing to notice is that at $\phi = -45°$ the value of the mean notch frequency is the same for all four plots for all three clusters. The centroids for three clusters for all cases lie around $6.5kHz$, $10kHz$, and $1.4kHz$, respectively. Another observation that we want to point out is that

all the cluster means $m_{k,\phi}$, $k = 1, 2$, exhibit a monotonically increasing behavior as a function of ϕ, despite some slight irregularities. These findings are consistent with the findings of previous studies, such as [13, 23, 42]. These irregularities are more prominent in the CIPIC database, especially for cluster 2. On the other hand, $m_{3,\phi}$ results to be almost constant in all the four considered cases. In a more general way, we observe that the slope of the clusters $m_{k,\phi}$, $k = 1, 2, 3$, is the highest for $m_{1,\phi}$ and almost zero for $m_{3,\phi}$. This behavior suggests that the pinna reflection causing a notch in the range of $m_{1,\phi}$ might be the most informative one for elevation perception in the median plane.

In the case of data extracted from the CIPIC database, we observe a peak around $\phi = 30°$ for the left ear, while the right ear exhibits a peak around $\phi = 40°$. However, in the SYMARE database, these irregularities are very mild and are present in just the right ear, while the evolution of the notch frequencies for the left ear is very smooth. Another thing worth noticing is that although we did not use any complex notch tracking scheme or method to group the notches as was used by [23] but still the notch evolution plots or notch tracks does not intersect or cross each other. Not just the tracks but also the spread for one track does not come into the range of the spread of the other track. In the next study, we compare the results for the left ear with the right ear.

Analysis 2
In the second analysis setup, we compare the results obtained for left ears with the ones obtained for the right ears for both databases. First, we convert the frequency centroids $m_{k,\phi}$ to the Bark scale [122] and then we compute their Euclidean distance between the centroids of one ear with the other. In the following we denote by $d_{k,\phi}$ the distance between the centroid of left and right ears for the k^{th} cluster and elevation ϕ. Results are reported in Figure 4.8.

It can be observed in both CIPIC and SYMARE, the maximum value for the distance between clusters is less than $0.5Bark$ for all the considered cases and for all the elevations. In the case of SYMARE, database distances have smaller values and a smoother distribution. In contrast, in the CIPIC database, distances are, in general, greater and less regular as a function of ϕ. We would like to point out that the differences exhibit minima in the horizontal plane ($\phi = 0°$) in all the considered cases and for all the clusters, suggesting that binaural cues are not relevant in the frontal image creation. On the other hand, it can be observed that the distances are greater moving away from the horizontal plane; this behavior suggests that both monaural and binaural cues are relevant for elevation perception in the median plane. This is something contrary to the original findings of many studies. Another possible reason for having these differences is that although the left and right ears of different subjects are almost similar, they are not always symmetric. These results show that the subjects in the SYMARE database has relatively better symmetry in their left and right ears and head shapes than CIPIC. Or the measurement facilities for SYMARE are more symmetric in terms of loudspeaker positions than CIPIC.

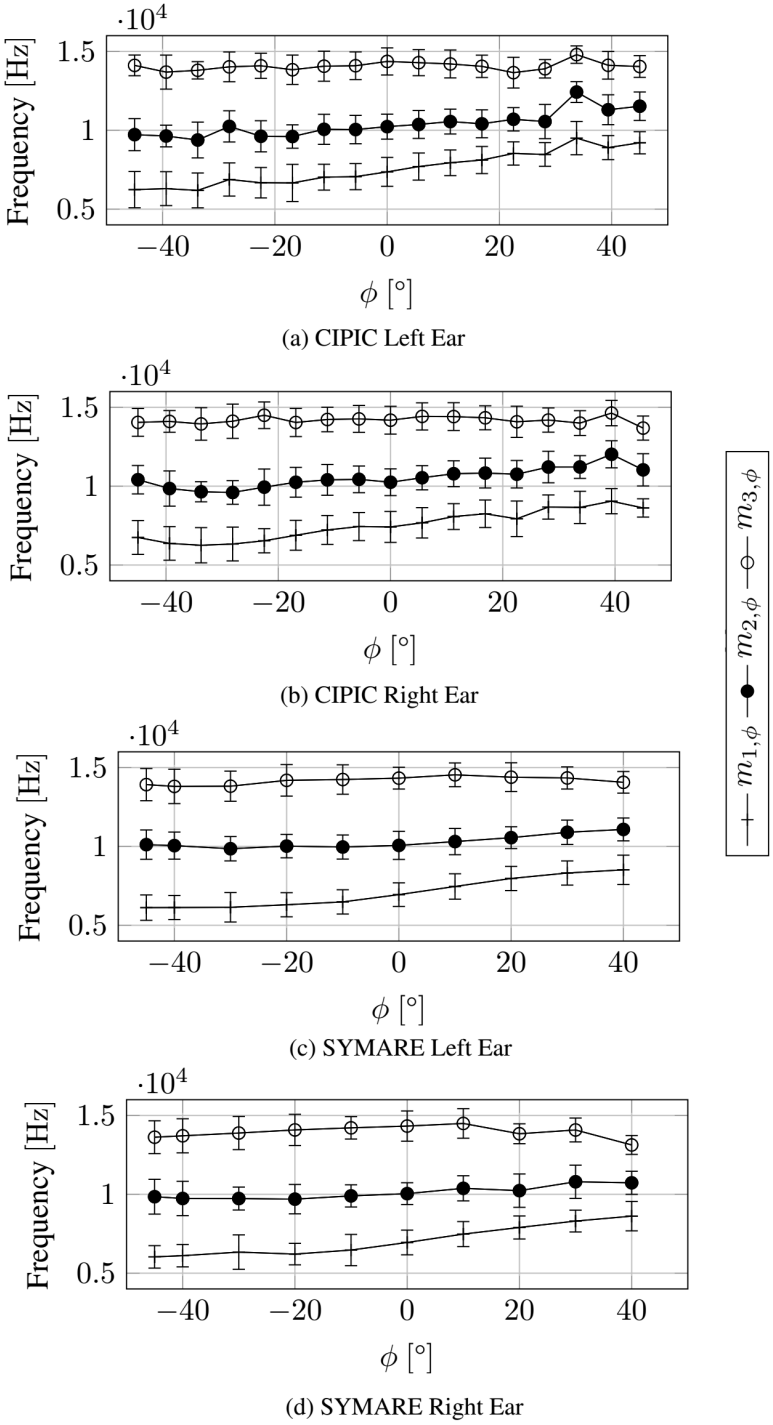
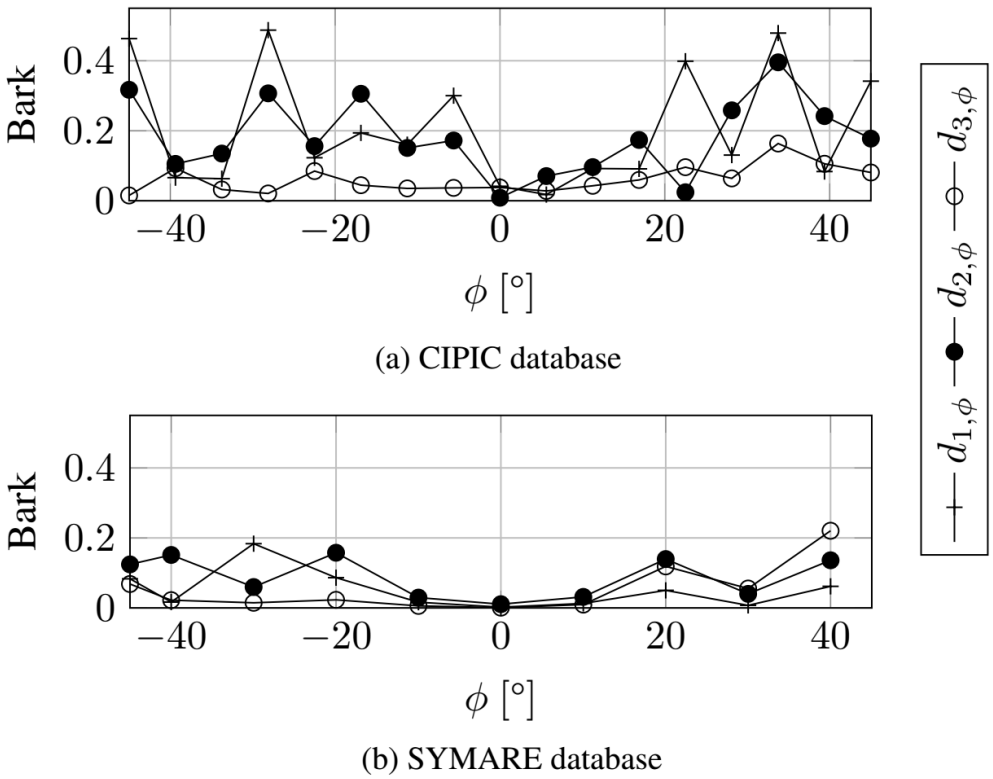


Figure 4.7: *Cluster centroids along with cluster standard deviation or spread as a function of the elevation angles in median plane HRTFs.*



(a) CIPIC database

(b) SYMARE database

Figure 4.8: *Distance between the centroids for the left and right ears as a function of elevations for median plane HRTFs.*

4.2 Weighted Sparse Representation

Recently a sparse representation based HRTF personalization method is proposed in [28]. This method treats the problem of HRTF personalization as finding the sparse representation. This approach applies an extreme assumption that the magnitude of an HRTF can be described by the same sparse linear representation as to the anthropometric features. Based on this assumption, HRTFs for a new subject, which is not present in the database, can be synthesized using the data in the database by sparse modeling. The results show that this method can improve the personalization of HRTFs and provide good results compared to database matching. Later [123], provided an overview of different post and preprocessing approaches, which can enhance the performance of the sparse representation based HRTF personalization.

Although these approaches [28, 123] provide a very nice performance, there is a problem with these approaches. All these approaches consider that all anthropometric parameters are equally important or relevant for HRTF personalization. However, past studies suggest that it is not the case. As Sec. 2.1.1, highlights the ears are more important for the complex feature generation in the HRTFs when compared to head and torso. This demands the inclusion of the relevance of each of the anthropometric parameters while calculating the sparse modeling. Following these lines, this study proposed an HRTF personalization method based on a weighted sparse representation of anthropo-

metric parameters. This also includes the findings of [123] and use the best pre- and postprocessing methods for sparse representation reported in this study. Another difference between our work and previously existing sparse representation based solutions is that we create a separate sparse model for both ears. The following are the details of this work.

4.2.1 Methodology

Sparse representation based HRTF personalization scheme treats the problem of magnitude response synthesis of the personalized HRTF by finding a sparse representation of the test subject's anthropometric features, i.e finding the linear combinations of the given anthropometric features which can generate the anthropometric features of the new subject. The approach is based on two strong assumptions:

- the HRTFs can be represented in the same sparse representation i.e., linear combinations as the anthropometric parameter vectors.
- The given training data set is rich enough to capture the anthropometric feature vector of any subject in the world.

Fig. 4.9 provides an abstract view on how this approach works.

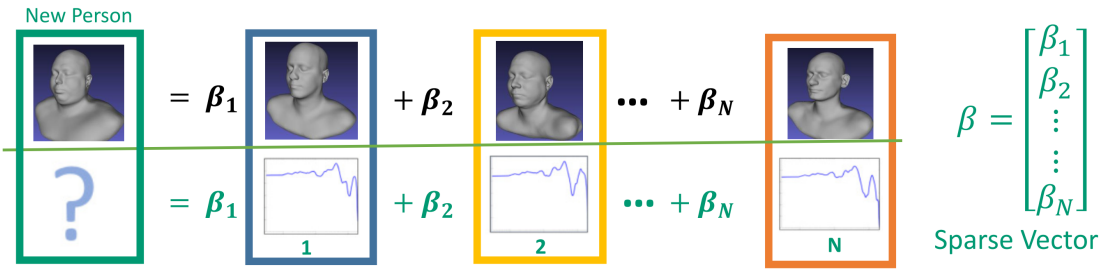


Figure 4.9: *The figure shows an HRTF personalization techniques based on sparse representation.*

In this work, we modify this simple approach to include the weights for every anthropometric feature and create a weighted-sparse representation based approach. Unlike the previous approaches which use a huge (96) number of anthropometric features and generate a single sparse representation for both ears, our approach uses fewer parameters. It also generates a separate sparse representation for both ears. The decision of using separate sparse representation for both ears is based on the findings made in Fig. 4.8, which shows that the HRTFs of left and right ears can be different and from the Fig. 4.10, which shows the mean of the absolute differences for few of the anthropometric features for ear shape. For being convinced to use the same sparse representation for both ears, the measurements for both ears must be the same, showing zero difference between the anthropometric parameters of the left and right ears. However, as can be seen from the presented figure, this is not the case. On average, the difference between left and right ear parameters can be as much as 10%, while the average of maximum differences is up to 30%. This suggests that different sparse representation solution for left and right ears is to be found and the HRTFs for right and left ears are different and not symmetric. Motivated from this, in this work, we used separate sparse representations for the magnitude synthesis of the left and right ears.

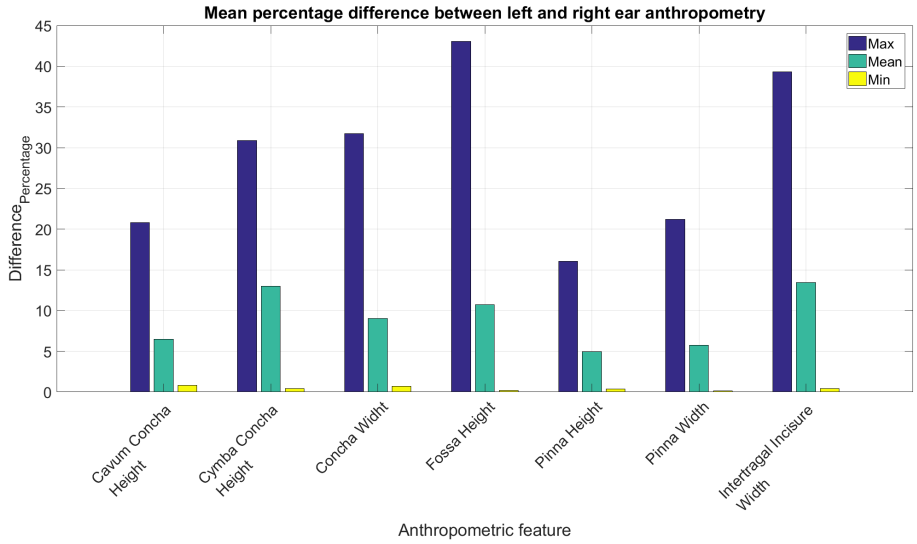


Figure 4.10: *The mean of the absolute difference between the anthropometric parameters of left and right ears of 36 subjects in the CIPIC database. This shows that the ear shapes are not symmetric and there is difference between the size of the ear shapes.*

The contributions of this work are: 1) Computation of the weights for different anthropometric parameters given in the CIPIC database, using a partially on-off method proposed in [15]. 2) Using these weights and proposing a weighted sparse representation. This approach, unlike the previous approaches based on sparse representation, considers the relevance of different anthropometric parameters through the found weights during the process of finding the sparse vector. 3) Providing an objective comparison of the results of the weighted sparse representation and different database matching approaches. Fig. 4.11 shows the block diagram of this work:

Database

All the experimentation conducted for this work is conducted on CIPIC database [14]. In the publicly available version, it contains acoustically measured HRIRs for 45 subjects for 1250 directions. In elevation plane it has a 50 angles starting from $-45°$ and going to $230.625°$, with a uniform step of $6.525°$. While the steps in azimuth plane are not uniform. The angles are coarser with larger steps towards the ears while are more dense with smaller steps in other directions. The 25 azimuthal angles available in the database are:

$$\theta = [-80, -65, -55, -45, -40, -35, \cdots, 35, 40, 45, 55, 65, 80]\,.$$

The database also comes with 27 anthropometric measurements of the subjects as shown in Fig. 4.13. Although the database contains the HRIRs for 45 subjects it only has the anthropometric data for 35 subjects. This study used only these subjects.

Anthropometric Feature Selection

Following the work of [15], this work suggests using only the 17 easily gatherable anthropometric features out of the 27 anthropometric parameters provided in CIPIC (12 for the head and torso and 7 for the pinna). Fig. 4.13 illustrates a simple setup to

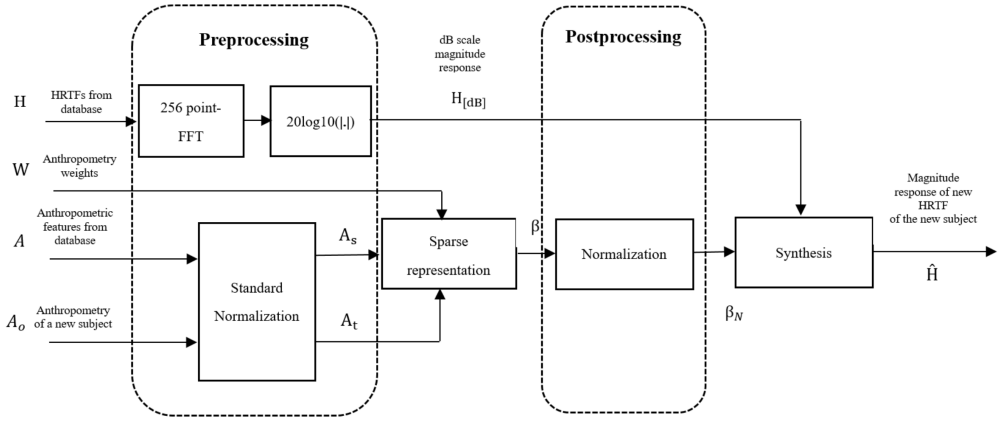


Figure 4.11: *Block Diagram of HRTF Personalization using weighted sparse representation of anthropometric features.*

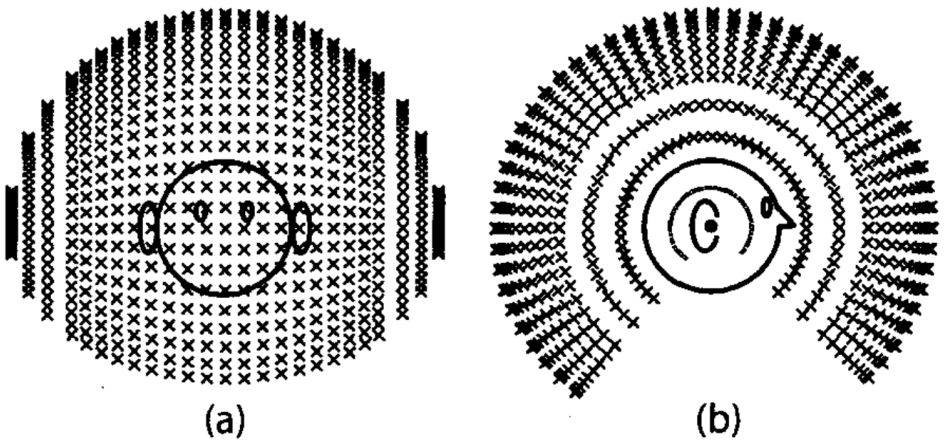


Figure 4.12: *Measurement points are relative to the head of the listener in the CIPIC database. Image taken from [14]*

Symbol	Anthropometric Feature	Symbol	Anthropometric Feature
x_1	head width	d_1	cavum concha height
x_2	head height	d_2	cymba concha height
x_3	head depth	d_3	concha width
x_4	pinna offset down	d_4	fossa height
x_5	pinna offset back	d_5	pinna height
x_6	neck width	d_6	pinna width
x_7	neck height	d_7	intertragal incisure width
x_8	neck depth		
x_9	torso top width		
x_{10}	torso top height		
x_{11}	torso top depth		
x_{12}	shoulder width		

Table 4.1: *Easily gatherable anthropometric features. According to the studies presented in [15] these* 19 *anthropometric features can be measured from three scaled pictures.*

acquire these anthropometric parameters using three scaled images. These parameters, which can be obtained from three images, are listed in Tab. 4.1. Note that this approach results in 19 features, but x_5 (pinna offset back) is not easy to measure as it highly depends on the flare angle of the pinna, and x_7 (neck height) strongly depends on the posture of the subject while the photograph is captured. Hence we ignore these two features out of 19 and use only the remaining 17 features for our study. Further details on acquiring anthropometric parameters from these three profiles are out of the scope of this work and can be found in [15].

Calculation of weights for anthropometric features

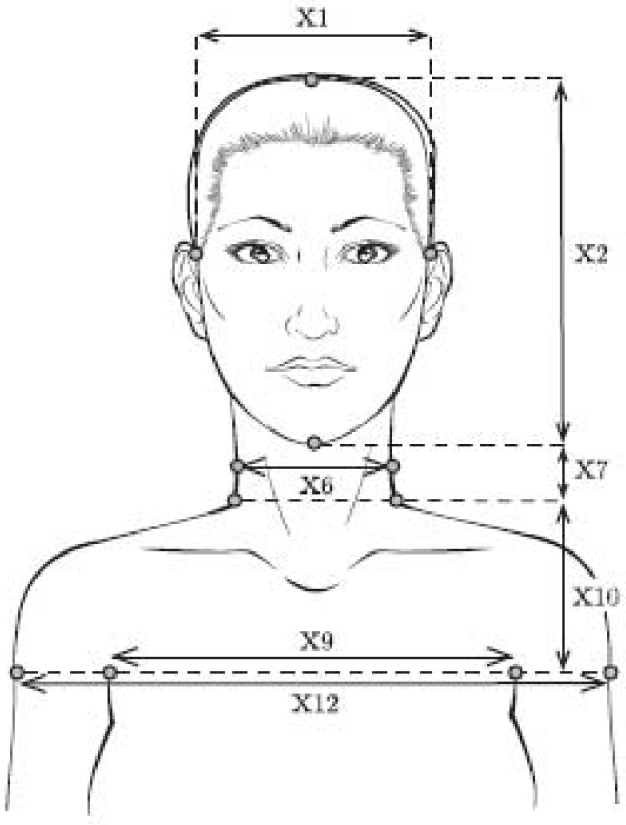
This section highlights how the weights for the anthropometric parameters are calculated. This work calculates the relevance metric or weights for given anthropometric parameters using the approach described in [15]. The process of weight calculation is shown in the block diagram presented in Fig. 4.14.

Using the anthropometric parameters in their original form is not very meaningful as different anthropometric features lie in different ranges and scales. For example, the height of the head is much larger than the height of the cavum concha. Hence to make these features meaningful for calculating the sparse representation, these features are normalized. This normalization process brings all the values on a common scale. This work uses min-max normalization to do so. For a given vector of random values x the min-max normalized vector x^N is determined as:

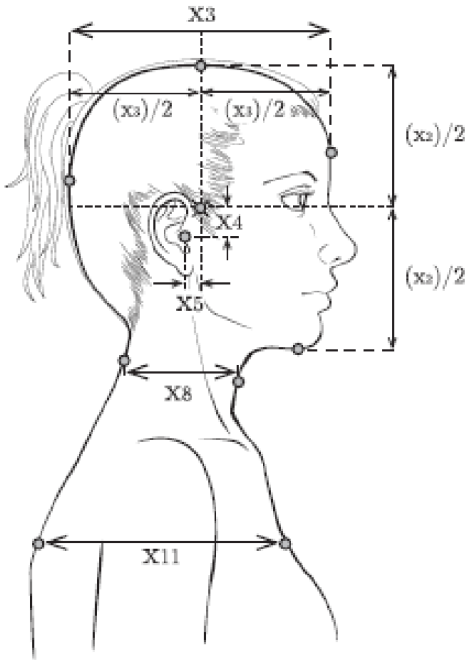
$$x^N = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{4.9}$$

All the min-max normalized anthropometric parameters are arrange in an anthropometry matrix A, for all the listeners as:

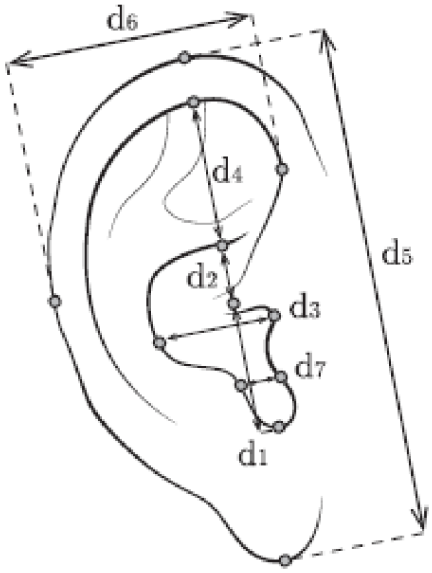
$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,25} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,25} \\ \vdots & \vdots & \ddots & \vdots \\ a_{35,1} & a_{35,2} & \cdots & a_{35,25} \end{bmatrix} \tag{4.10}$$



(a) *Front View*



(b) *Side View*



(c) *Pinna View*

Figure 4.13: *Set of anthropometric features that can be measured using three images. The above figures show a) Front view, b) side view and c) zoomed in pinna view. Image reprinted from [15]*

where, $a_{i,j}$ denotes the j^{th} anthropometric parameter for i^{th} listener. Using the min-max normalization method provided in equation 4.9, the i^{th} column in A denoted by $A_{\{,i\}}$ is normalized as:

$$A^N_{,i} = \frac{A_{\{,i\}} - \min(A_{\{,i\}})}{\max(A_{\{,i\}}) - \min(A_{\{,i\}})}, \tag{4.11}$$

where $A^N_{\{,i\}}$ denotes the normalized column vector containing the normalized i^{th} anthropometric parameter for all listeners. The normalized anthropometric parameter matrix is denoted as A^N.

We assume that if an anthropometric parameter is relevant to the HRTF personalization process, two listeners having the same or closely matching values for that particular anthropometric parameter must have relatively good agreement between their HRTFs as well. To put this theory to test, this work uses an extensive search for relevance metrics based on a partially on-off approach. This means that a set of $2^{25} - 1$ different combinations is tried, where at any given time, different anthropometric parameters are controlled by 25 on-off switches. The reason for subtracting 1 is the exclusion of the case where all parameters are off. A mismatch matrix M is computed for each of these combinations, which contains the difference between the anthropometric parameters from the anthropometric parameters of all other listeners. The k^{th} iteration of M is calculated as:

$$M^{(i,j,k)} = \left\| \sum_{a=1}^{25} (A^N_{i,a,k} - A^N_{j,a,k}) \right\| \quad \forall k = 1, 2, \cdots, 2^{25} - 1 \tag{4.12}$$

where $M^{(i,j,k)}$ corresponds to the mismatch between anthropometric parameters of i^th and j^{th} listener in CIPIC in the k^{th} partial on-off combination. In total there will be $2^{25} - 1$ iterations for this process each resulting in a matrix of size 35×35.

Next, we calculated the global average spectral distortion (GASD) matrix of HRTFs, from each listener to all other listeners as:

$$SD(H_i, H_j) = \sqrt{\frac{1}{D}\frac{1}{F}\sum_{d=1}^{D}\sum_{f=1}^{F}\left(20\log_{10}\frac{\left\|H_{(i,d)}(f)\right\|}{\left\|H_{(j,d)}(f)\right\|}\right)^2}, \tag{4.13}$$

where $H_{(i,d)}(f)$ and $H_{(j,d)}(f)$ correspond to the HRTF of i^{th} and the j^{th} subject in direction d. F denotes the number of frequency bins. As we performed 256 point FFT, the value of F in our case is 128. D is number of directions for which HRTFs are available for and is equal to 1250. As we have 35 listeners $GASD$ matrix will be of size 35×35 matrix of SD.

To see which combination is good for a given subject, we used Pearson's correlation between the $GASD$ and M for all possible $2^{25} - 1$ combinations of partial on-off anthropometric parameters as follows:

$$\rho^{(i,k)} = corr\left(M^{(i,.,k)}, SD^{i\times 35}\right), \tag{4.14}$$

where $\rho^{(i,k)}$ corresponds to the Pearson's correlation coefficient of the i^{th} subject in the k^{th} combination, and $M^{(i,.,k)}$ represents a column vector of the anthropometric distance

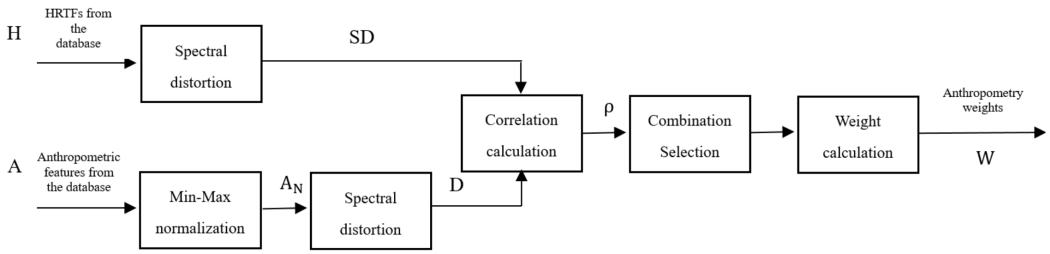


Figure 4.14: *Block Diagram for weight calculation procedure.*

between subject i and all other subjects for $k-th$ combination. We are looking for a combination that results in a maximum value for this correlation. This combination is selected as the best combination for the given subject i. Once we have the best combination for all the subjects, the weight vector can be calculated by dividing the number of occurrences of any given anthropometric parameter in different cases with the total number of combinations which is 35 as we have 35 subjects:

$$W^{(n)} = \frac{t^{(n)}}{35}, \tag{4.15}$$

where $W^{(n)}$ corresponds to weight of n^{th} anthropometric feature. t_n is number of times of n^{th} anthropometric parameter appeared in the best combinations for different subjects.

4.2.2 Preprocessing for anthropometric features and HRTFs for Sparse Representation

A recent study [15], investigated the effects of different preprocessing and postprocessing methods on the performance of sparse representation based HRTF personalization. The results suggest that using the standard normalization for the anthropometric parameters results in a better performance. Consider all the anthropometric parameters for the listeners in the dataset are given by A as shown in Eq. 4.10, and the anthropometric features for a new user are given as A_o, a row vector. The superset of the anthropometric parameters is created by concatenating the matrix A with vector A_o as:

$$A_d = \begin{bmatrix} A \\ A_o \end{bmatrix} \tag{4.16}$$

The standard normalized anthropometric parameter matrix A_s and standard normalized test listeners anthropometric parameter vector A_t are then found as:

$$\begin{aligned} A_s &= \frac{A - \text{mean}[A_d]}{\text{std}[A_d]}, \quad \& \\ A_t &= \frac{A_\circ - \text{mean}[A_d]}{\text{std}[A_d]}. \end{aligned} \tag{4.17}$$

The preprocessing study also suggested that for acoustic data when the dB-scale magnitude responses of HRTFs are used instead of using the time domain HRIRs, the performance is better. So the HRTFs from the HRIR data are calculated applying 256 points Fast Fourier Transform (FFT) and dB-scale magnitudes are then obtained as:

$$H_{[dB]} = 20 \log_{10} |H|, \tag{4.18}$$

where $|H|$, denotes the magnitude response of the HRTF H.

4.2.3 Sparse representation of anthropometric features

Having the normalized anthropometric parameter matrix of the training set and test listener, the sparse representation based modeling of test subject's anthropometry is obtained as:

$$A_t \approx A_s^t \beta \tag{4.19}$$

where A_s is the standard score of the anthropometric parameters A in the database.

In the sparse vector $\beta = [\beta_1, \beta_2, \cdots, \beta_S]^T$, each element corresponds to the weight of a subject in the linear superposition, where S is the number of subjects in training set. Thus, the problem of looking for an optimal sparse vector can be considered as a minimization problem:

$$\beta = \underset{\beta}{\operatorname{argmin}} \left(\left\| (A_t - A_s^T \beta)^T W \right\|_2^2 + \lambda \|\beta\| \right), \quad \textbf{s.t.} \quad \beta^{(i)} \geq 0, \tag{4.20}$$

where W represents the weights of different anthropometric parameters. In line with [15], we added a non-negative constraint on β e.g. $\beta_i \geq 0$. Where the regularization parameter λ of this minimization problem is a non-negative parameter.

To ensure that the synthesizing process has consistent amplitudes at the output, as in the database, we normalized the values of β vector such that the sum of the β vector is equal to 1, such that:

$$\beta_N = \frac{\beta}{\sum_{s=1}^{25} \beta^{(s)}}. \tag{4.21}$$

This changes the process of HRTF synthesis to a weighting average.

4.2.4 HRTF Synthesis

Once normalized sparse vector β_m is obtained, this can directly be applied to the log-scale HRTF data $H_{[dB]}$ in the database:

$$\hat{H}_{[dB]} = \beta_N H_{[dB]}. \tag{4.22}$$

However, the new synthesized HRTF $\hat{H}_{[dB]}$ is expressed in dBs. The magnitude response of the HRTFs is found as:

$$\hat{H} = 10^{\frac{\hat{H}_{[dB]}}{20}}. \tag{4.23}$$

4.2.5 Experiments

The performance of the proposed approach is evaluated by applying the "leave one out cross-validation" approach proposed in [124]. Each of the 35 subjects is taken out one-by-one as the test subject, and the remaining subset is considered as the training set.

Sparse representation using	Left Ear	Right Ear	Average
17 parameters (weighted)	5.5235	5.5351	5.5293
17 parameters (traditional)	5.6298	5.6359	5.6328
27 parameters (traditional)	5.5770	5.5705	5.5738

Table 4.2: *Results of sparse representation based HRTF personalization methods for traditional and weighted cases. This table shows that weighted sparse representation provides a better performance compared to the traditional sparse representation even when fewer number of anthropometric parameters are used.*

Evaluation Criteria

For evaluating the difference between synthesized HRTFs $\hat{H}$ and the original HRTFs H of the test subject, we employed a widely used error metric spectral distortion as our evaluation criteria.

$$SD^{(d)}(H,,\hat{H}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(20\log_{10}\left\|\frac{H^{(d)(f_n)}}{\hat{H}^{(d)}(f_n)}\right\|\right)^2} \tag{4.24}$$

where $H^{(d)}$ is the original HRTF in the d-th direction, and $\hat{H}^{(d)}$ is the synthesized HRTF in same direction. N is the number of frequency bins ($N = 128$). The SD for all directions is combined as:

$$SD(H,\hat{H}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(SD^(d)(H,\hat{H}))^2} \quad [dB], \tag{4.25}$$

where D is 1250.

Results and discussion

The results of the approach are presented in Tab 4.2, and Tab. 4.3. Tab 4.2 compares the results of weighted sparse representation based approach with the non-weighted one. The results show that even when more anthropometric parameters are used traditional approach can not beat the weighted one.

This study also compared the results of the proposed approach with few of the most popular database matching personalization methods proposed in [15,102,125]. Furthermore, two baselines are created to understand the performance of database matching method by creating "Best-baseline", when the selected HRTF and the individualized HRTF has the least amount of disagreement in terms of SD and the "Worst-baseline", when the selected HRTF and the individualized HRTF has the largest error. These two baselines define upper and lower limits for any closest matching algorithm. The results are reported in Tab. 4.3. These results show that none of the closes matching based approach can outperform the proposed approach.

4.3 Chapter Conlusion

The first study presented in this chapter has provided an analysis methodology to analyze the evolution of the notch frequencies in the median plane HRTFs in two publicly

HRTF Personalization Method	Left Ear	Right Ear	Average
Weighted Sparse representation with 17	5.5235	5.5351	5.5293
Closes-match based on Pinna Contours [102]	7.3403	7.3403	7.3403
Closest-match based on anthropometry and PCA [125]	7.6287	7.1844	7.4065
Closest-match based on weighted anthropometry [15]	7.5451	7.2239	7.3845
Closest-match "Best" baseline	6.2306	6.0317	6.1311
Closest-match "Worst" baseline	9.5628	9.0821	9.3324

Table 4.3: *Comparison of the results for weighted sparse representation and some of the most popular closest matching based HRTF personalization techniques. The results show that even for best baseline, closest match find can not beat the performance of our work.*

available databases. In particular, this work describes a technique which extracts the pinna related notches and their center frequencies from the HRTF data and classifies them into three clusters. These three clusters correspond to specific contours in the pinna, namely the helix, antihelix, and outer wall of the concha, respectively. The results of our approach are validated the proposed with acoustically measured HRTFs from the CIPIC and SYMARE databases. We also provide a comparative analysis of the evolution of notch frequencies in the median plane in CIPIC and SYMARE databases showing the results for both databases side by side. Although the spatial resolution in the median plane is different for both databases, the results somehow are still identical. To be more specific, the results show a strong dependency of the notches in the HRTFs on the elevation angles in the median plane. Moreover, we also studied the clusters in both databases binaurally by analyzing the differences between the mean of the notch frequencies for both ears. The results of this analysis revealed that not only monaural but also binaural cues are essential for elevation perception.

Furthermore, this chapter calculates the importance metric for each of the anthropometric parameters, which provides a measure of their relevance in the HRTF personalization method. The calculated importance vector is then used to create a weighted sparse representation based personalization method for HRTF magnitude responses. Unlike previous studies, this method requires only 17 anthropometric parameters, which can all be gathered from three scaled pictures. The evaluation of the approach shows that this approach outperforms the previous approaches of this kind. The results are also compared with the existing closest matching based personalization solutions and suggest that the proposed method can beat any closest matching based personalization method.

CHAPTER 5

Studying the Morphoacoustic of Affine Transformations on Ear Shapes

This chapter investigates if the geometric shape variations in the ear shapes and their corresponding acoustics can be studied independently of the scale, rotation and translation variations. In order to perform this study a unique database of ear shapes is created out of SYMARE database. In this synthetic database all the ear shapes are affine matched using LDDMM framework to the multi-scale template ear shape calculated using [17] to have same size, orientation and position. This was the core assumption of [17] for using the affine matched ear shapes for the creation of morphable models. In this work we investigate this hypothesis and propose a simple and complete 3D model with head, ear and torso shapes, in which the ear shapes for individuals has to have correct shape but are rigid matched to the template ear while the head and torso shapes of the template are used. We call these 3D models as affine matched models or affine models. The acoustic transfer functions for the affine models are then computed using FM-BEM simulations. The detailed and step by step procedure for these BEM simulations is provided in Sec. 2.3. To the best of the authors knowledge, this dataset is a unique dataset, one of its own kind created using powerful LDDMM framework to study the variations in scale, rotation, translation and geometry of the ear shapes. An outcome of this study and one of the main contribution of this chapter is the this dataset.

The study presented in section 7.2 of Ph.D. thesis [1] could be considered similar to this study, but our study differs from that study as that study simply used the affine matched ear for the morphable model creation and did not provide any analysis on the compensation/correction procedure for these affine transformations. Also, although they studied the acoustic simplifications introduced due to affine matching the ear shapes, they did that only for six subjects. Another difference is that they analyzed the differences between only the ear shapes without considering the differences that can

come from different head and torso shapes. author in that study did not analyzed the corrections for procedure neither did performed the study for the full database.

Using the synthetic database of affine models and their corresponding acoustics, this chapter tries to answer following research questions:

- How the acoustic transfer functions of the affine matched ear shapes vary compare to the acoustic transfer functions of the original ears? If there have been some simplifications introduced, can we quantify these simplifications?
- Can the introduced simplification, which are the artifacts in the individualized HRTFs, be corrected or compensated for, through simple frequency scaling (as proposed by [25]) and rotations of HRTFs?
- If yes, then how to find an optimal scaling for aligning the features of the HRTFs of affine matched ear shapes for low frequency range (head contributions) and for high frequency range (ear contributions)?
- How the obtained optimal scaling factors relate to the physical scaling factors to propose a simple scale and rotation correction approach for future studies, where optimal scaling factor can be derived simply from head and ear scaling factors?

The results of the studies conducted in this work suggest that the affine matchings of the ear shape and using same head and torso shapes reduces the inter-subject variations in the SYMARE population by almost 10%. Furthermore, the size and rotation angles of the ear and head shapes have mean values which are very close to the values for template. The analysis on corrections shows that the simple corrections such as frequency scale corrections and rotation of the HRTF acoustic surfaces can result in significant improvements in the matching of the HRTFs of the affine matched ear shapes with the ones for original ear shapes. Finally an optimal scaling factor search method is proposed which provides three optimal scaling factors for frequency axis, which best align the HRTF features for:
1) the whole analyzed frequency range, i.e. from 0.2 to 17 kHz,
2) for the head contributed frequency range i.e. upto 5 kHz, and,
3) for ear contributed frequency range for 5-17 kHz.
Finally a simple regression based mapping between the optimal scales and the physical scales shows that all these three scales can be approximated using the physical head and ear scales.

The rest of the chapter is arranged as follows: Sec. 5.1 provides a detailed review of the preliminary studies for this conducted in [1]. This background includes on creating the affine model, extracting and analyzing the scale and rotation information from the affine transformation matrices, and analyzing how template sits in the SYMARE database population, and quantifying the acoustic simplicities introduced by affine transformations for six subjects. Sec. 5.2 provides a view on the acoustic simplifications achieved for the full database when the HRTFs of the affine models are studied instead of the HRTFs of original 3D models. Sec. 5.3 provides details on a simple yet effective method of compensations and corrections to estimate the individualized HRTF directivity patterns of the original head, torso and ear shapes from the affine model acoustic data using simply scale factor and rotation matrices of the ear shapes. Finally, this section discusses the process of selecting an optimal scale and an

analysis of obtaining this scale factor from simple anthropometric measures such as head and ear scale factors is provided.

5.1 Background

The inspiration for this work comes from the work presented in [25, 98] originally. In [25, 98] the investigative studies were proposed which suggest that the differences amongst the individualized HRTFs can be reduced by simply compensating the scale differences in the ear shapes. Relying on these studies the authors in [1, 17, 32] used the affine matched ear shapes for the creation of the morphable model for the ear shapes instead of using the original ears. Another motivation for them to prefer the affine matched ear shapes over the original ear shapes was to simplify the modeling process for LDDMM and KPCA limiting the variations in the ear shapes only to the shape variations excluding the size, orientation and translation.

Following these lines authors in [1] affine matched all the ear shapes in the SYMARE database using the method provided in 2.8. The following section provides some insights to the results of different analyses performed by the past thesis study to compare the size and rotation of the created template to the other ears in the SYMARE population.

5.1.1 Scale Factors and Rotation Angles for Ear Shapes

The scaling factor and rotation angles from the template shape to given ear shape for subject l to template ear shape can be computed by decomposing the affine transformation matrix M_l, which contains the information for rotation and scaling to match the two shapes. The detailed process on the extraction of the scale and rotation information is provided in Sec. 2.3.1. Here we only report the statistics for the scale factor and angles. The results are given in the Fig. 5.1.

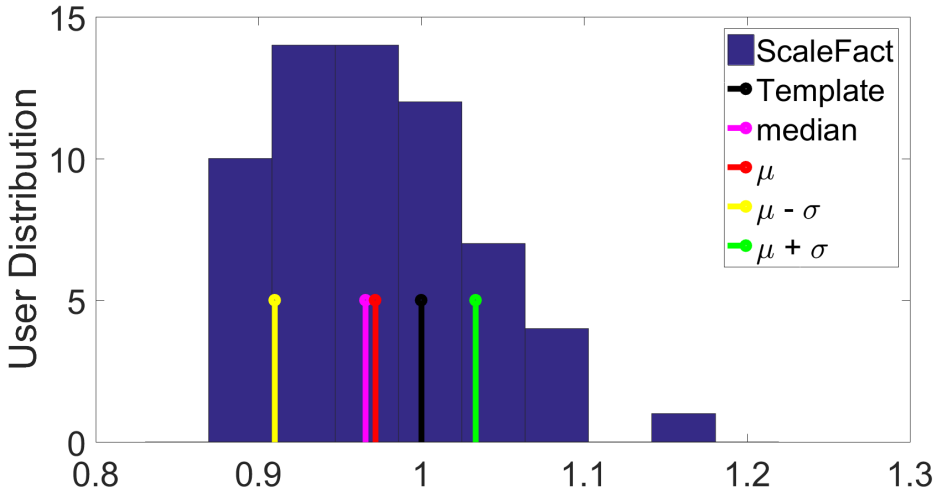


Figure 5.1: *Histogram for the scale factor from the template ear to 62 ear shapes.*

Note that, as per the results presented in this figure, 62% of the subjects are smaller than template ear while 38% are larger. The mean and median of scale factors are 0.97

and 0.96 respectively, both of which are very close to unity. This shows that template shape indeed is a good representative of the population.

The stats for the rotation information are presented in Fig. 5.2. Again these stats show that the mean and median value for the angles θ_X, θ_Y, and θ_Z are close to 0°, which shows template ear shapes orientation represents a nice representative of the average of the ear shape population. These scale factors and rotation angles are used to correct and compensate for the effects of affine transformations and to retrieve the HRTF directivity patterns of original shapes.

5.1.2 Scale Factor for Head Shapes

Head shapes are much simpler compared to the ear shapes, so the scale factors for the head shapes was directly calculated by measuring the head width, depth and height. To do so author of [1] created a set of landmark points on the head, torso and ear models as were created by [103]. These landmark points are shown in Fig. 5.3.

The three dimensions for the head, namely, head-height, head-width, and head-depth, can be calculated using these points. Where the head-height, head-width and head-depth are given by the Euclidean distances between P_{28} & P_{30}, P_2 & P_{15} and P_{27} & P_{31} respectively. The scale factor for these three measures are then calculated by taking the ratio of the measures of the subjects with the measures of the template ear shape, as stated below in Eq. 5.1:

$$\begin{aligned} SF^l_{H_H} &= \sqrt{\frac{||P^l_{28} - P^l_{30}||^2}{||P^{T_{MS}}_{28} - P^{T_{MS}}_{30}||^2}} \\ SF^l_{H_H} &= \sqrt{\frac{||P^l_{28} - P^l_{30}||^2}{||P^{T_{MS}}_{28} - P^{T_{MS}}_{30}||^2}} \\ SF^l_{H_H} &= \sqrt{\frac{||P^l_{28} - P^l_{30}||^2}{||P^{T_{MS}}_{28} - P^{T_{MS}}_{30}||^2}} \end{aligned} \tag{5.1}$$

where superscript l denotes the subject id. The distribution of three scale factor for 62 subjects are given in Fig. 5.4. These figures show that width and height of the template head are very close to the average of head width and height of the population while the depth of template head is not average. In fact it is 10% smaller than the average of the population.

The analysis provided in these studies provide us with a deep insight on the physical sizes of template with respect to the sizes of the population, it would be really nice if we could have a single scale factor for the head shape instead of three. To do so in this study we rely on the findings of a well known study [126], which models the head shape using a simple sphere shape. This study performs some experiments and suggest that the effective radius of the head can be calculated using three head measures, namely head-width, head-depth and head-height through an expression as follows:

$$r_{eff} = 0.51\frac{H_W}{2} + 0.019\frac{H_H}{2} + 0.18\frac{H_D}{2} + 3.2(cm) \tag{5.2}$$

In order to find a single scale for the head size we calculate the effective radius of all the subjects and the template head and then calculate the scale factor of the head as a

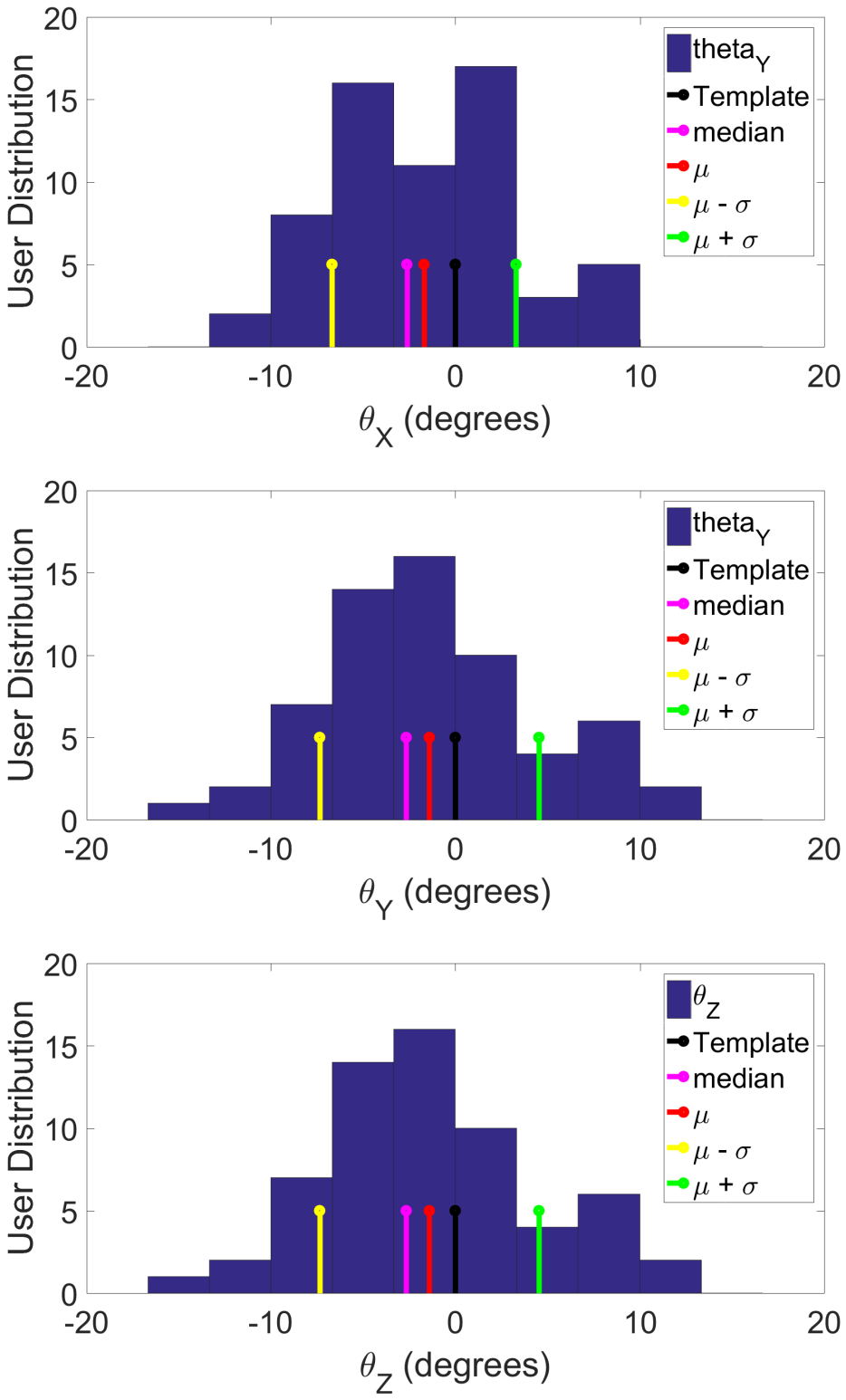


Figure 5.2: *Histogram for the Tait Bryan angles θ_x, θ_y, and θ_z. Note that these angles are computed according to z-axis, then y-axis, and then z-axis.*

ratio of the effective radius of the subjects and radius of the template.

$$SF_H = \frac{r_{eff}^l}{r_{eff}^T} \tag{5.3}$$

The analysis of the scale factor of the head is given below in the Fig. 5.5. This figure shows that the mean of the scale factor is close to 1.

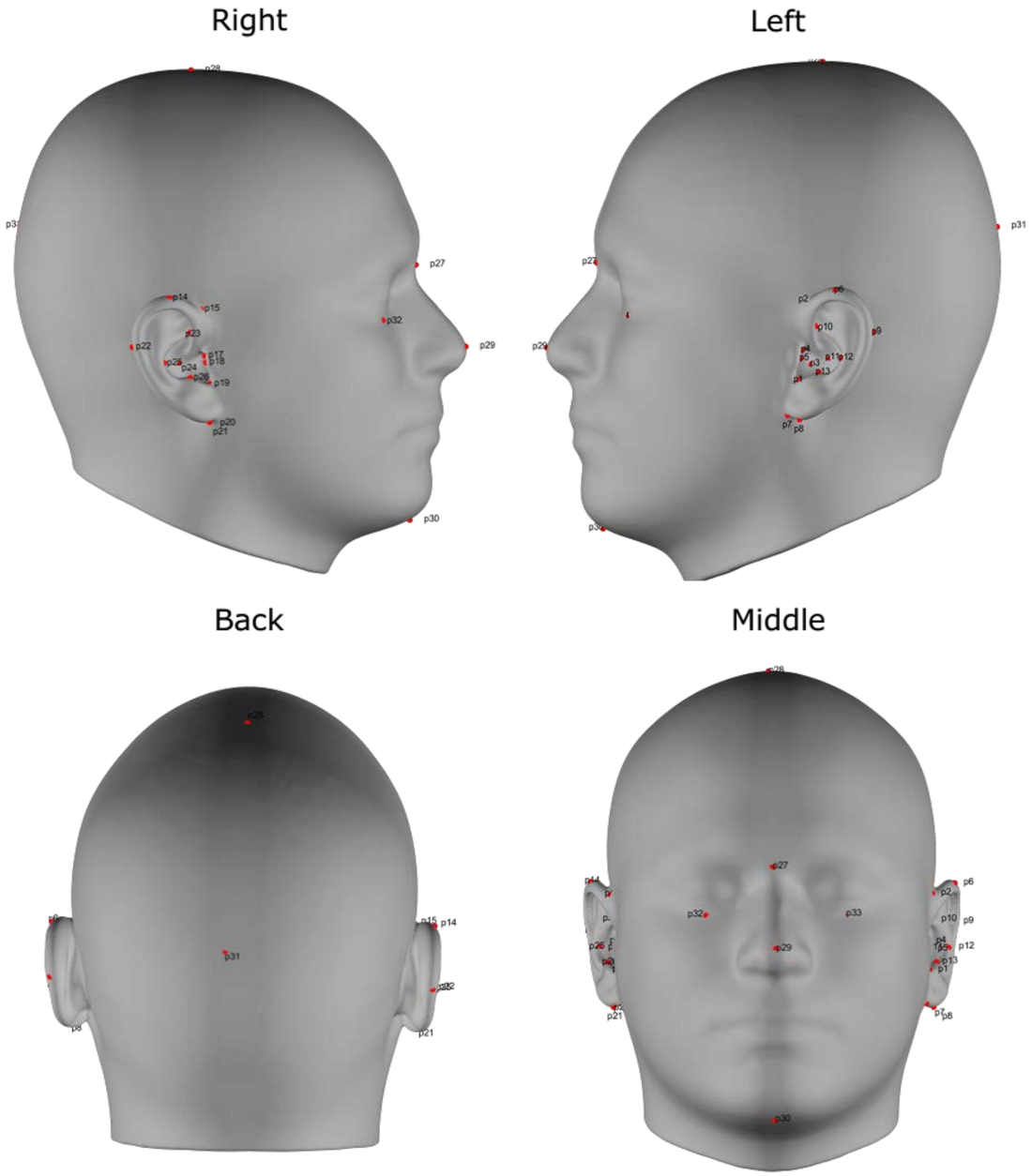


Figure 5.3: *Land mark points on the template ear and head shape used for anthropometric measurements. Similar points were measured for subjects in the SYMARE database. Points P_1-P_{13} are for the left ear shape, points P_{14}-P_{26} are for the right ear shape, and points P_{27}-P_{33} are for the head shape. (Picture taken from [1])*

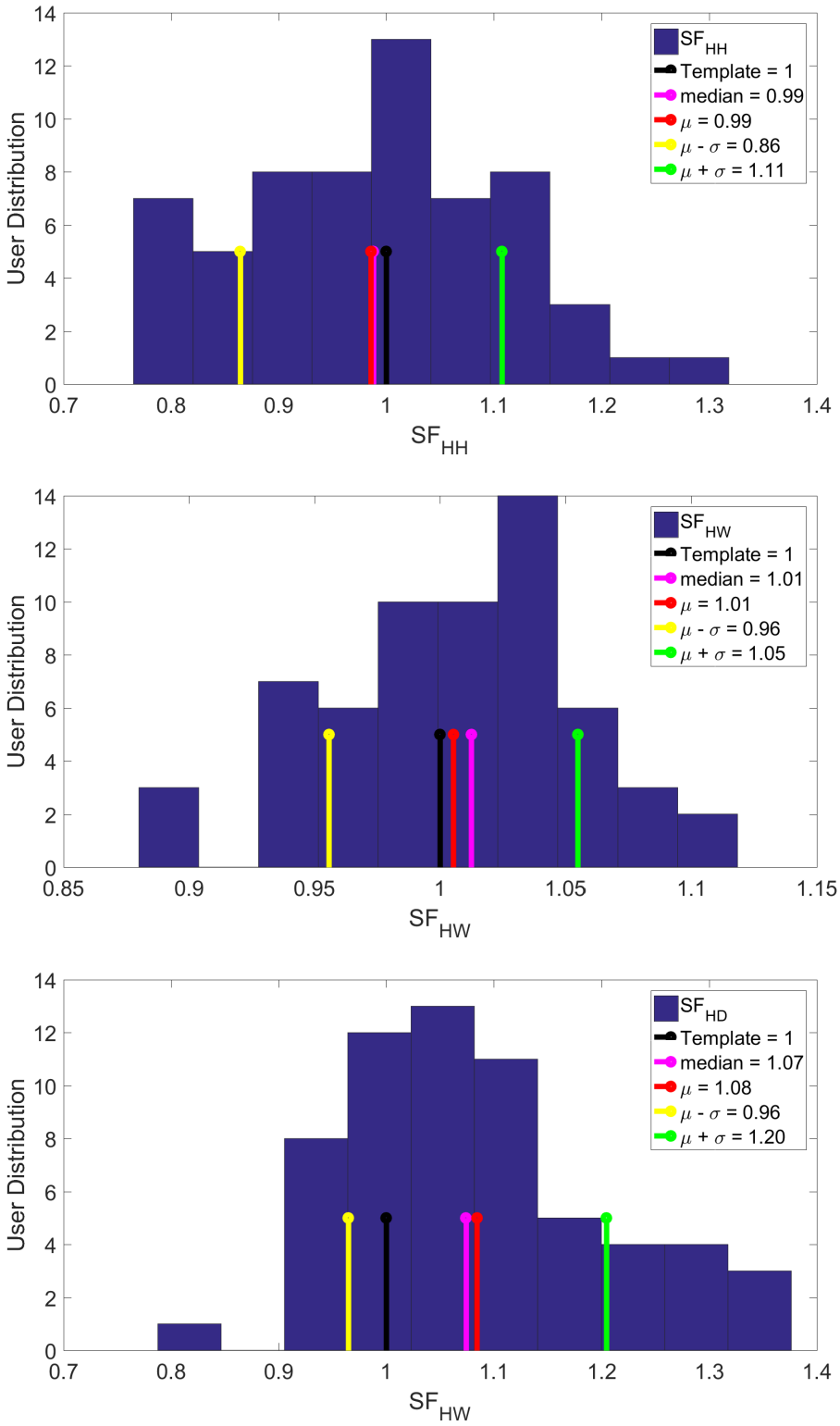


Figure 5.4: *Histogram for the scale factors for head-height, head-width, and head-depth, along with the μ, median, $\mu - \sigma$ and $\mu + \sigma$ values.*

The scale factors and rotation angles extracted in these studies are later used to

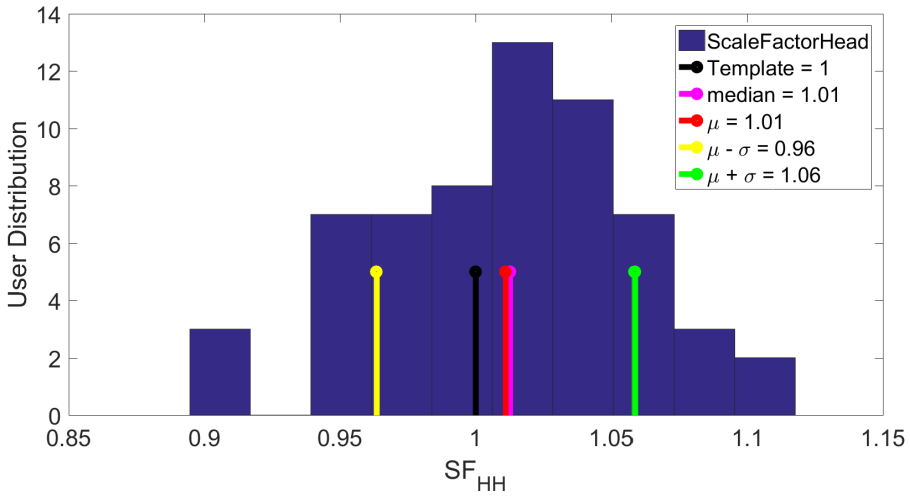


Figure 5.5: *Histogram for the scale factors for head, along with the μ, median, $\mu - \sigma$ and $\mu + \sigma$ values.*

correct and compensate for the effects of the affine transformations on the ear shapes and using of template head and torso shape instead of using original ones.

5.1.3 Attaching Affine Matched Ears to the Template Head and Torso

Once all the ear shapes are affine matched to template ear shape, their acoustic transfer functions can be calculated using BEM by attaching them to head and torso shapes. As the main aim of this chapter is to study the morphoacoustic properties of the ear shapes, keeping any other kind of variations in both shape and acoustics minimum, the affine matched ear shapes for all the subjects were placed on the same template head and torso shape. This section describes how the affine matched ears are attached to the template head and torso shape using the LDDMM framework. These head, torso, and ear shapes are called as the affine models for the subjects. The steps to attach the affine matched ear shapes to the template head, and torso using the LDDMM framework are provided in Algo. 6.

Algorithm 6 Attaching affine matched subject ears on template head and torso

inputs: $HTE_{Temp}, E_l, E_{Temp}, \sigma_V, \sigma_W$.
outputs: $HT_{Temp}E_l^{AM}$
$E_l^{AM} \leftarrow \mathcal{RM}(E_l, E_{Temp})$, will be used later for the corrections.
$\alpha_l^{AM}(t) \leftarrow \mathcal{M}(E_Temp, E_l^{AM}, \sigma_V, \sigma_W)$
$E_l^{AM} \leftarrow \mathcal{F}(E_{Temp}, \alpha_l^{AM}(t))$
$HT_{Temp}E_l^{AM} \leftarrow \mathcal{F}(HTE_{Temp}, \alpha_l^{AM}(t))$

First the ear shape for subject l, denoted by E_l is affine matched to the multi-scale template ear denoted by E_{Temp} using the approach described in Sec. 2.8. The affine matching process returns a shape denoted by E_l^{AM} which is matched to the template ear in size, orientation and center of mass position, along with the roto-scale matrix M_l and translation vector b_l. In the next step the template ear E_{Temp} is mapped to the affine matched ear E_l^{AM} using LDDMM. This process returns the momentum vectors $\alpha_l(t)$, describing how each of the vertices of template are are moved to new positions to match

the shape with the affine matched ear at different time steps t. Having these momentum vectors and the template ear shape in hand one can easily trace all the matching process for the vertices in template ear to the affine matched ear using the flow function of LDDMM framework. A geodesic flow is then applied on the template HTE_{Temp}, head and torso shape through the initial momentum vectors which will provide us with head torso shape of template with an affine matched ear shape of subject l replacing the template ear. This end shape is denoted as $HT_{Temp}E_l^{AM}$. This shape now can be used to run the FM-BEM simulations.

These affine models are unique and very special. By using the template head and torso shape for all the subjects, all the inter-subject variations due to different head and torso shapes are eliminated. While the ear scale, rotation, and position variations are removed by affine matching all the subjects to template ear shape. This leaves in the database the only variations due to the geometric shape variations of the ear shapes. To the best of the author's knowledge, this attribute of this work and the dataset produced as a result are unique, and no work has ever attempted this before. This is one of the biggest contributions of this study.

The transfer functions for these affine modeled shapes are then calculated using the process described in Sec. 2.3.

5.2 Acoustic Simplification introduced by Affine Model

One of the counter product of creating the affine models is the simplifications of the acoustics of the database. This section pursues an analytical study to answer the first research question of the chapter, i.e. quantification of the simplifications achieved by using the affine models instead of using the original head, ear and torso shapes. Furthermore, this section also provides some basis insights on how relevant the shape vs right scale, or rotation are, as the studies in [25], suggest that simple frequency scaling can provide a great way of HRTF personalization. This section critically analyse these claims.

Authors in [1] also studied the simplifications introduced due to the affine transformations for six subjects in section 7.2 of the study. However, there are two differences in this work compared to their work. The first difference is that while studying the simplifications they compared the acoustic responses of the affine model for six subjects to the normal ears translated to template position and put on the template head and torso. In other words for them both the original and affine transformed ear shapes were placed on the template head and torso ears, which simplify the problem. While in our study we are comparing the acoustic transfer functions of the original head, torso and ear shapes with the acoustic transfer functions for the affine models. As the final aim of this study is to see if we can compensate the process and obtain the original HRTF of the listener, i.e. when original head, torso and ear shapes are available. The second difference is that they analysed the variations for just six subjects while in this study we do this analysis for the full database of 62 subjects. Otherwise the used method to quantify the simplifications is the same.

As we have 62 (SYMARE database have only 61 subjects, but internally we have 62 subjects, the 62nd subject was not added in the database as we do not have the acoustically measured data for this subject) ear shapes in total in SYMARE database, pairing them up in all possible combinations we get $C_2^{62} = 1891$, combinations in total.

The acoustic differences or similarities were calculated between all pairs when affine matched vs non-affine matched or original shapes. More specifically we compared the acoustic responses of the affine-matched and original shapes, i.e. in real SYMARE database and in synthetic and simplified SYMARE database.

To calculate the simplifications in the HRTFs this work uses the same method as was used by [1]. We use a distance measure called standard deviation of spectral differences (SDS). SDS provides the standard deviation of the spectral difference between the acoustic responses of two ear shapes in all frequencies for a given direction i, denoted by θ_i, ϕ_i. The SDS measure between the acoustical responses of two shapes S_1 and S_2 is denoted by V_{S_1,S_2} and is given by:

$$\begin{aligned} x_n &= H_{S_1}(f_n, \theta_i, \phi_i) - H_{S_2}(f_n, \theta_i, \phi_i) \\ u_n &= \frac{1}{N}\sum_{n=1}^{N}(x_n) \\ V_{S_1,S_2}(H_{S_1}(f_n, \theta_i, \phi_i), H_{S_2}(f_n, \theta_i, \phi_i)) &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - u_n)^2} \end{aligned} \tag{5.4}$$

The function V_{S_1,S_2} in equation 5.4 is the standard deviation of the difference between the HRTFs for the two subjects and is calculated over a range of discrete frequencies f_n for the given direction i, denoted by θ_i, ϕ_i. We can also define a global measure as mean SDS measure which is useful to find the overall acoustical quality of matching between the two shapes as,

$$\bar{V}_{S_1,S_2} = \frac{1}{M}\sum_{i=1}^{M} V_{S_1,S_2}(H_{S_1}(f_n, \theta_i, \phi_i), H_{S_2}(f_n, \theta_i, \phi_i)), \tag{5.5}$$

where M, denotes the total number of directions. Note this measure is different from the measure used by [25] as this doesn't apply any preprocessing step to convert the linear frequency range to equivalent frequency bands (ERBs) on the HRTF data, as well as it has a unit of dB unlike the measure used by [25], which has units in dB^2. The sample plot of SDS is given in Fig. 5.6.

The SDS of 1891 pairs is calculated and the histogram of the mean SDS for all these pairs is given below in Fig. 5.7. Fig. 5.7a, plots the histogram of the SDS for all the pairs when they are in original condition, i.e. non-affine matched ear shapes on the original head and torso. While Fig.5.7b show that for the HRTFs produced from the BEM simulations ran over the meshes obtained from affine models for the same ear shapes. Finally the Fig. 5.7c, shows the histogram for the ratios of the two. The results show that almost 10% reduction in the variation in terms of mean SDS is observed when the affine models are used.

The results for six subjects calculated by [1] are provided in Fig. 5.8. The results provided in this figure are almost same as the results obtained by us. The slight difference could be due to the fact that the study in [1] compared the transfer functions of the original and affine matched ears both put at the template head and torso while we compared the differences between the original and affine ears when original ears are put on the original head and torso shapes.

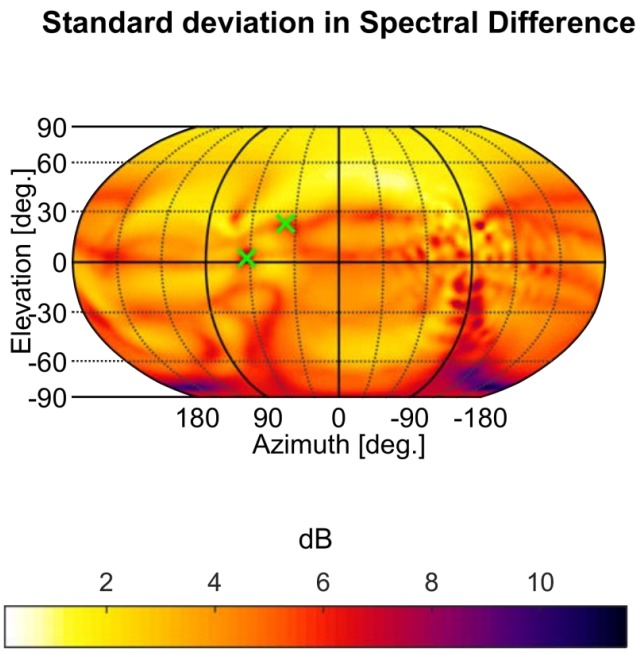


Figure 5.6: *A graph for function V between two subjects S_1 and S_2 is presented in this figure. The green crosses point the directions for maximum values.*

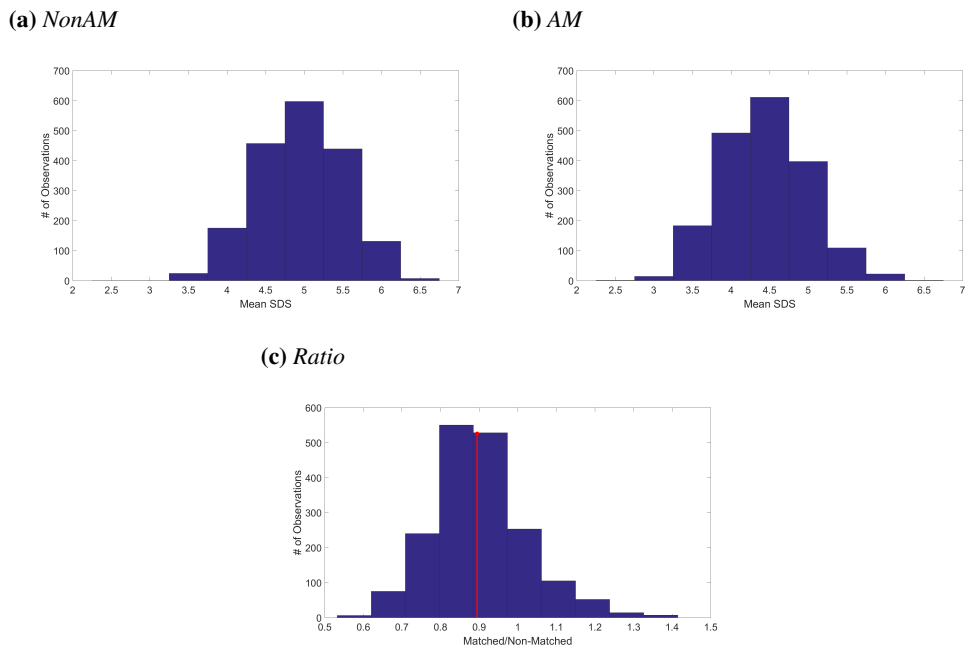


Figure 5.7: *The histograms of the SDS between every pair in the database are plotted when the a)Non affine matched (original) 3D head, ear and torso shapes are used, b) When affine models of the subjects are used, c) the ratio of a) and b).*

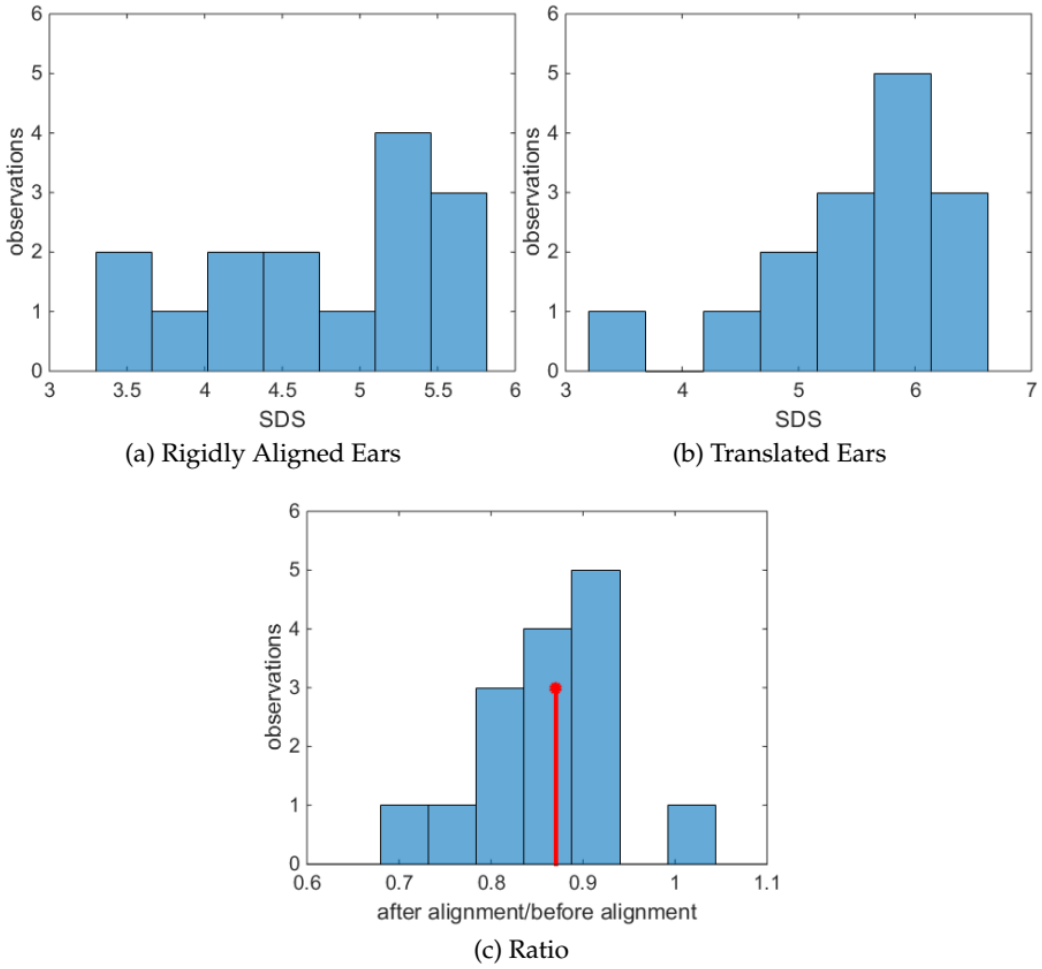


Figure 5.8: *The histograms of the SDS between six pairs in the database are plotted when the a)Non affine matched (original) 3D head, ear and torso shapes are used, b) When affine models of the subjects are used, c) the ratio of a) and b). Image taken from [1].*

The results presented in these two figures refute the claims made in [25], that the scale corrections only could be very useful in terms of personalization of HRTFs. These results clearly suggest that although the scaling and rotation matching does provide some sort of personalization, but the amount of provided personalization could account for only a small amount of total personalization possible. This indirectly suggests that the main contributors in the personalization of the HRTFs are the morphological features defining the shape of the ear not the scale and rotation.

5.3 Studying the Corrections and Compensations for Affine Matching

This section is the main section of this chapter and answers the last three research questions of the chapter. Which are

- Can the artifacts introduced in the process of affine matching and using the same head, torso, and ear shapes be corrected through simple frequency scaling (as proposed by [25]) and rotation of HRTFs?

- If yes, then how to find an optimal scaling for aligning the features of the HRTFs of affine matched ear shapes for low-frequency range (head contributions) and for high-frequency range (ear contributions)?

- How the obtained optimal scaling factors relate to the physical scaling factors to propose a simple scale and rotation correction approach for future studies, where optimal scaling factor can be derived simply from the head and ear scaling factors?

This work starts by creating an initial hypothesis on the basis of the work presented in [25], which suggests that the scaling differences in the shape corresponding to a scaling of the frequency axis. Knowing the scale in one space can help to fix the scale in another. However, the work presented here is different from that work in [25] in two main aspects.
1) The previous work studied the effects of the scaling factor between two subjects, which have a different head, ear, and torso shapes. However, in this thesis work, a unique dataset is created to study the effects of affine transformations of the same subjects, thanks to the powerful LDDMM framework to enable this work, this is a unique and novel work. 2) The second difference between the work presented in this thesis and Middlebrooks' work is that in this work, we run a simple simulation to identify the frequency regions which are affected by scaling of the head only, vs. the regions which are affected by the scaling of ears only. This helps us understand the frequency ranges to be scaled for head scaling and for ear scaling. Following these findings this study analyses the HRTFs for the frequency range from 0.2-17 kHz range to estimate three optimal scales, a) the optimal scale which matches the whole frequency range from 0.2-17 kHz, b) optimal scale which matches the frequency ranges for head, i.e., up to 5 kHz, c) the optimal scale which best matches the ear dependent frequencies, i.e., from 5-17kHz. The details of the process are given below. Finally, just like [25], we create a mapping between the obtained optimal scales and the physical scales using regression.

5.3.1 Understanding Head and Ear Scale Contributions

This section describes a preliminary study conducted to understand the contributions of head and ear scaling as a function of frequency. These studies are conducted on subject 55 in the SYMARE database as this subject has a very large head compared to the template head, while the size of the ear is the same as the size of the template ear. Three 3D models are created using the LDDMM framework: (i) Affine model of subject 55, i.e., actual ear shapes from the SYMARE database that are affine transformed to the template ear shapes using the affine matching process explained in Sec. 2.8 and attached on the template head. This mode is referred as $\overline{HT}E_{AM}$; (ii) Affine-transformed ear shapes on the scaled template head, $\overline{HT_S}E_{AM}$. The head is scaled with the scale factor corresponding to the biggest head in the dataset, which is 1.11 times or 11% larger than the template head. The process of calculating the scale factor is described in Sec. 5.1.2. The scale factor for the head is denoted by ξ_H; and, finally (iii) $\overline{HT}E_{Ac}$, which are the actual ear shapes attached the template head and torso shape.

The scale factor ξ_E and the rotation transformation $T(Rx, Ry, Rz)$ is obtained by affine matching from E_{AM} to E_{Ac} as described in Sec. 5.1.1. Here, ξ_E is the ratio of the size of the actual ear to the size of the affine-transformed ear. Rx, Ry and Rz are extrinsic rotation angles along the x, y and z axis. The x axis lies in the horizontal plane from the center of the head towards the nose tip. The y axis passes through the ear canals, and the z axis points upwards (see Figure 5.9). (i) and (iii) use a similar procedure to that of described in Sec. 5.1.3 to attach ears onto the template head and torso. Here in this section, only the procedure of generating the shape for (ii) is described, i.e., the scaled head and torso shape with affine matched ear shape.

To generate $\overline{HT_S}E_{Af}$, the affine aligned ears E_{Af} and the template ear $\overline{E}$ scaled with ξ_H are matched. The momentum vectors calculated are:

$$\{\boldsymbol{\alpha}_n(t)\}_{1\leq n\leq N}^{0\leq t\leq 1} = \mathscr{M}(\xi_H \cdot \overline{E}, E_{Af})\,. \tag{5.6}$$

Next, the flow of diffeomorphisms is applied to the template ear, head and torso $\overline{HTE}$ which is scaled by $\xi_H = 1.11$:

$$\overline{HT_S}E_{Af} = \mathscr{F}\left(\xi_H \cdot (\overline{HTE}), \{\boldsymbol{\alpha}_n(t)\}_{1\leq n\leq N}^{0\leq t\leq 1}\right)\,. \tag{5.7}$$

As a final result the affine matched ear shape E_{AM} is smoothly attached to the scaled head and torso shape as is shown in Figure 5.9.

Once these shapes are created the HRTFs for these shapes were then calculated using Coustyx FM-BEM simulations as described in Sec. 2.3. One thing to be noted is that in this section we are using only the directional transfer functions DTFs, by removing the common transfer function CTFs (average of the HRTFs) from these HRTFs. To analyse the influence of scaling the head, $D_{\overline{HT}E_{Af}}$ and $D_{\overline{HT_S}E_{Af}}$ were compared for four directions in the medium plane of azimuth, elevation $(0°, 45°)$ and $(0°, -45°)$ in front of the head and $(180°, 45°)$ and $(180°, -45°)$ at the back. The results are shown in Fig. 5.10, which shows the DTF data for (a) $(45°, 0°)$, (b) $(-45°, 0°)$, (c) $(-45°, 180°)$ and (d) $(-45°, 180°)$; directions. In every subfigure from top to bottom we have the DTFs of affine transformed ear on the template head; affine transformed ear on the scaled template head; scale corrected ear on the template head and actual ear on the template head. $D_{\overline{HT}E_{Af}}$ are corrected to provide us $\widetilde{D_{SC}}$, which are calculated by

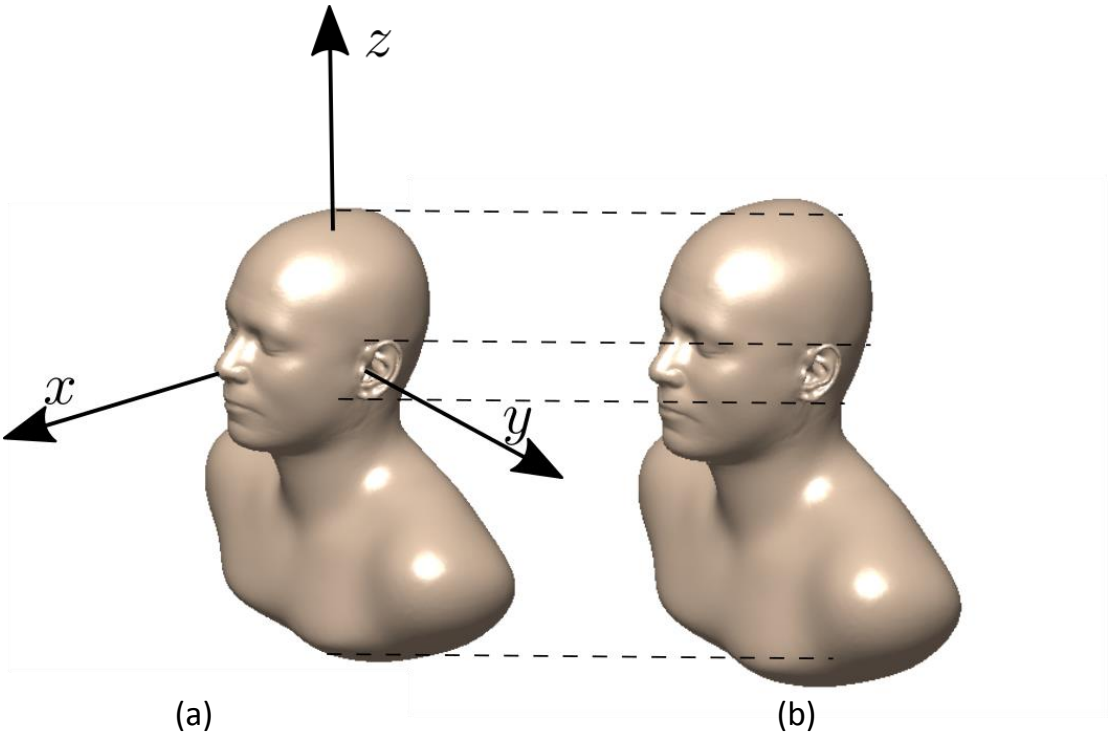


Figure 5.9: *This figure shows two of the models generated to study the contributions of the head and ear scalings. Showing two shapes in the experiment for subject 55, (a) template head and torso shape with affine matched ear (affine model for subject 55) (b) affine matched ear on a scaled template head and torso.*

simply scaling the frequency axis of the DTFs to compensate for the affine matching. We first translated the coordinate system from the centre of the head to the ear canal. Then, $\widetilde{D_{SC}}$, the DTFs after applying the scale correction were calculated:

$$\widetilde{D_{SC}}(f,\theta,\phi) = D_{\overline{HTE}_{Af}}(\xi_E \cdot f,\theta,\phi). \tag{5.8}$$

where f is the frequency. Furthermore a simple rotation correction was applied to obtain $\widetilde{D_{SRC}}$:

$$\widetilde{D_{SRC}}(f,\theta,\phi) = \widetilde{D_{SC}}(f,\theta,\phi) \circ T(R_x, Ry, Rz), \tag{5.9}$$

where $T(R_x, R_y, R_z)$ is the same rotation transformation that was applied to the ear. To demonstrate the performance of the scale correction, DTFs for some directions for subject 55 ($\xi_E = 0.94$) appear in the last two rows in Figure 5.10 These results clearly show two things.

1. The scaling of the head has no effect on the HRTFs of the higher frequencies and scale on the frequency axis for lower frequency region, namely up to 4-5 kHz. For example, the notch labeled as N_{2_i} stays unchanged in its location for both cases which have scaled or normal head and torso shape.

2. The ear scaling doesn't work on the lower frequency content. In fact, it makes the matching worse by scaling the frequency axis in this region. However, it manages to align the frequency content in the high-frequency region i.e., $\geq$5 kHz. This

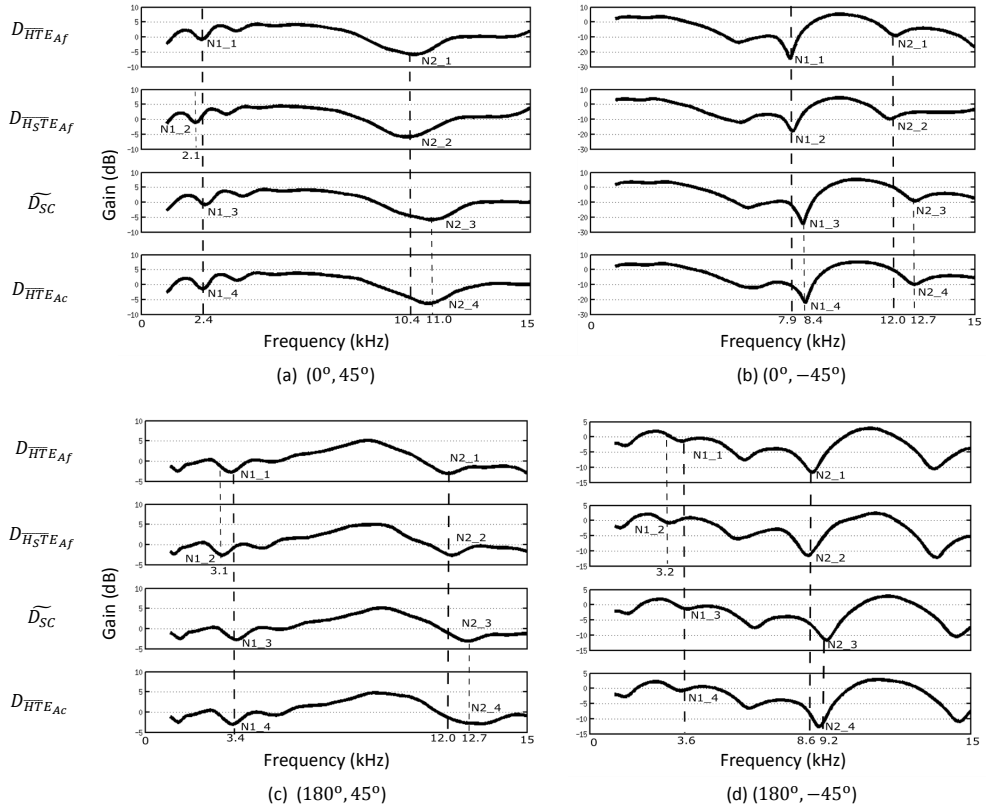


Figure 5.10: *DTFs on the medium plane of subject 055 for (a) $(45°, 0°)$, (b) $(-45°, 0°)$, (c) $(-45°, 180°)$ and (d) $(-45°, 180°)$; from top to bottom, each figures shows DTFs of affine transformed ear on the template head; affine transformed ear on the scaled template head; scale corrected ear on the template head and actual ear on the template head.*

clearly is against the claims made by Middlebrooks in his study, who reported formula for the optimal scaling suggesting the contribution of head scales to be higher than of the ear scale.

These two findings suggest that although single frequency scaling corresponding to the ear scaling can perform very well in the higher frequency content, for lower frequencies, head scaling is to be used, unlike using a single scaling as proposed by [25]. In the following section, we provide a simple analysis to compute the optimal scaling for the affine models for all the subjects in the SYMARE population. Furthermore, it presents a simple mapping inferred using simple linear regression, which will let us compute the optimal scaling for two frequency ranges based on the simple anthropometric measure.

5.3.2 Finding an Optimum Scaling Factor for Frequency Axis

Following the findings of the above-mentioned experiment, the aim of this section is to study the optimal frequency axis scaling, which best matches the affine models HRTFs to the original HRTFs by studying only the acoustic data. For the demonstration purpose this section uses four subjects from SYMARE namely subject 21(has almost same head and ear scale), subject 23(have the largest ears in the population with slightly large head than of the template), subject 37 (has smaller ears while the head size is almost

Subject number	Head Scale Factor	Ear Scale Factor
21	87.70	88.00
23	1.06	1.15
37	1.01	87.81
56	1.11	1.00

Table 5.1: *Head and ear scale factors for subjects [21, 23, 37, 55].*

same as the template), and finally subject 56 (has the largest head in the SYMARE population, while the ear size is same as of the template). Tab. 5.1 provides the head and ear scale factors for these subjects. These scale factors are calculated using the methods described in Sec. 2.8.1. This section finds an optimal scale for every frequency by studying the directivity patterns for every frequency one by one. The directivity pattern of an HRTF for a given frequency presents gains and losses for that frequency in all directions. An SFRS plot of a sample directivity pattern for a subject from SYMARE is given in Fig. 5.11. To find an optimal scaling, we need to define a similarity or mismatch measure for directivity patterns, and the optimal scaling factor will be a scaling factor that will result in the highest similarity or lowest dissimilarity between the affine model and original HRTFs. For this purpose, this study employed the measure called Spatial Correlation Metric (SCM) given in [32]. The SCM is denoted mathematically as $C_{D1,D2}(f)$, signifying that it is a function of frequency f, and parameterized by the two inputs (in our case the directivity patterns $D_1(f)$ and $D_2(f)$). The function C provides a measure of how similar two directivity patterns are $D_1(f)$ and $D_2(f)$ are across all the directions in 3D space. It has a single scalar value, which is between 0 and 1 (which can also be translated between 0% to 100% by simply multiplying it with 100) for any given pair of directivity patterns. A value of 0% indicates now matching at all, while a value of 100% means to directivity patterns have the exact same shape. As we are going to perform analysis across frequency for two sets of HRTFs, it will have a vector of values, one for each frequency as a function of frequency. A similar kind of metric was also used by [32, 51, 61].

Given the log-magnitude for the directivity patterns $D_1(f, \theta, \phi)$ and $D_2(f, \theta, \phi)$ the SCM between these two directivity patterns is computed as:

$$\overline{D_1(f,\theta,\phi)} = \frac{1}{M}\sum_{i=1}^{M} D_1(f,\theta_i,\phi_i) \tag{5.10}$$

$$\overline{D_2(f,\theta,\phi)} = \frac{1}{M}\sum_{i=1}^{M} D_2(f,\theta_i,\phi_i) \tag{5.11}$$

$$\zeta(D_1,f) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(D_1(f,\theta_i,\phi_i) - \overline{D_1(f,\theta,\phi)})^2} \tag{5.12}$$

$$\zeta(D_2,f) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(D_2(f,\theta_i,\phi_i) - \overline{D_2(f,\theta,\phi)})^2} \tag{5.13}$$

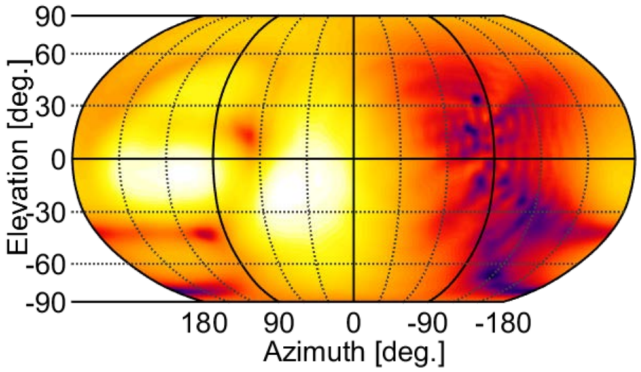


Figure 5.11: *Directivity pattern for a random subject at 12 kHz represented as and SFRS.*

The SCM between two HRTFs is then obtained as a value for every frequency using these expressions below:

$$C_{D_1(f),D_2(f)} = \frac{\sum_{i=1}^{M}[D_1 - \overline{D_1(f,\theta,\phi)}][D_2 - \overline{D_2(f,\theta,\phi)}]}{\zeta(D_1,f)\zeta(D_2,f)}. \tag{5.14}$$

For the sake of making the differences amongst the affine matched and non-affine matched HRTFs, we are using the Squared SCM or SSCM, which is just the square of the SCM measure. A global measure for the whole frequency range under consideration can be found by simply taking an average of the SCM for all the frequency bins under consideration as:

$$\overline{C_{D_1,D_2}} = \frac{1}{N_f}\sum_{i=1}^{N_f} C^2_{D_1(f_i),D_2(f_i)}. \tag{5.15}$$

The expression in Eq. 5.15 is referred as Global Squared Spatial Correlation Metric or (GSSCM). Following we provide an analysis on finding the optimal scale for some subjects and the improvements in terms of average SSCM achieved towards individualized HRTFs using simple scaling and rotation corrections. Furthermore, we also show the dependencies of these optimal scales on the head and ear scale factors. Three optimal scales are calculated in this work:

1. optimal scale for whole frequency range, i.e., an optimal scale which results in the highest GSSCM for the frequency range from 0.2-17 kHz.

2. optimal scale for head-dependent range, i.e., optimal scale, which results in the highest GSSCM for the frequency range from 0.2-5 kHz.

3. and finally, the optimal scale for ear dependent range, i.e., an optimal scale which results in the highest GSSCM for the frequency range from 5-17 kHz.

To find the optimal scaling factor, we sweep across a range of the scaling factors (51 scales to be exact). The range of scaling factors explored is from $1.01^{-25} = 0.78$ to $1.01^{+25} = 1.28$, with every scale to be a power of 1.01. The reason to use this multiplication factor is that this way, we only scale for 1% at every step. Fig. 5.12,

Fig. 5.13, Fig. 5.14 and Fig. 5.15 show the optimal frequency scaling analysis for subjects [21,23,37,56]. Every figure has 17 subfigures in total. The first subfigure in top-left presents the evolution of GSSCM as a function of scaling factors in three different frequency ranges. The diamonds show the resulting GSSCM for given frequency ranges for the scaling factor of the frequency range. The second subfigures in every figure report the SSCM functions for all the frequencies for five cases each showing SSCM between directivity patterns of actual 3D model of the subject with the (from bottom to top) i) directivity patterns of affine model of the subject, and then scale and rotation corrected directivity patterns when ii) only the scaling based on ear scale is applied, iii) optimal head scale is applied, iv) optimal ear scale is applied, v) optimal scale for whole frequency range is applied. In the next 15 subfigures, the directivity patterns corresponding to 5 frequencies (2 kHz, 4 kHz, 6 kHz, 9 kHz, and 15 kHz) are presented with one row for each frequency, and from left to right these have directivity patterns of actual models, affine models, and (optimal) scale and rotation corrected directivity patterns. To generate the directivity patterns in these figures, optimal head scaling was used for the directivity patterns at 2 kHz and 4 kHz, while for the directivity patterns at higher frequencies, the optimal scaling for the ear is used as these frequencies belong to head and ear regions respectively. Following, we provide an analysis of each of the figures and see how these figures answer the posed research questions in this chapter.

It can be seen for subject 21, where the scale factors for head and ear are almost identical with values $\xi_h = 0.89$ and $\xi_E = 0.877$, that the optimal scale factor for head is very close to the physical scale factor for head ξ_H, and the optimal scale factor for ear and overall scaling factors are (0.85) 2% smaller than that of the physical scale factor ξ_E. The performance of all the four scalings is almost identical in the head range (might be because of small differences and low-frequency range), while in the high-frequency range, the differences start to appear and ear scaling plays a more vital role. Furthermore, in this range, the global and ear based optimal scale performs slightly better than of the simple ear scaling. This, is also apparent while looking at the directivity patterns for all frequencies. In lower frequency range, the changes are not as sudden as in the higher frequencies; still, the simple rotation and scale correction manage to capture the small subtleties in shape, and the scale corrected directivity pattern appears to be correctly oriented and shaped. While for the directivity patterns from 4 kHz onwards, these changes are more apparent, especially at 9 and 14 kHz, where the affine model has a directivity pattern that looks completely different from the actual directivity patterns. This shows the power of simple scale and rotation corrections and also validates the use of our affine model with proper corrections in place.

Subject 23 is a special case as it has the largest ear shapes in the SYMARE database with ear scaling factor $\xi_E = 1.15$, while the head size is only 6% larger than of the size of the template head with a scaling factor $\xi_H = 1.06$. For this subject again, the global and ear optimal scales are very close to each other and are a little bit smaller than the physical scale of the ear with a value of 1.125. While the optimal head scale in this region is 5% smaller than of the template head, which is exactly the opposite of what the physical head scale indicates. The only explanation for this that came to mind is that it may be a larger size of the ear that causes some interplay between the head and ear shapes ending up creating this artifact. This is also clear from the second subfigure, which clearly shows that even with the optimal head scaling, the values for SSCM are

still worse than the affine model indicating the optimal frequency search failing for this subject for the head. Furthermore, the scale corrections are very good in the frequency ranges above 4 kHz, as is clear from the second figure as well as from the directivity patterns.

Subject 37 has ear which is 13% smaller than the template ear with a scaling factor of $\zeta_E = 0.8781$ while the head size is almost the same as the template head with a scaling factor of $\zeta_H = 1.01$. The optimal scale factors and optimal ear range scale factors again exhibit the same behavior as they did for the previous two subjects being very close but slightly lower than the physical ear scale factor ζ_E, while the head factor is smaller than 1 with a value of $\zeta_H = 0.93$. Also, for this subject, the scaling does not work in the head range i.e., up to 5 kHz, while it works very well for the ear frequency range. This further confirms the speculation that when the scaling factor for the ear is very large (being ears to be very small or very large), the scaling will not work in the lower frequency range. Also, for this subject, the scaling works very nicely for higher frequency ranges, as is shown in directivity patterns of 6, 9, and 14 kHz.

Finally, we analyze the results for subject 56, which is also a special case having the largest head in the SYMARE database, with a scale factor of 1.12, while the size of the ear is almost the same as the template ear with a scaling factor $\zeta_E \approx 1$. The optimal scale factors, both global and ear region, are again in agreement with their previous behavior being slightly smaller than that of the actual ear scaling factor and global scaling factor to be slightly larger than the optimal ear scale. Also, the head scale factor is the same as before smaller than that of the actual head scale. For this subject, the optimal scaling for ears causes a little improvement in the higher frequency region.

One might have a question of why the optimal scaling for the head is resulting in degradation of the SCM even when compared to no scaling at all. The reason for that is the correction for rotation, which is coming from the ear shapes always and is kept the same for all the scaling factors. This generally seems to work, so we did not remove it from the head range as well.

5.3.3 Quantifying the Improvements Achieved through Simple Scale and Rotation Corrections

This section quantifies the improvements achieved by performing simple scaling and rotation corrections. The Fig. **??** shows the histograms of the GSCM between the actual HRTFs and HRTFs with scale, rotation corrected for three cases:

1. simple ear scale correction.
2. optimal scale correction (with optimal global scale).
3. composite optimal scale correction (optimal scale for head until 5 kHz and optimal scale ear after 5 kHz to 17 kHz).

The results show that the directivity patterns of the affine models, on average, match the actual models of the subject to 61% of extent. By simply scaling the frequency axis of HRTFs by the relative scale of the ear shapes, the matching improves by a factor of almost 21% and becomes 74%. The optimal scaling increases this matching by 5% more as the average GSSCM becomes 77%. Then composite optimal scaling further improves it by providing a gain of 2% more, making the total GSSCM value 78%.

This answers the third and fourth questions posed in this work showing how much corrections in terms of improved matching can be achieved using simple scaling and rotation corrections, as well as create three scaling factors for three frequency ranges.

5.3.4 Deriving Scaling Factors from simple Anthropometry

Finally this section presents a simple regression analysis on the physical scales ξ_E and ξ_H to derive the global, ear and head optimal scaling factors ξ_{opt}, $\xi_{H_{opt}}$ and $\xi_{E_{opt}}$. Simple regression analysis shows that these scale factors can be estimated using the following expressions:

$$\zeta_o pt = 1.0198\zeta_E - 0.0105\zeta_H - 0.0348C \quad (5.16)$$

$$\zeta_{H_{opt}} = 0.6062\zeta_E + 0.0719\zeta_H + 0.3338 \quad (5.17)$$

$$\zeta_{E_{opt}} = 1.0247\zeta_E - 0.0127\zeta_H - 0.0402 \quad (5.18)$$

When estimating the optimal scale factors through these expressions, the average error between the original and estimated scale values for global scale, ear scale, and head scale are 0.8%, 0.8%, and 7%, respectively. The possible reason for the large value of the estimation error for head scale could be that the rotation coming from the ear matching is also used for the head frequency range directivity patterns while finding an optimal scale.

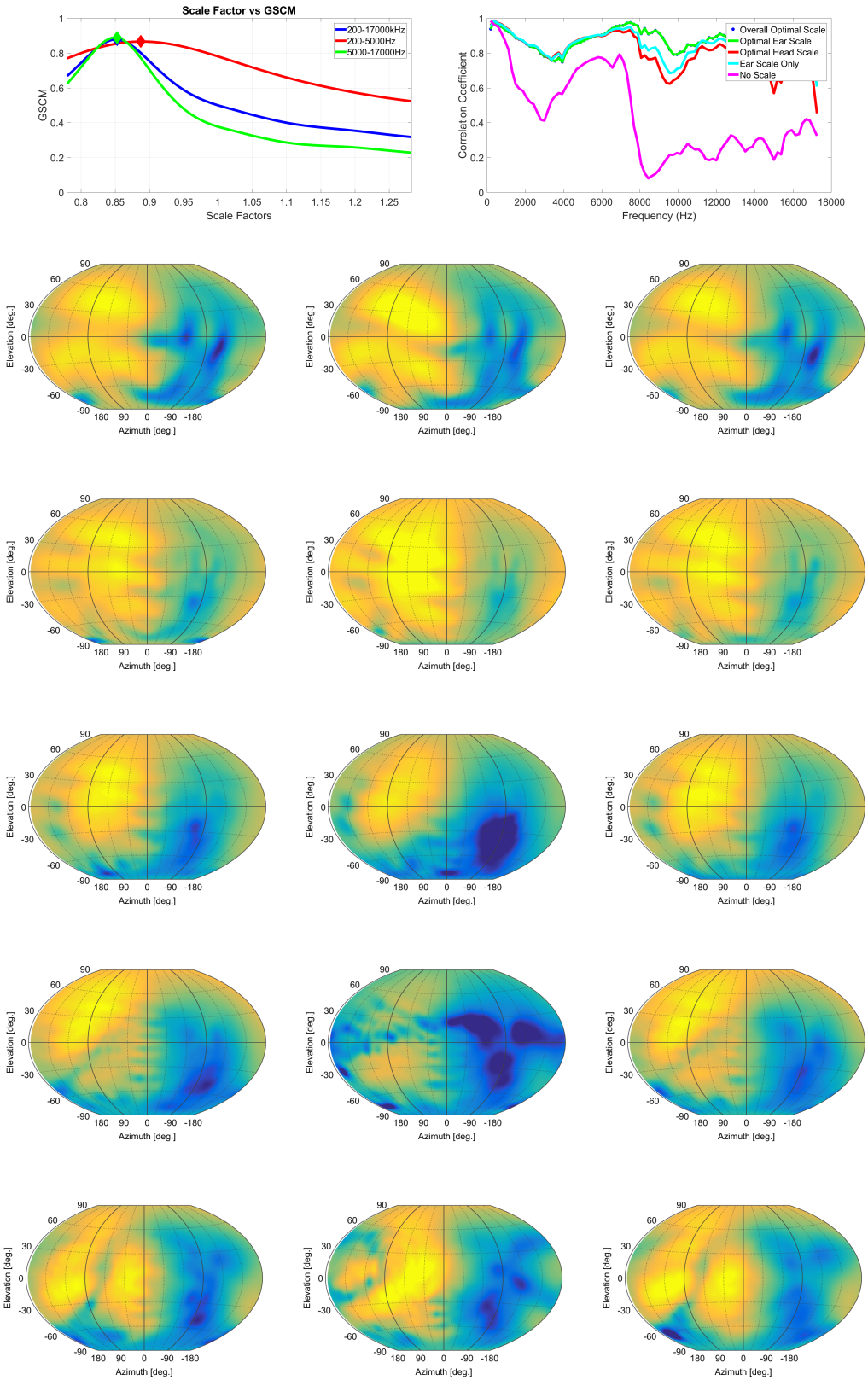


Figure 5.12: *This figure shows the optimal scale searching study for subject 21. In top row(left) GSCM (average SCM) vs explored scale factors, (top right) cross correlation vs frequency for different scale corrections. Row 2-6 show directivity patterns for 2, 4, 6, 9, and 14 kHz with actual, affine matched and optimal scale corrected directivity patterns in left, middle and right column respectively.*

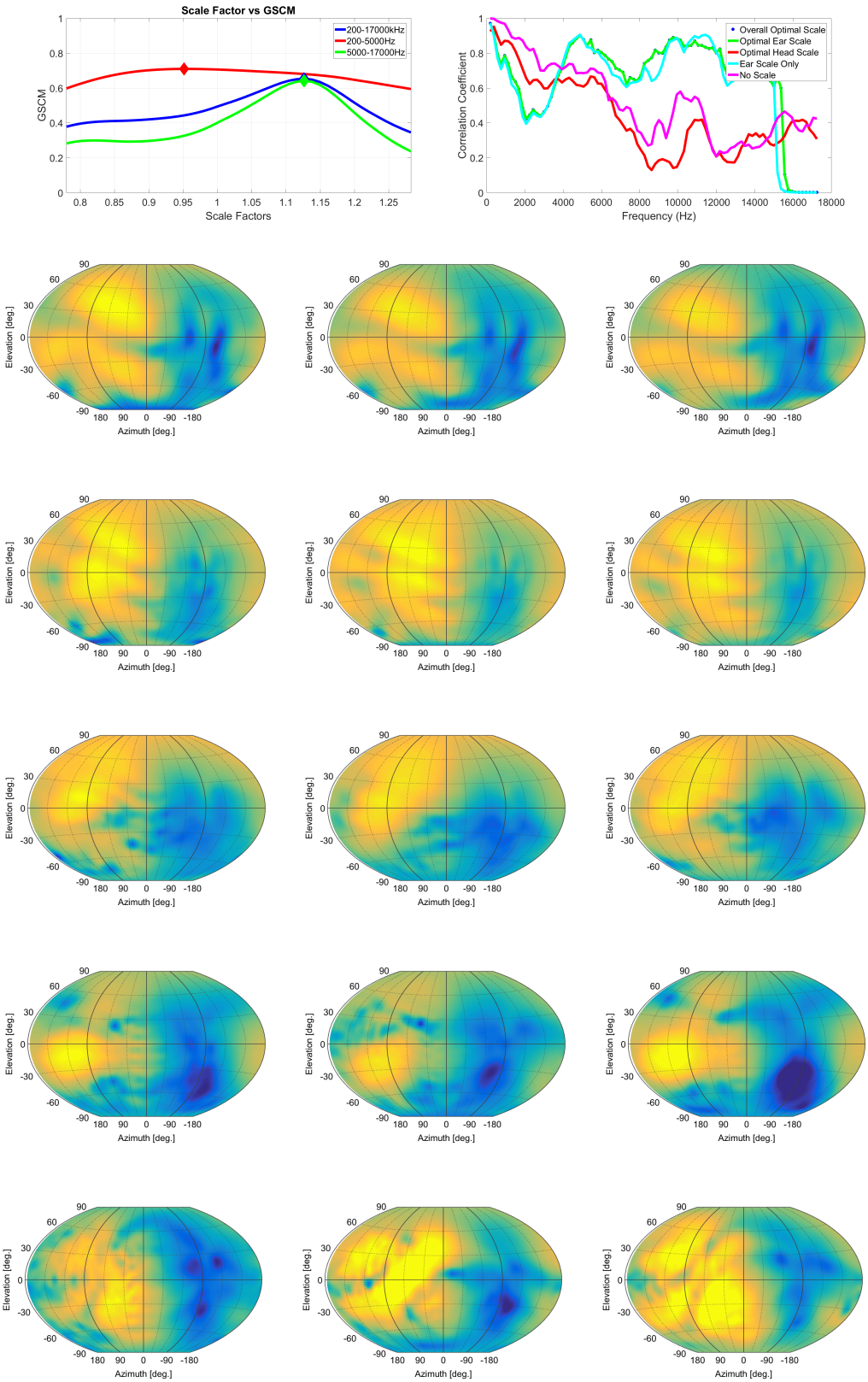


Figure 5.13: *This figure shows the optimal scale searching study for subject 23. In top row(left) GSCM (average SCM) vs explored scale factors, (top right) cross correlation vs frequency for different scale corrections. Row 2-6 show directivity patterns for 2, 4, 6, 9, and 14 kHz with actual, affine matched and optimal scale corrected directivity patterns in left, middle and right column respectively.*

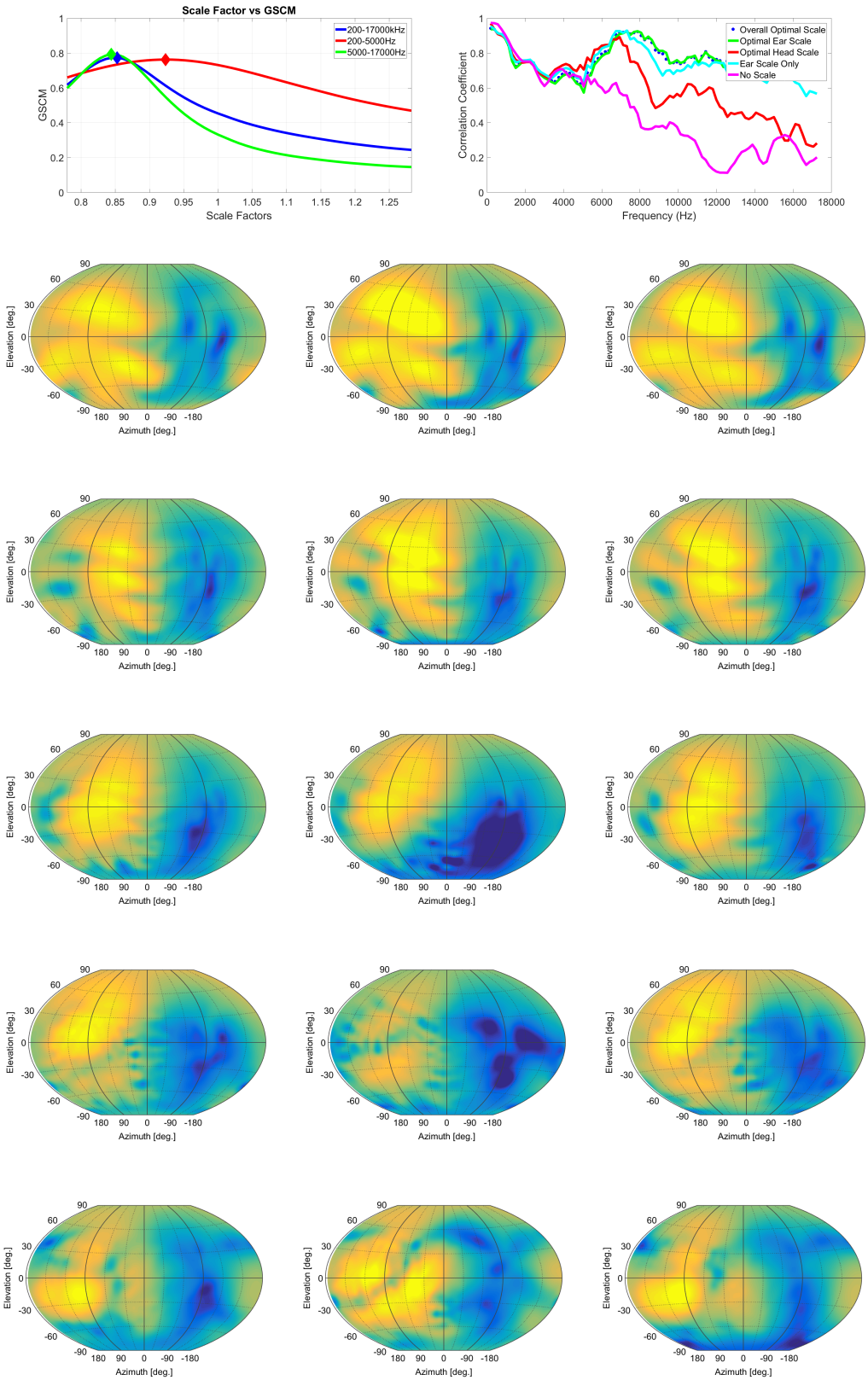


Figure 5.14: *This figure shows the optimal scale searching study for subject 37. In top row(left) GSCM (average SCM) vs explored scale factors, (top right) cross correlation vs frequency for different scale corrections. Row 2-6 show directivity patterns for 2, 4, 6, 9, and 14 kHz with actual, affine matched and optimal scale corrected directivity patterns in left, middle and right column respectively.*

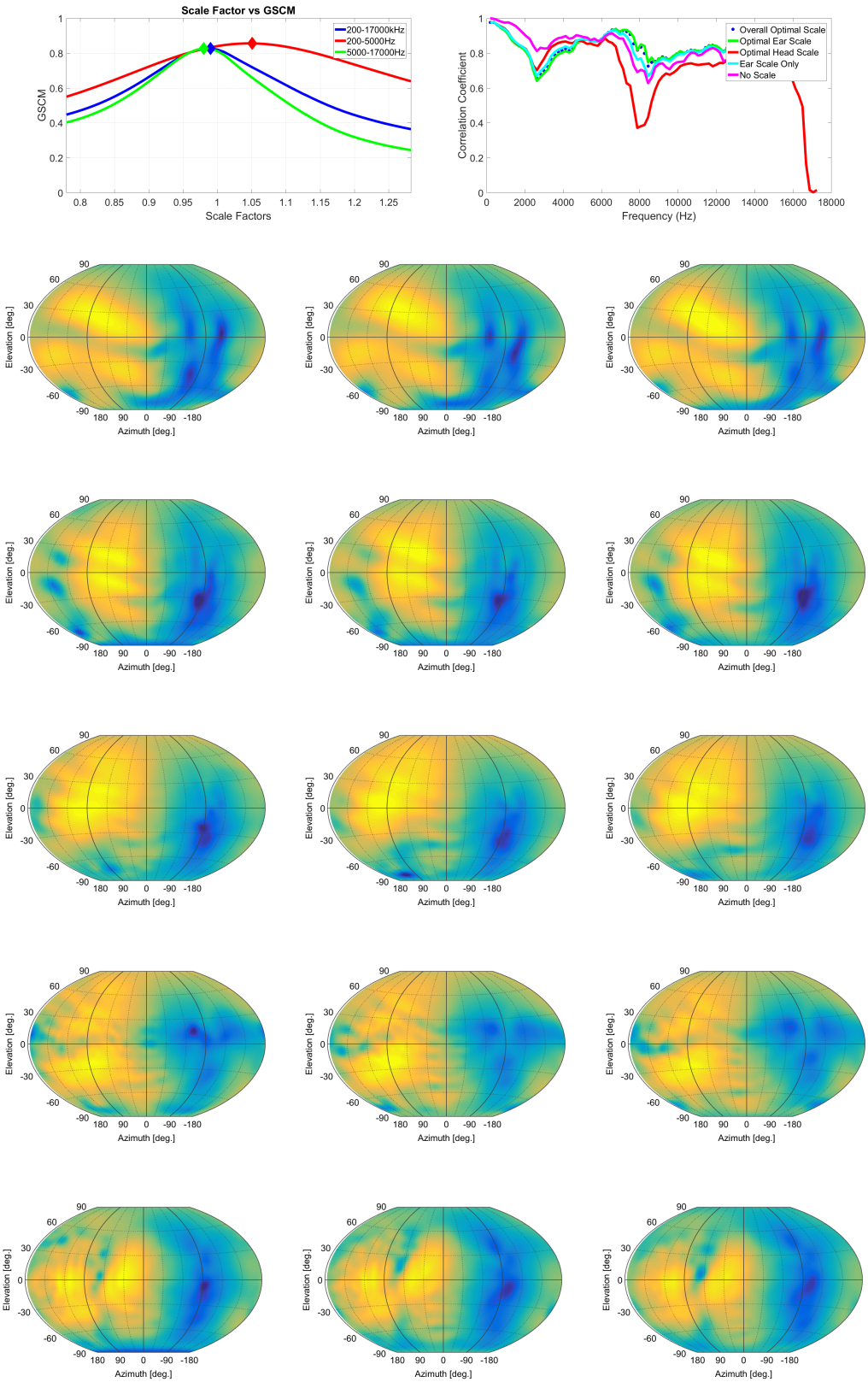


Figure 5.15: *This figure shows the optimal scale searching study for subject 56. In top row(left) GSCM (average SCM) vs explored scale factors, (top right) cross correlation vs frequency for different scale corrections. Row 2-6 show directivity patterns for 2, 4, 6, 9, and 14 kHz with actual, affine matched and optimal scale corrected directivity patterns in left, middle and right column respectively.*

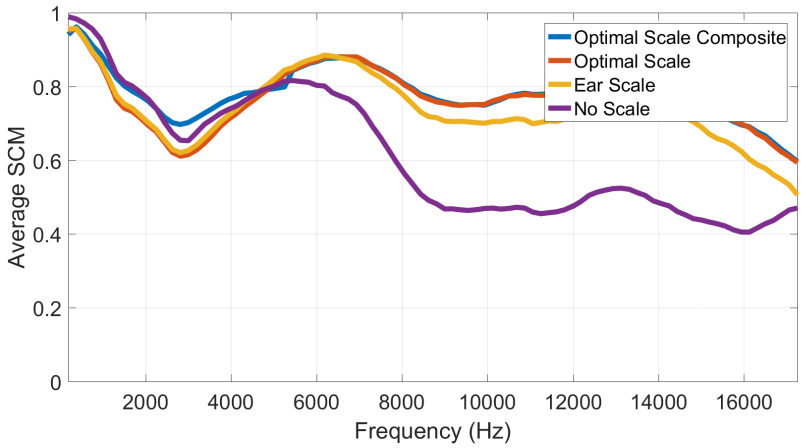


Figure 5.16: *Summary plot for SSCM between the real and scale and rotation corrected directivity patterns of SYMARE database as a function of frequency.*

5.4 Conclusion

In this chapter, we presented a simple yet powerful approach to creating a simple 3D model for the sake of HRTF personalization. We showed how this model could be made by just having the 3D shape and sizes of the subject without the need to capture the full head and torso shapes. The quantification of the acoustic simplification was also calculated, which is one of the main aims of doing this study. It showed that using the affine model reduces the acoustic complexity of about 10% compared to original HRTFs, which can be very useful in order to model the acoustic data. Furthermore, affine matching of the ear shapes was performed before creating the morphable model of the ear shapes in [17], and the corrections performed in this chapter are essential to retrieve a good estimate of the actual ear and head shape model when this morphable model is used. Finally, this chapter presented a simple study to find the optimal scaling factors to be used for the whole frequency range, or for head and ear ranges separately. The last section also shows a simple way of calculating these scale factors from the physical scale factors of head and ear shapes compared to the template. The results show that these scale factors mainly depend on the scaling factor of ear shape even in the range of the head-dependent frequencies. Future studies for this work include performing independent analysis of head and ear scale and including the rotation of head and torso as well as torso scale factors in the studies.

5.5 Chapter Conclusion

This chapter performs a comprehensive study using LDDMM framework to analyze the effects of the affine transformations on the ear shapes on corresponding acoustics. This chapter answered five research questions. The findings of the chapter were followings: the use of affine models results in 10% simplifications of the HRTFs. The template resides well in the SYMARE database when the scale and orientation of the ear population is analyzed compared to the template ear shape. A simple database is created to study the ear scale, rotation, and shape morphoacoustics separately. The scaling of ear

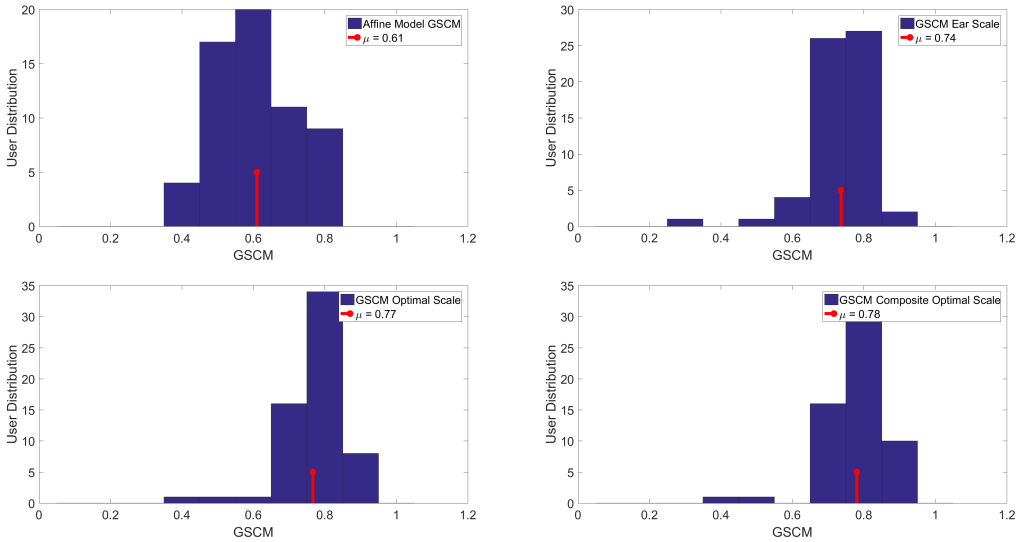


Figure 5.17: *Histograms of the GSCM for four cases. a) GSSCM between actual and affine models of the SYMARE population, b) GSSCM of actual and scale rotation corrected affine models (ear scale is used for correction) models.c) GSSCM of the actual and affine model with optimal scale and rotation correction, d) GSSCM of actual and composite optimal scale and rotation correction. i.e., the optimal head scale is used for frequencies up to 5 kHz and after that optimal ear scale is used.*

shapes only show changes in frequencies above 5 kHz. Using optimal scaling factors simple frequency axis scaling and rotation corrections for ear shapes can result in a 28% better matching of the HRTFs for affine models with the original HRTFs. Finally the optimal scaling can be predicted using the physical scaling factors for head and ear shapes.

CHAPTER 6

Principal Component Analysis on Head-related Transfer Functions

In this chapter, an end to end HRTF personalization method is proposed, which produces the HRTFs for an affine matched 3D ear shape without running the BEM simulations. More specifically, this chapter models the variations in the morphology using the proposed morphable ear shape model and models the variations in the acoustics using PCA. Then it creates a simple linear regression-based mapping to relate the variations in one domain to the variations of the other domain.

The research questions explored in this chapter are as follows:

- How can the inter-subject variations be modeled using PCA as a function of frequency?
- How many principal components are required to model the variations for a given frequency, and what kind of reconstruction they provide for the given frequency in a data constrained manner? (We can not use more than one-eighth of the data).
- What kind of variations in the ear shape be captured using only one-eighth of the data?
- Can linear regression model the relationship between morphology and acoustic variation models to create an HRTF personalization method?
- How does morphological weighting improve the prediction, and can it be used as a tool to understand the relative contributions of each of the ear part?

The contributions of this chapter are as follows:

1. A procedure to model the HRTFs for affine models of the SYMARE database using frequency by frequency PCA of the acoustic directivity patterns. As this

PCA model the spherical or spatial surfaces, we termed this as Spatial Principal Component Analysis (SPCA).

2. Quantification of the number of SPCA components required to model the directivity patterns of a given frequency. The results show that even when only one-eighth of the data is used to model the dataset, this can still provide very good results even for 17 kHz directivity patterns.
3. An HRTF personalization method using simple multiple linear regression on the ear parameters.
4. A novel Weighted KCPA based model to improve the personalized predictions of the directivity patterns. This model can also be used as a potential variant of the morphoacoustic perturbation analysis using LDDMM.

The rest of the chapter is organized as follows: Sec. 6.1 described the process of preparing the acoustic directivity patterns to be used for this study, along with the preprocessing steps performed. Sec. 6.2 describes the application of spatial principal component analysis on the directivity data. Sec. 6.3 describe the process of quantifying the number of principal components required to model the directivity patterns of a given frequency. Sec. 6.4 starts by providing a simple way to analyze the variations of the ear shapes using a morphable ear shape model in [17] and use the parameters in the morphology domain to create a linear regression-based mapping between the morphology and acoustic domain parameters providing a simple HRTF personalization approach. Finally, Sec. 6.5 propose a novel yet straightforward method that uses all the developed methods in this chapter and improves the personalized prediction of the acoustic parameters. Furthermore, its use as a potential tool to understand the contributions of each of the ear components in the HRTF generation is explored.

6.1 Preparing the Acoustics

This work uses PCA to analyze the acoustic variation of the affine models of SYMARE subjects. However, unlike previous studies, it studies the directivity patterns of the acoustic data instead of studying the whole HRTFs direction by direction. For a given frequency, the directivity patterns present the pattern of acoustic gains and attenuations across space for a given ear shape. The ear shapes become more and more directive as the HRTFs go high in frequency, making the directivity patterns to have more structured features as frequency increases, as shown in Fig. 6.1. This figure shows the directivity patterns for subject 2 for different frequencies. This gives the reader an idea of how the directivity patterns for a given subject evolve as a function of frequency. Although we have a very high resolution of HRTFs when we measure them using BEM simulations, we down-sample the data to create the directivity patterns with sampling the imaginary sphere with equally spaced sensors representing 642 positions, which means a directivity pattern will have only 642 samples. Furthermore, as the HRTFs have very low value at the contralateral side (please refer to Sec. 2.1.1 for more details), here we only study the ipsilateral HRTFs, where the ear has high signal-to-noise ration and does not suffer from head shadows. Because the directivity pattern on the contralateral side can be varied and noisy but is likely not significant [127], we have applied gentle spherical Gaussian smoothing (std.: 5.7 degrees of spherical angle) to the directivity pattern on

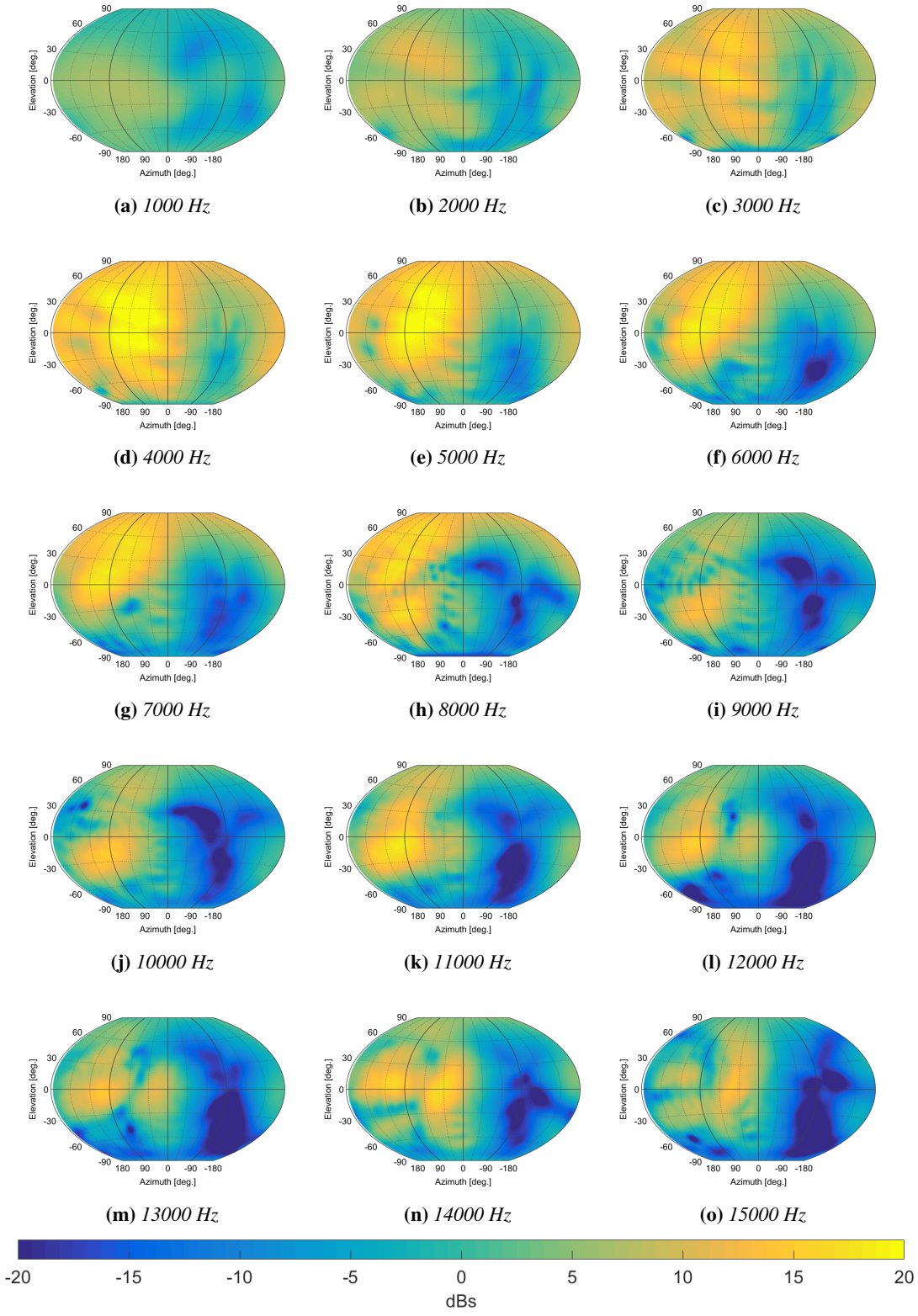
the contralateral side. The directivity pattern data is then treated mathematically as a vector, and standard principal component analysis is applied across subjects for a given frequency. Fig. 6.1, shows the directivity patterns of the affine model of Subject 2 in the SYMARE database.

6.2 Spatial Principal Component Analysis

The mathematical detail on the PCA is provided in Sec. 2.5. PCA has long been used for HRTF modeling. [30, 103, 105, 115, 128, 129], are only few of these studies to mention. Different studies in the past have used the PCA on different modalities of HRTF. For example, some of the studies model the time-domain equivalent HRIRs, while other model the complex magnitudes of the frequency domain HRTFs, and some only model the magnitude responses of the frequency domain data relying on the findings of previous studies that the HRTFs are minimum phase filters and the phase information can be retrieved by simply using the Hilbert transform once the magnitude response is modeled.

In this work, we model the dB scale magnitude responses of the HRTFs, but unlike the previous studies, we model the directivity patterns of the HRTFs. In previous studies, the whole HRTF of a direction is modeled as a single variable while in this work, we model the data for each frequency separately. We think the reason for that is very intuitive and entirely natural. While listening to a sound, we do not simply listen to the sound of a single direction, but our brain actually works on the sound coming from all the directions at once. Furthermore, we move our head to resolve the directional confusion while localizing the sound source as if we are painting an acoustic pattern with our ears. For this reason, we model the directivity patterns instead of modeling the HRTFs of every direction separately. Another reason for doing this is that while data at the lower frequencies are mostly similar to each other, the inter-subject variation really increases in the higher frequencies. A set of sample directivity patterns for different frequencies for a set of subjects is shown in Fig. 6.2 So when modeling the HRTFs for all frequencies at once using PCA, we are required to use more components for the whole range to accurately construct the data for the higher frequencies. Finally, as we have created a synthetic database in which the size and rotation of each of the ear shapes are matched to template ear shape and inherently to each other, following [25] findings the frequencies for every subject would have aligned as well. Hence we can capture most of the variations in the directivity patterns with very few components. We term the PCA for directivity patterns as spatial principal component analysis or SPCA. In this section, we provide a simple analysis of applying the SPCA on the directivity patterns of few frequencies.

As mentioned in Sec. 6.1, that we are using the directivity patterns for the affine matched HRTFs for 642 directions. So the directivity pattern of a given frequency for a subject can be modeled using a sequence of 642 numbers. Having data for 62 subjects, we created matrices of size 624×62 for a given frequency and modeled the data with PCA. Using 62 components will result in a reconstruction with zero loss, while if we use fewer components, we will have some loss. In Fig. 6.3 present actual and reconstructed directivity pattern data for four subjects [3, 49, 30, 52] (from left to right), for four different frequencies 6, 9, 12, and 15 kHz. In all of the following cases, we used only three principal components for reconstructions. The first thing



(a) *1000 Hz* **(b)** *2000 Hz* **(c)** *3000 Hz*

(d) *4000 Hz* **(e)** *5000 Hz* **(f)** *6000 Hz*

(g) *7000 Hz* **(h)** *8000 Hz* **(i)** *9000 Hz*

(j) *10000 Hz* **(k)** *11000 Hz* **(l)** *12000 Hz*

(m) *13000 Hz* **(n)** *14000 Hz* **(o)** *15000 Hz*

Figure 6.1: *Directivity patterns for affine model of subject 2 in SYMARE database for frequencies, 1-15 kHz.*

that I would like to point in this work is the fact that it is using just three principal components. We capture the variance in the directivity patterns of the subjects very well. The results for directivity patterns at 6 kHz show almost perfect reconstruction for the presented subjects while the reconstruction starts to present the sign of a little bit of struggle as we move higher in frequency. Despite these small signs of performance deteriorations, the results still are very impressive. This is probably because all the ears are affine matched to template ear shapes and have the same size, orientation, and position, making it easier to model the directivity patterns with very few principal components, speculation which was made in Ch. 5. Another thing to notice here is for some frequencies; we probably need more components to model the directivity patterns; for example, for 9 kHz, the PCA really struggles with the last two subjects. While for some frequencies like 6 kHz, we might have done a good job even with using fewer components. The next section talks about a simple method to quantify the number of principal components to be used to model the directivity patterns of a given frequency.

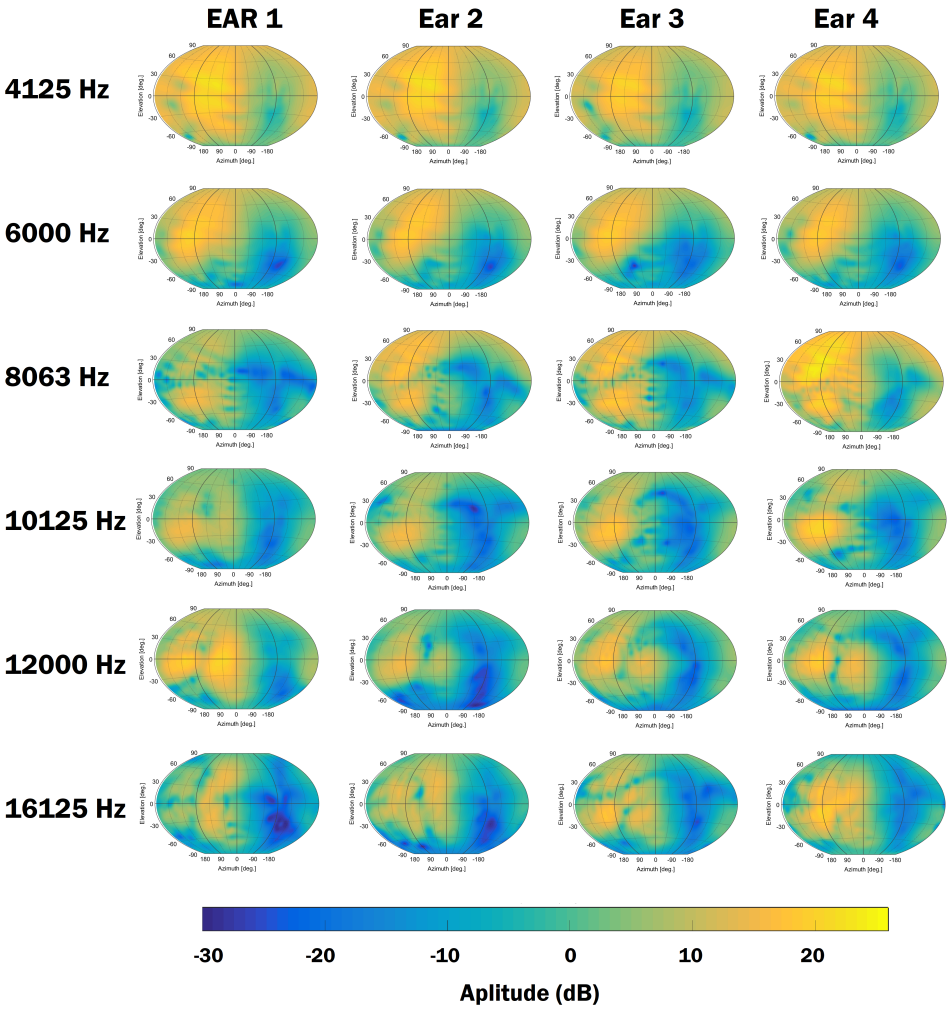


Figure 6.2: *Directivity patterns of four subjects from SYMARE database at 4, 6, 8, 10, 12, and 16 kHz*

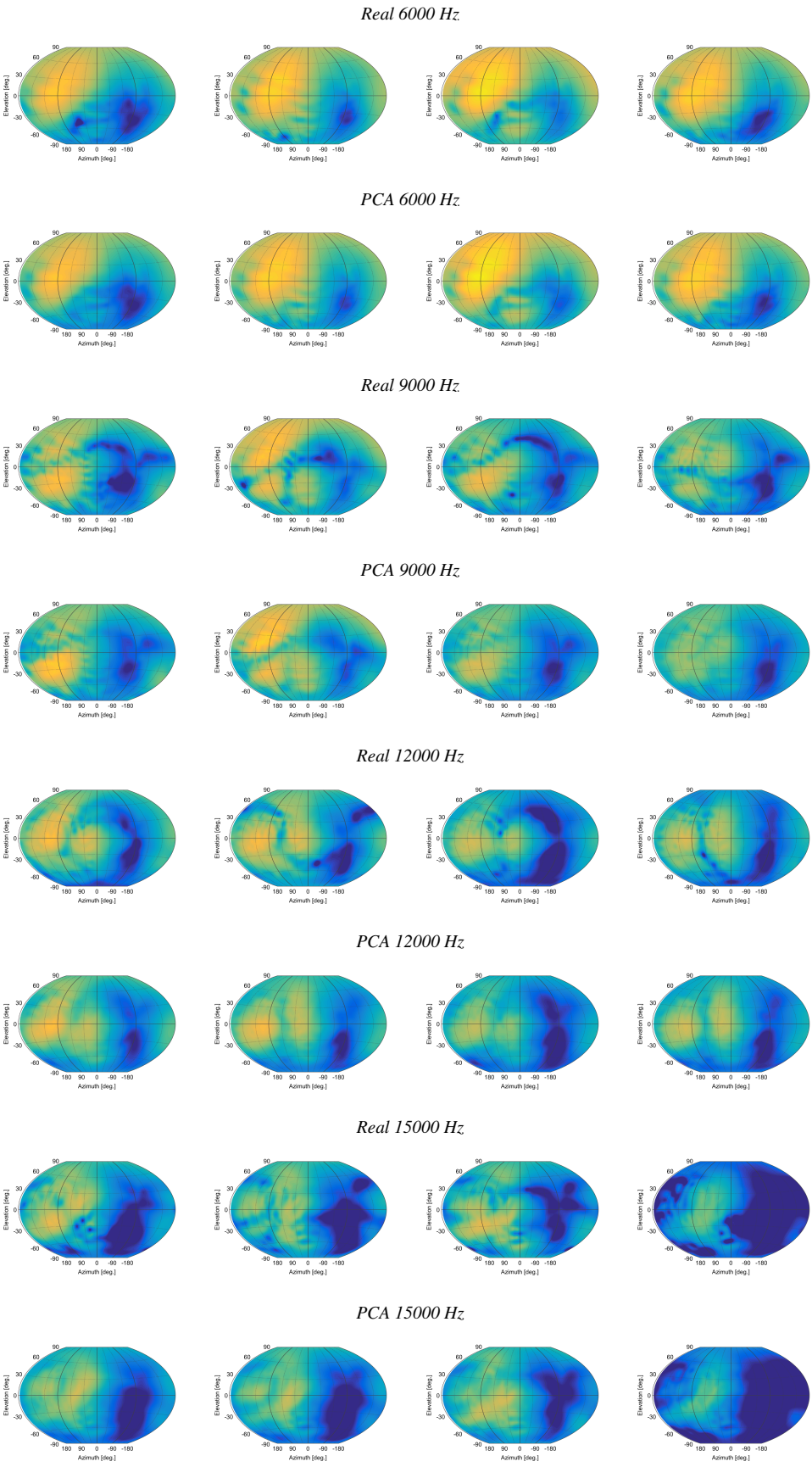


Figure 6.3: *Real and PCA modeled directivity patterns for subject [3, 49, 30, 52] (left to right)*

6.3 Quantification of Number of Parameters Required for Every Frequency

This section provides a simple method to account for the number of principal components required to model the directivity patterns of a given frequency reliably without losing too much information for the sake of using lesser principal components. To find out if we have enough components to model the variations int he data, this study relies on the cumulative captured variance measure used by [103, 115, 130–132]. The cumulative captured variance of PCA is measured by the following expression:

$$cumVar(N) = \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{T} \lambda_i}, \tag{6.1}$$

where, λ_i denotes the i$^{\text{th}}$ eigenvalue of the covariance matrix computed for the PCA, N denotes the number of principal component under consideration and T denotes the total number of principal components. Just like many studies in the past, we decided to keep increasing the number of principal components until we can capture 80 % of the variance in the data. However, one thing to be understood is that to make the PCA more realistic, this study limits the number of the maximum principal component can be used. As we have data for 62 subjects only, we limit ourselves to use eight principal components at max, which is roughly one-eighth of the number of samples we have in hand. In Fig. 6.4, we show a scaled color image plot of the cumulative variance captured for a given frequency when we use a given number of principal components. At the same time, the red line over it indicates the number of principal components required to capture 80 % of the variance, and the green line shows a number of principal components we are going to use as we can not go over eight for data limitation reasons. The line plot in the second row of the figure shows the number of principals components required as a function of frequency. While the third figure reports the standard spectral differences (SSD) between the real and reconstructed directivity patterns for the SYMARE population for a given frequency. The SSD is calculated using the expression provided in Eq. 5.4. The vertical dotted lines indicate the frequencies we analyzed (3, 6, 9, 12, 15, and 17 kHz) in more detail further.

Fig. 6.5 present an overview of the SSD and SPCA study for the directivity patterns at frequency 3 kHz, by digging down a little bit deeper to provide a detailed view of what is happening at this frequency when the reconstructions for the directivity patterns for these subjects are made by using SPCA. The first scattered plot figure in top presents the reader with the SSDs between the real and SPCA reconstructed directivity patterns for all the subjects in the SYMARE population. The values are pretty low and are below 1 dB for all the subjects. This is astonishing when we look at the fact that only the first principal component is being used for the reconstruction. The reason for all this is that we are using the affine models where the head and torso shapes are the same for all the subjects. In the following two lines of the graphs, we present the real and SPCA based reconstructed directivity patterns of four subjects, namely subject (3,49,30,52). The directivity patterns in the third line are the predicted directivity patterns on the basis of morphology. This will be discussed in the next section. Similarly, we present the result for 6, 9, 12, 15 and 17 kHz data in Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig. 6.9, and Fig. 6.6 respectively. These SPCA based directivity patterns at these frequencies are re-

constructed using four, size, eight, eight, and eight principal components, respectively. The results show that even with using only eight principal components, one can model the variations in the directivity pattern at very high frequencies pretty accurately, with average SSDs always less than 3 dBs, all thanks to affine models.

6.4 Personalization of Directivity Patterns

So far, we have seen how to effectively model the HRTFs directivity patterns in a parametric way using SPCA. To create a mapping between the morphology of the subjects and corresponding HRTFs, we need to model the ear shapes as well. For this reason, we rely on the KPCA and LDDMM based morphable ear shape model proposed in [1]. At this point we ask the readers to familiarise themselves with the necessary concepts used in LDDMM, and KPCA specifically in the context of ear shape modelling provided in 3.2, 2.7, and 3.2.2.

In this section, we use these concepts to analyze the morphological variations in the affine matched ear shapes of a few subjects in the SYMARE database. The Fig. 6.11 shows the original and reconstructed ear shapes when only eight KPCA components are used. This figure shows the ability of KPCA based morphable model to capture the ear shape variations quite reasonably in most of the ear shapes even when only the first eight KPCA components are used. However, in some cases, the reconstruction clearly requires more components, for example, fifth and seventh ear shape in the figures (going from left to right). This is due to the reason that these two ears have very different features compared to other ears. For example, ear5 is very different from other ears in the fossa region, while ear7 has a very different shape overall. Again remember all the ear shapes are affinely matched to the template. Hence KPCA can perform well as it only has to capture the shape variation instead of capturing shape, rotation, translation, and scale information. We used these principal components to infer the weights for the principal component of HRTF directivity patterns using simple linear regression. The results for the prediction are showing in the third rows of figures Fig. 6.5, Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig. 6.9, Fig. 6.10.

6.5 Weighted Morphable Model for Ear Shapes

This section explores the potential for morphological weighting of different regions of the ear shape to improve the prediction of HRTF directivity patterns of the listeners. It uses the previously developed LDDMM and KPCA based morphable model [17]; however, it modifies it by applying a weighted kernel principal component analysis to model the pinna morphology. In this process, different regions of the ear shape can be weighted differently before the application of KPCA is applied. This section analyzes the performance of this morphable model for prediction by varying the weights applied to the various regions of the pinna. The results show that this study can not be just used to get a better prediction for the HRTFs, but by varying different weights assigned to different regions, we begin to learn the relative importance of the various regions to the acoustic directivity of the ear shapes as a function of frequency showing us which of the regions in the ear shape are responsible for generation of which spectral cue in the HRTFs.

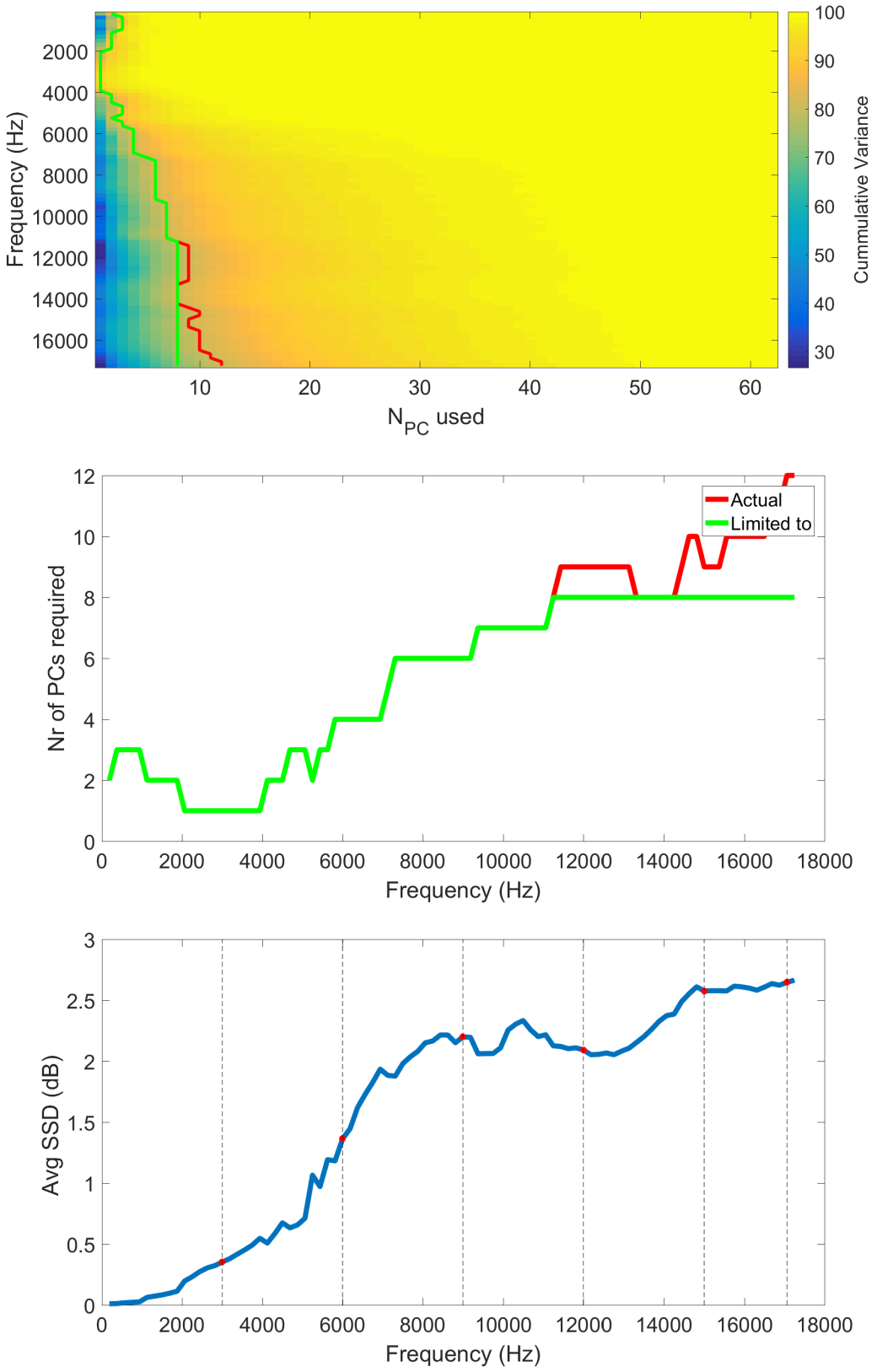


Figure 6.4: *Cumulative captured variance captured when different number of PCs are used/*

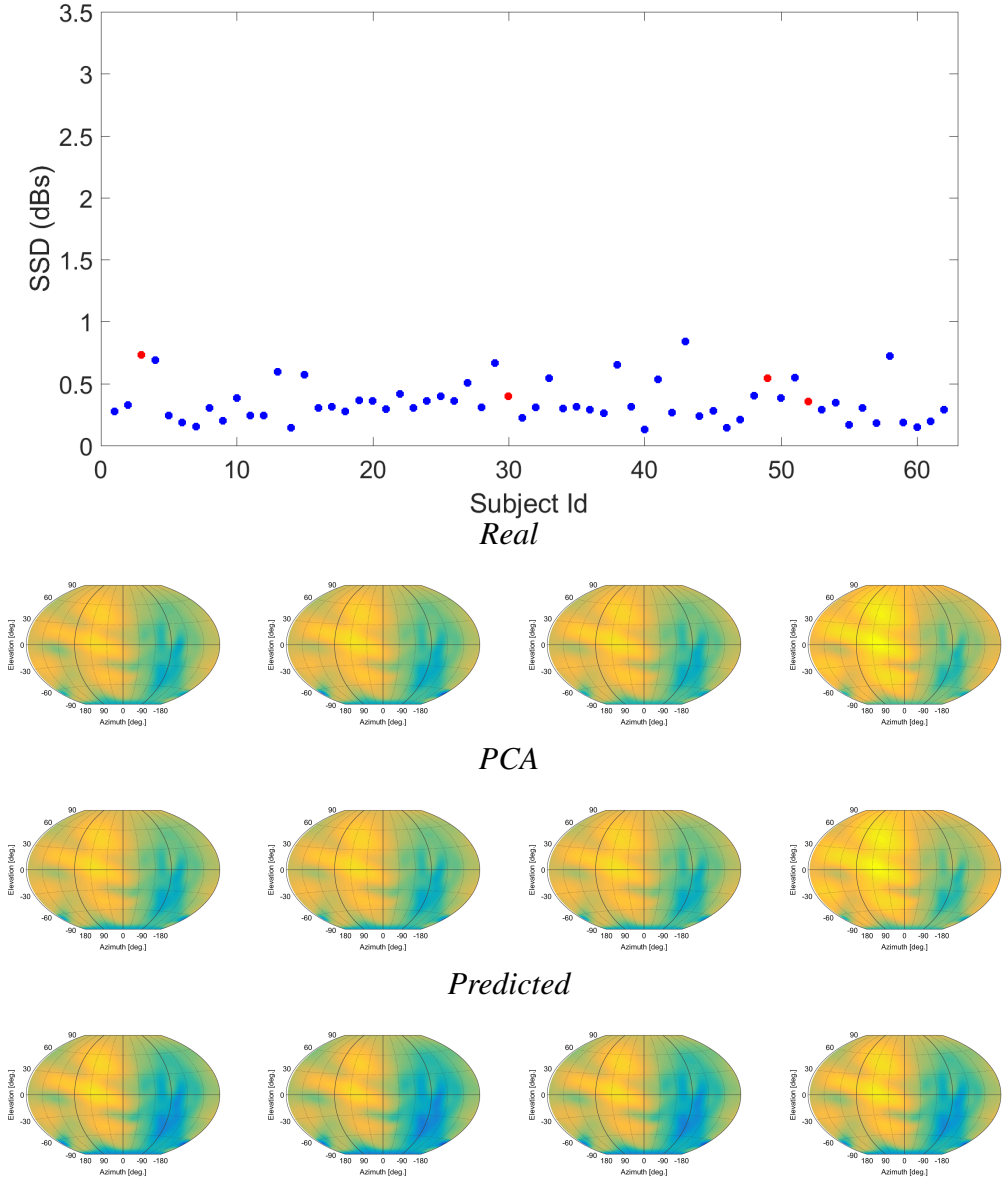


Figure 6.5: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real and PCA based directivity patterns (constructed using a single PC) for 3000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only one principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

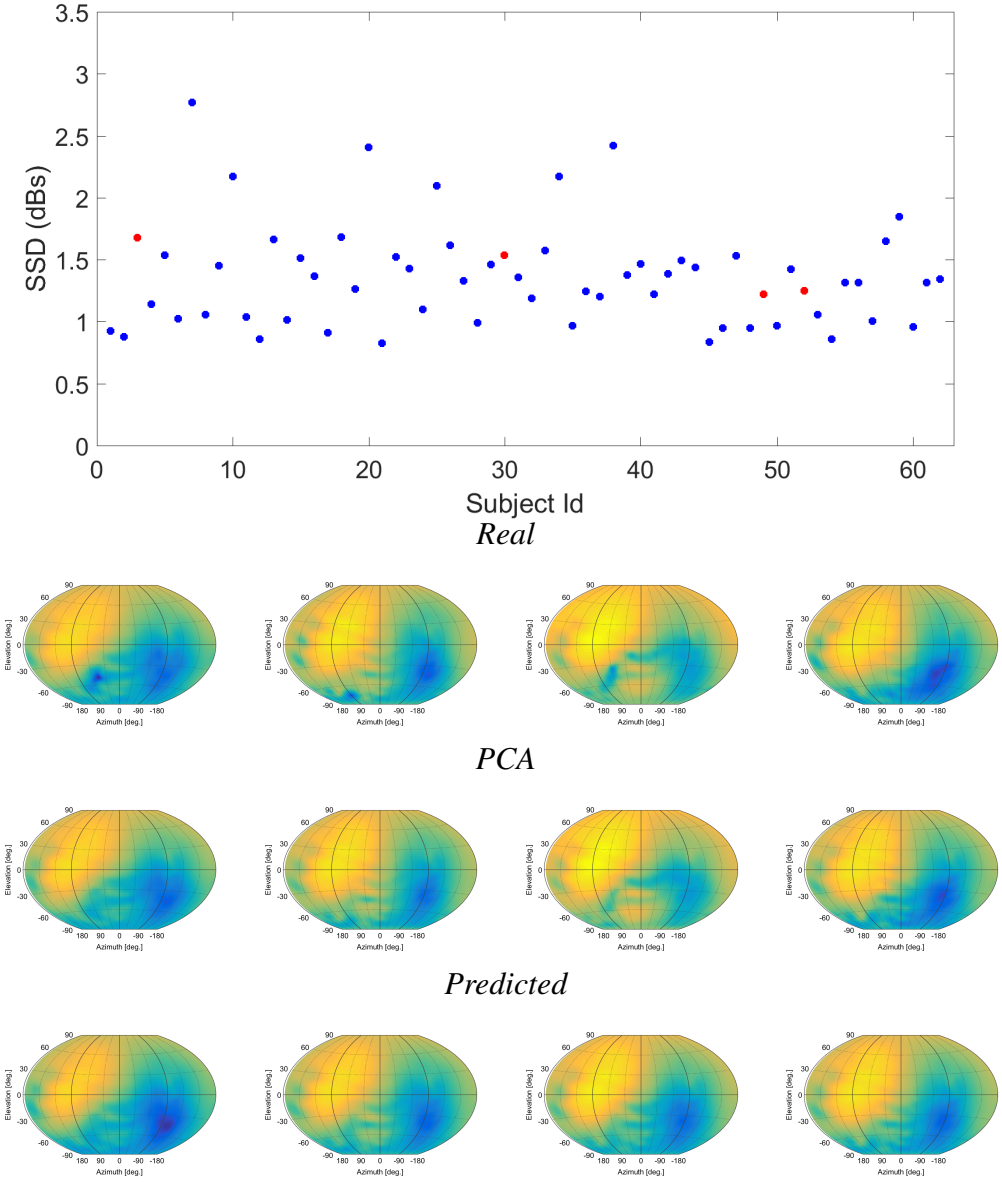


Figure 6.6: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real and PCA based directivity patterns (constructed using first four single PC) for 6000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only four principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

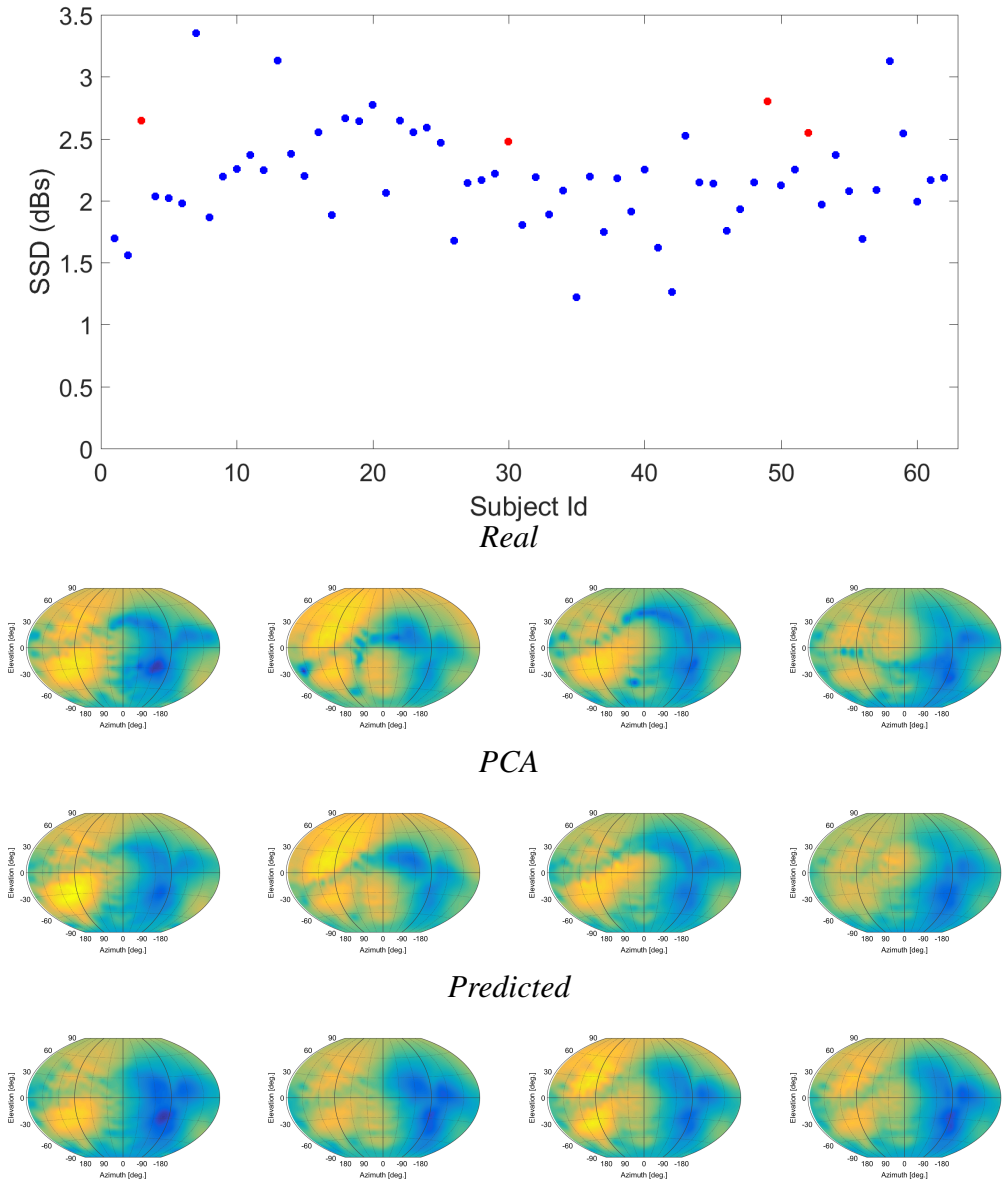


Figure 6.7: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real and PCA based directivity patterns (constructed using first six single PC) for 9000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only eight principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

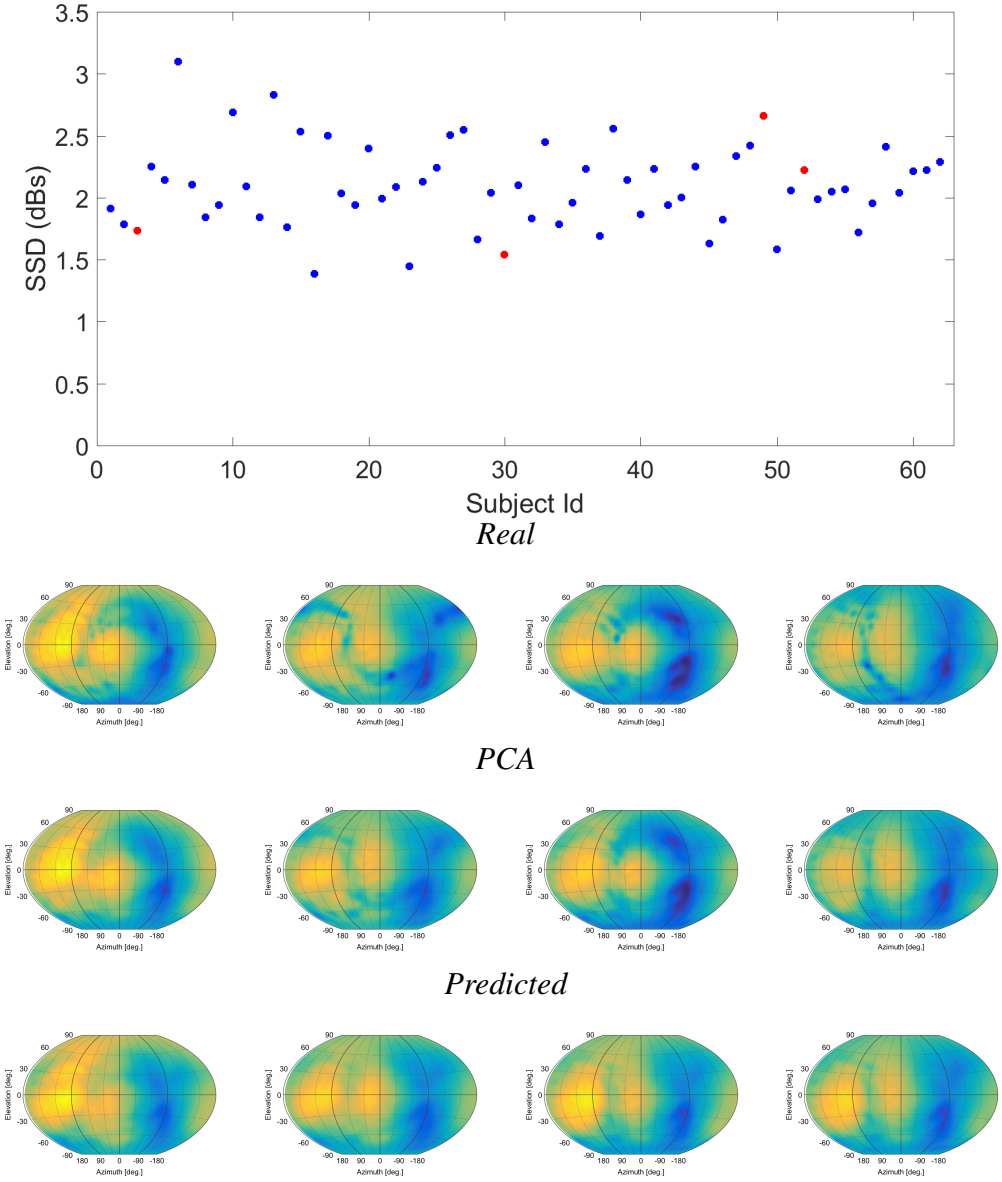


Figure 6.8: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real and PCA based directivity patterns (constructed using first eight PC) for 12000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only first eight principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

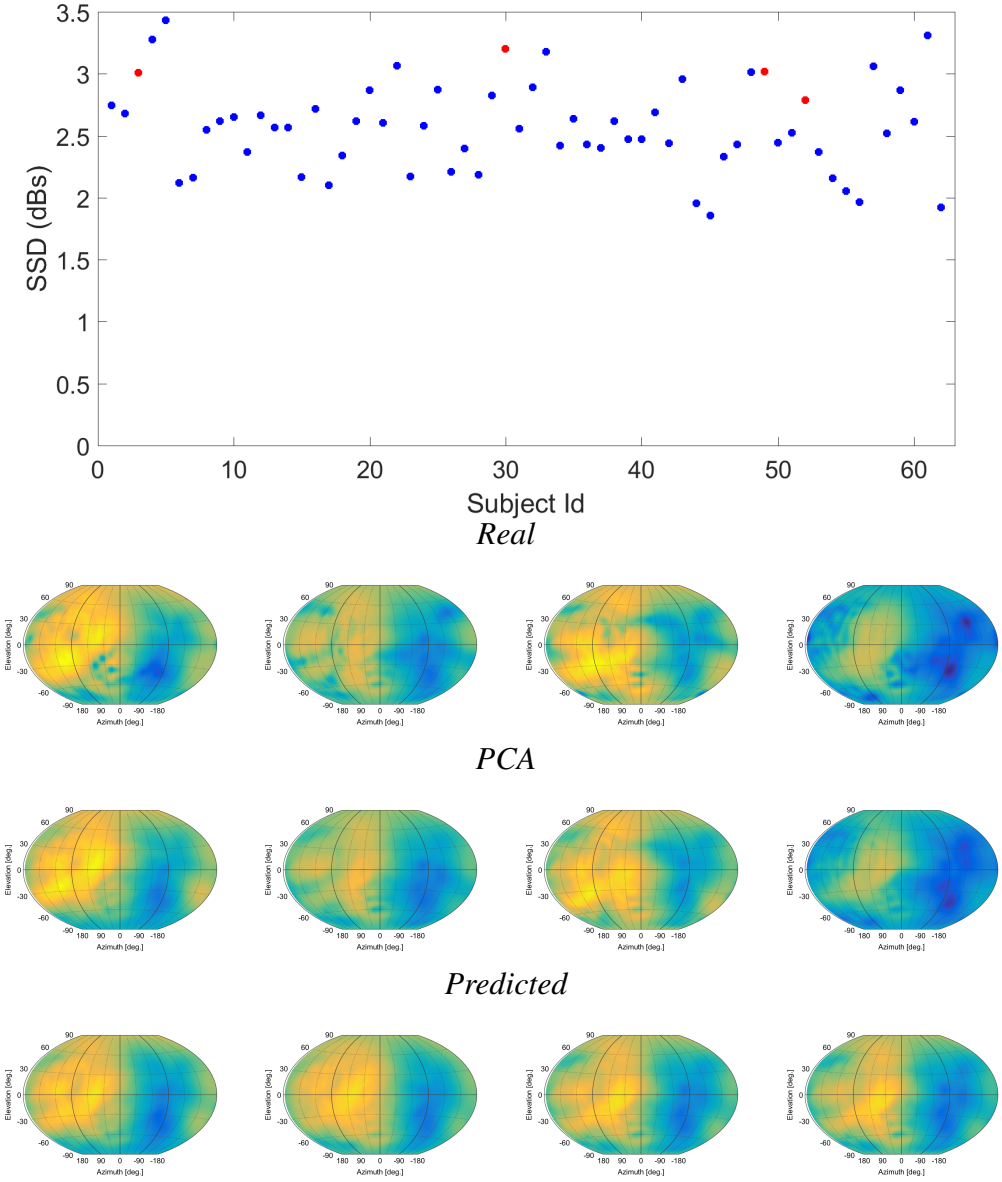


Figure 6.9: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real and PCA based directivity patterns (constructed using first eight PC) for 15000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only first eight principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

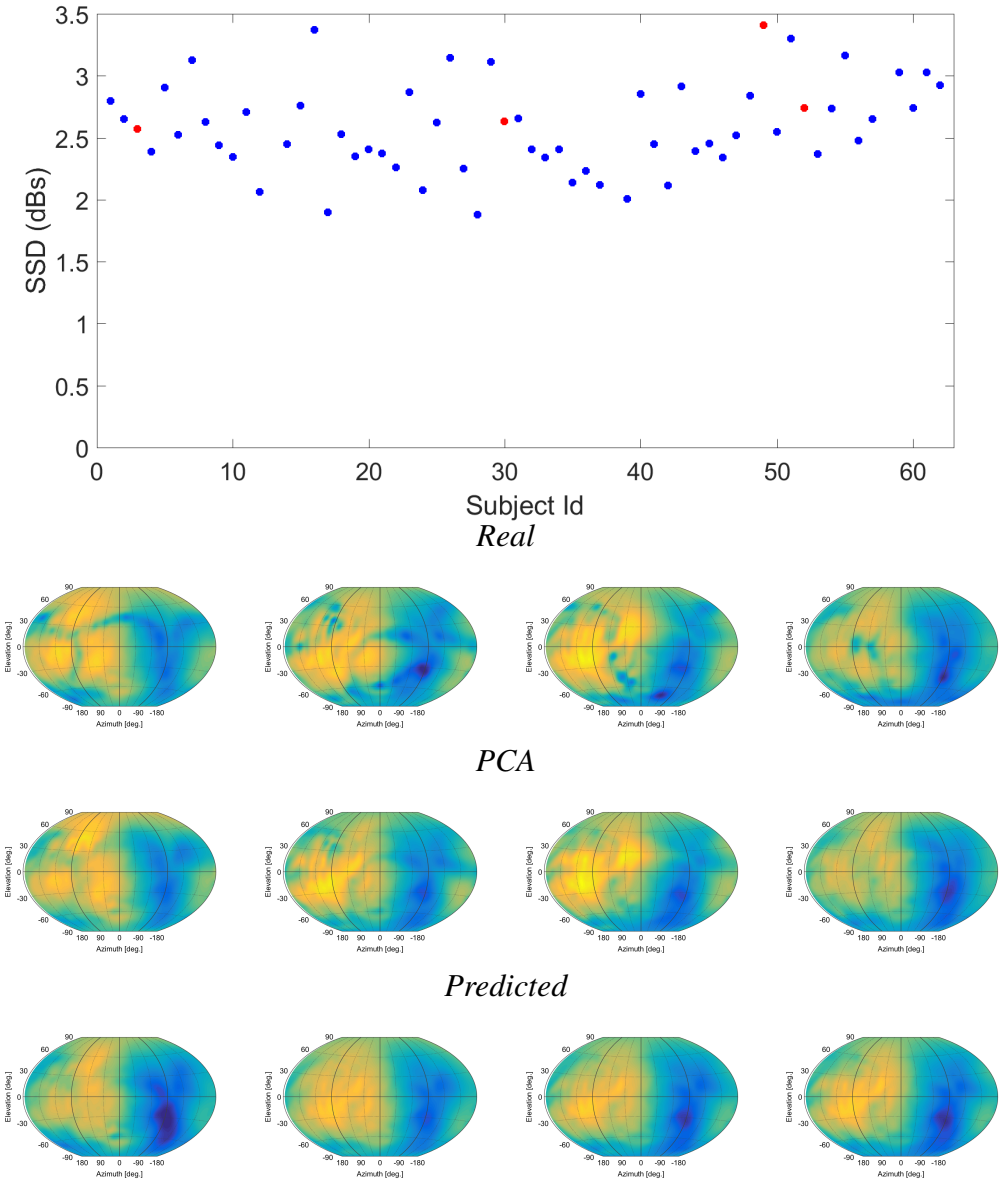
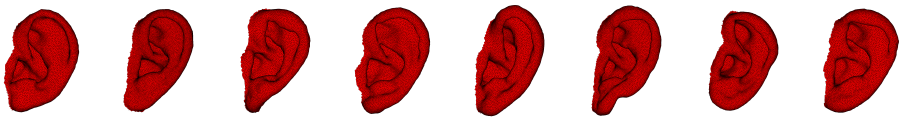


Figure 6.10: *SSD plot for 62, subjects in the top, red stars indicate the subjects analyzed. In bottom two rows we have real, PCA based directivity patterns (constructed using first eight PC), and multiple linear regression based predicted directivity patterns for 17000 Hz for four subjects [3, 49, 30, 52] (from left to right). The results show even with using only first eight principal components one can model the variations in the directivity pattern with average SSDs less than 3 dBs all thanks to affine models.*

Real Ears



KPCA Ears

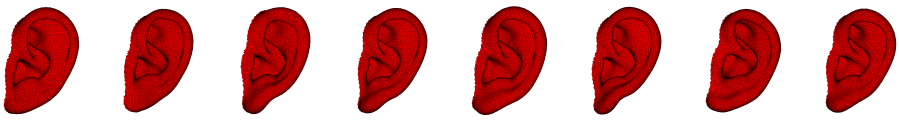


Figure 6.11: *Ear Shapes reconstructed using morphable ear model on the basis of just eight kernal principal components.*

6.5.1 Weighted KPCA (WKPCA)

The morphable model described in [17] has a limitation. While performing the KPCA, it assumes that each of the mesh vertexes contributes equally to the variations and is equally important for the analysis of morphological variations in the ear shapes in the prospect of the acoustic contribution of this vertex. Nonetheless, the previous studies have shown that the relative areas of various regions of the ear do not necessarily accurately represent the importance of their contribution to the acoustic properties of the ear. For example, the back of the ear likely plays a role which is low to no importance when compared with the role played by the concha region . [23]. To explore this issue in more detail, we have apportioned the ear into various sections (refer to Fig. 6.12), which enables a weighting to be applied during the KPCA. In this case, the kernel function in the KPCA is modified, as shown below:

$$k'_V(x,y) = \frac{w(x)w(y)}{1 + \frac{1+\|x-y\|^2}{\sigma_v^2}}, \tag{6.2}$$

where $w(x)$ and $w(y)$ denote the weights for vertices x and y respectively. This way the kernel function does not just depend on the distance between two vertices but also on the weights associated to them.

Ear Portion	Area	Non-Weighted		Concha Weighting		Fossa Weighting	
		w	$w \times$ area	w	$w \times$ area	w	$w \times$ area
Back of Ear	0.25	1.00	0.25	0.75	0.19	0.60	0.15
Ear Lobe	0.05	1.00	0.05	0.75	0.04	0.06	0.03
Scaphoid Fossa	0.03	1.00	0.03	0.67	0.02	5.65	0.17
Helix	0.24	1.00	0.24	0.75	0.18	0.60	0.14
Cymba Concha	0.03	1.00	0.03	1.67	0.05	0.06	0.02
Cavum Concha	0.08	1.00	0.08	1.50	0.12	0.61	0.05
Triangular Fossa	0.05	1.00	0.05	0.80	0.04	5.80	0.29
Concha Rim	0.21	1.00	0.21	1.50	0.32	0.57	0.12
Concha Ridge	0.06	1.00	0.06	0.83	0.05	0.50	0.03

Table 6.1: *Relative vertex weightings, w, and region contributions, $w \times$ area, are shown for three conditions. Note that sum of $w \times$ area column is already unity.*

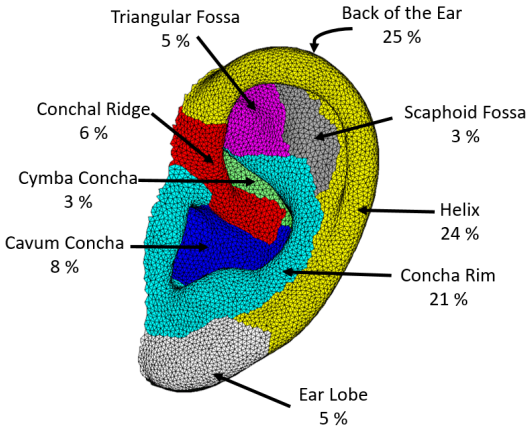


Figure 6.12: *Various regions of the pinna are identified along with their respective fractional contribution to the total surface area.*

6.5.2 Results

We explored this morphological weighting for two regions for all the affine models for the SYMARE population. We on purpose explored two of the most important regions in the intricate shape of the ears namely, the concha and the fossa. In our view the regions with larger surface area have greater influence on the LDDMM mapping algorithm (in original morphable model), we determined the surface area for each of the regions created in the division of the ear (as shown in Fig. 6.12. The relative contribution of the region's contribution is then calculated by multiplying the weighting assinged to the area with the relative contribution of the area of the region. In our view this provides a more realistic way of assigning the contributions to different regions. At this stage, we only used a simple approach to morphological weighting: e.g., a region's contribution was multiplied by a small, arbitrary factor to see if this weighting creates some improvements. As per our analysis a moderate weighting factor is gen-

erally proved to better for regions which contribute more towards the area while the smaller regions require a larger morphological weights to create a significant impact in driving the weighted morphable model as is shown in Table 6.1. Another thing to be noticed is that all the weights are normalized so that the region contributions (region area multiplied by morphological weight) always sum up to unity.

The impact of the morphological weighting was measured simply using linear regression. In this work, we have only considered the first principal component for the acoustic directivity patterns. While this assessment is limited, it is important to keep in mind that the ears are affine-matched, and only a few principal components are required to adequately describe the acoustic directivity patterns, as shown in Sec. 6.3. So a given morphological weighting was evaluated by applying linear regression to find the best linear relationship between the first principal components of the acoustic data to the eight principal components of the weighted KPCA based LDDMM ear model. Example results are shown in Fig. 6.13. In this case, we examine a relatively low frequency (approximately 4 kHz) and find that morphological weighting applied to the concha makes a small improvement.

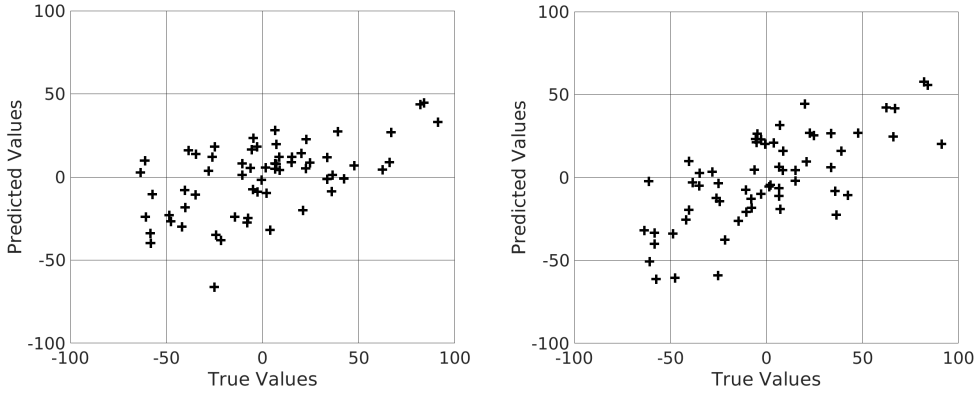


Figure 6.13: *Scatter plots show the predicted and true values for the first principal component of the acoustic directivity patterns corresponding to a frequency of 3938 Hz. Plots are shown for data both without (a) and with (b) morphological weighting. The respective R^2 values are 0.32 and 0.52.*

What is much more interesting in this study is that this simple study lets us analyze the influence of morphological weighting as a function of frequency. An analysis of this kind is shown in Fig. 6.14. In this analysis, we applied the linear regression-based prediction model to predict the first principal component of the directivity patterns for all the subjects for every frequency. Prediction errors were then measured in units of one standard deviation for the whole SYMARE population data. Fig. 6.14a), shows the improvements as a result of the concha weighting given in Tab. 6.1. We see that concha weighting results in improvements at various frequencies around 3 kHz, 7 kHz, 10 kHz, and 13 kHz. We interpret the broad range of frequencies as indicating the concha may influence resonance modes at various frequencies. To further support these findings, we examined the percentage of cases with improvements and found a similar pattern across many subjects (see Fig. 6.14b). The morphological weighting for the fossa produced similar results, albeit at slightly higher frequencies (refer to Figs. 6.14c and 6.14d. We do not intend for these data to indicate the concha and fossa play independent roles. Rather, the morphological weighting enables one to explore at which frequencies a

particular region of the ear may have a particular influence on the acoustic directivity of the ear.

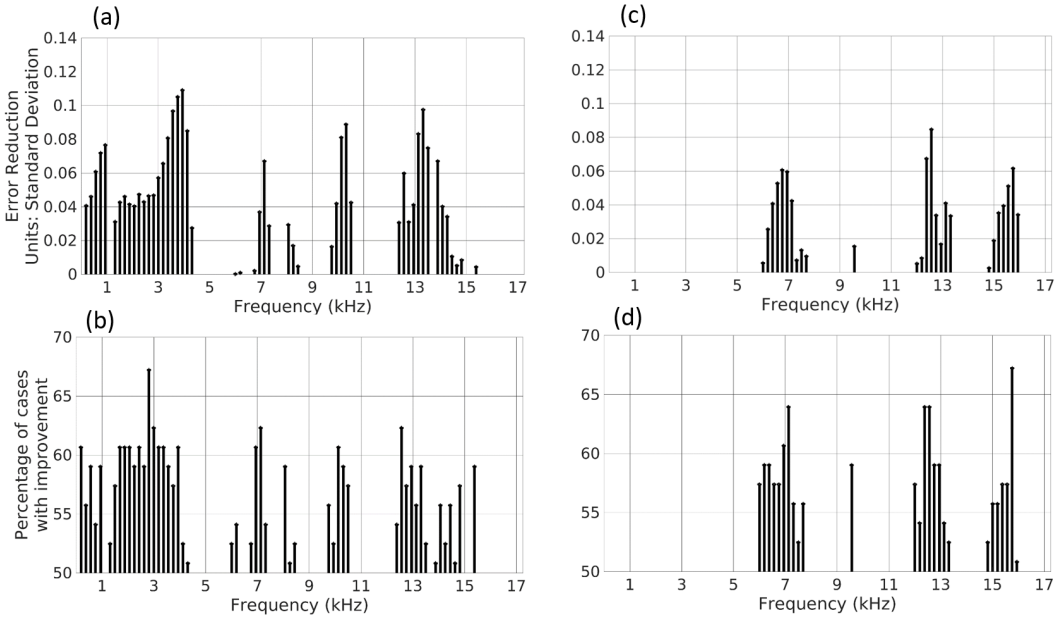


Figure 6.14: *Stem plots show the impact of morphological weighting as a function of frequency. Data for the concha are shown in (a) and (b), while data for the fossa are shown in (c) and (d). The mean reduction in prediction error is shown in (a) and (c) using the population standard deviation as a unit measure. The percentage of ears for which the prediction improved is shown in (b) and (d).*

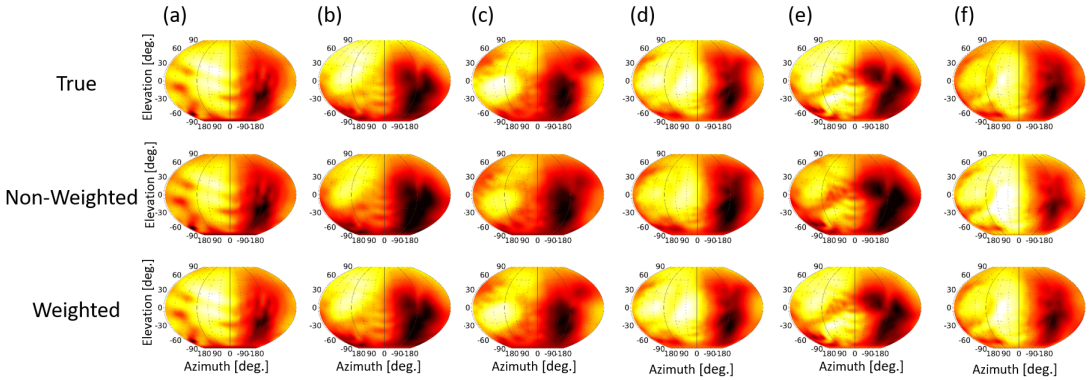


Figure 6.15: *Changes in the acoustic directivity patterns that occur based on the prediction of the first principal component are shown. Azimuth and elevation angles are shown in degrees. The top row shows the true data; the second row shows the data* without *morphological weighting, and the third row shows the data* with *morphological weighting. Data are shown for the concha at frequencies: (a) 3938 Hz; (b) 7125 Hz; (c) 10312 Hz; and (d) 13313 Hz. Data are shown for the fossa at frequencies: (e) 6938 Hz and (f) 12563 Hz. Best viewed in color online to see subtle differences.*

The influences of the improvement in the prediction of the first principal component on the resulting acoustic directivity patterns are shown in Fig. 6.15. Because we only explore the first principal component, all other principal components are held fixed at their true values. We find that the improved prediction of the first principal component

does result in small, but visible improvements in the acoustic directivity patterns.

6.6 Discussion and Conclusion

In this chapter, we provided a simple method of parametrical modeling of the acoustic directivity patterns of the HRTFs. As we are modeling the spatial directivity patterns, we called this PCA as spatial principal component analysis or SPCA. We used SPCA to analyze the variations in the directivity patterns as a function of frequency. Furthermore, using a simple measure of captured cumulative variance, we quantified the number of principal components required for each of the frequencies. The results show that using only eight principal components; one can model the directivity patterns of even very high frequencies around 4 kHz pretty accurately. The performance of the SPCA was also evaluated using SSD, and the results show that even when we are modeling the HRTF directivity patterns with only eight principal components average, SSD never increases from 3 dBs, which is pretty impressive for high-frequency content. Furthermore, we showed a simple way of the HRTF personalization method based on linear regression. We presented a simple analysis of the KPCA based morphable model for the ear shapes and shown how using only eight principal components can effectively model the variations in the ear shapes. We used these eight morphology principal components and using simple linear regression derived the PCA weights for the acoustics. The results show that even with this simple prediction method, we can efficiently predict the directivity patterns. Finally, we used all the tools used in the chapter so far and used them to propose a simple yet powerful tool based on morphological weighting. Morphological weighting is shown to provide an interesting tool to explore the morphoacoustic properties of the human outer ear. However, it is worth to mention that at this stage, our understanding is limited. Each frequency and each principal component may find improvements with different morphological weightings. We do not find this unreasonable because the acoustic properties of the outer ear result from the structure as a whole, and the strength of any particular resonance mode may result from complicated interactions between various morphological elements. We have not yet explored a general optimization algorithm for morphological weighting, nor explored whether additive combinations of morphological weightings would make any sense. It is not even clear how many physical regions one should divide the ear into, nor what the possible interactions may be. Further, it is not yet clear whether a particular morphological weighting should be applied for all ears or just a particular class of ears. Nonetheless, we have made a start and believe there is much more to be learned and will so direct our future attention.

CHAPTER 7

Conclusion and Future Work

This thesis provides a set of studies to understand the relationship between the morphology of the ear shapes and its relationship to the corresponding individualized acoustic filters called HRTFs. Based on the findings of the most recent works, this thesis work studied two different paradigms for HRTF personalization. One of these paradigms relies on the anthropometric parameters, while the other uses a morphoacoustic approach to understand the underlying complex phenomenons generating the HRTFs.

This thesis work started by providing the reader with the introduction to the study under review with a brief background of the problem along with motivation and problem statement and contributions in Ch. 1. In Ch. 2, it provided a general background with the required important concepts which are essential to understand the work performed in this thesis. Ch. 3 provided a comprehensive review on state of the art. The next three chapters from Ch. 4 to Ch. 6 are the contribution chapter of this thesis in which answers to the following questions are explored.

Question1: Given the findings of [13, 23] one knows that the notch frequencies for CIPIC on average have a monotonically increase as the elevation angle increases and fall in the range of 6 kHz to 9 kHz, 10 kHz to 12 kHz, and 13 kHz to 14 kHz respectively, do the notches in the SYMARE database evolve similarly? Are the notch frequency evolutions symmetric for left and right ears?

Answer: Given this question, this thesis performs a statistical analysis on the notch frequencies of the median plane for elevation angles from -45° to 45° for CIPIC and SYMARE databases. The results of this analysis show that the evolution of notches is consistent in both databases. Also, the notch frequency ranges are in agreement with the results of [23]. The comparative analysis for two ears show that the notch frequencies for left and right ear are not symmetric and although the differences are not huge, they still suggest that binaural cues can also play a part in elevation plane localization

Question2: Given a simple, sparse representation-based approach for HRTF personal-

ization and having a hypothesis that different anthropometric features have a different amount of importance when it comes to personalization of HRTFs, how to calculate the weights for these anthropometric parameters, and use them to optimize and improve the sparse representation?
Answer: Pursuing this question a simple study was conducted presented in Sec. 4.2. The weights for every anthropometric feature is computed by using an extensive partially on-off based HRTF personalization technique. These important factors are then used to calculate the weighted-sparse representation of the anthropometric features and then to calculate and synthesize the HRTFs. The results show that using the weighted sparse representation not just improves the result for the performance but also requires a fewer number of anthropometric features compared to the traditional sparse representation based approaches. This study also identifies and uses the only anthropometric features which can be easily obtained from a set of three scaled images as proposed by [15]. The weights calculated for the anthropometric features can be used for various other studies.
Question3: Can the variations in the ear shape geometry, scale, and rotation separately be studied? What are the effects of affine transformations of the ear shapes on the corresponding acoustics? Can the effects of these affine transformation be modeled or corrected without redoing the BEM simulations on the 3D affine model? What are the frequency ranges which are affected by the head and ear scaling, and can an optimal scaling be found to fix the problems in the whole HRTF frequency range?
Answer: To answer these questions, performed a comprehensive study in Ch. 5. We created a synthetic database using the shapes in the SYMARE dataset using the LDDMM framework. This is a big contribution to this work and will be added to the SYMARE database in the future. This is a unique dataset, where all the ear shapes were affinely matched to match the scale, rotation, and position with the multi-scale template ear shape. The results show that using the affine model simplifies the acoustics by 10%. The matching between the affine matched HRTFs, and original HRTFs can be improved by almost 29% on average using simple corrections such as frequency scaling and rotation of the directivity patterns. Furthermore, the used scaling and rotation matrices can be easily obtained from the rotation and scaling of the head and ear shapes. These ear shapes are called affine matched ear shapes.
Question4: Having this set of simplified HRTFs how to model the inter-subject variations in the HRTFS as a function of frequency?
Answer: We performed a simple study based on PCA to model the inter-subject variations amongst the HRTF directivity patterns of SYMARE subjects as a function of frequency. As we are modeling the directivity patterns, which are like spatial surfaces, we termed this PCA as Spatial Principal Component Analysis. The results of this study show indeed, the affine model helps one to mode the HRTFs in a very simple way. We are able to model the directivity patterns of the HRTFs even at 17 kHz with only eight principal components with a standard spectral difference (SSD) error of less than 3 dBs on average. This study also uses the extracted parameters of the directivity patterns and create a simple mapping between the eight parameters of the morphable ear shape model for every affine matched ear in the SYMARE population using multiple linear regression (MLR). This provides a simple approach for the personalization of the HRTFs.

With this question answered, we have a full end to end scheme for HRTF personalization, which provides one with the personalized HRTF of a given subject with the shape of the ear in hand without running the lengthy and power-hungry BEM simulations. However, we studied another interesting question in this thesis study with the quest to improve the personalization method.

Question5: Having the KPCA, and LDDMM based morphable and parametric model of the ear shapes in hand along with the linear regression-based HRTF personalization method, can we have something like a weighted morphable model for ear shapes using weighted KPCA similarly like weighted sparse representation to improve the performance of the personalization?

Answer: It is well understood that different regions in the ear shapes play a different role in producing the spectral coloration modeled by HRTFs. For example, the back of the ear plays a much simpler role in the prediction of HRTFs when compared to the concha or foss cavity. This gave us a notion that there is an inherent limitation in the traditional KPCA based parametric model the ear shapes, which considers all the vertices in the 3D mesh of the ear shapes to be equally important. To change that we hand cut different parts of the ear shapes and created a simple weighted KPCA based ear shape model, which allows one to assign higher weights to a given region of the ear shape and study its effects on the corresponding acoustics obtained through either personalization procedure or through BEM simulations. What is more interesting is to look at the frequency wise improvements suggesting the relevance of each of the ear region in the generation of the acoustic cues. This provides us a simple yet powerful technique that can be used as a variant for morphoacoustic perturbation analysis.

The following section highlights the challenges faced while conducting these studies.

7.1 Challenges

Running the simulations to study the variations in the morphology and acoustics of the outer ear shapes at this scale was a challenging task. Although the advent of multi-core computers, large memory space and improvements in the simulations techniques, has made the numerical simulations for HRTFs a realistic task, by cutting down the simulation times multi-fold, but still simulating an FM-BEM simulation on a high-resolution head-torso-ear mesh with thousands of the triangular faces for frequencies up to 24kHz and with a spatial resolution of 3° takes almost a day for one subject for one ear shape. If only the head and ear mesh are used, removing the torso mesh, this simulation time is cut to almost half. However, for the sake of completion, we did not do this. This essentially means that just to run the BEM simulations for our new synthetic database, it took us over two months.

The other challenging task was to identify the portions of the ears which work the best for weighted KPCA for a new morphable model for ear shapes and to assign weights to these portions. The simulations and visual analysis for one weight combination can very easily take more than one day. The incisions for cutting the ear shapes were performed with hand and took us multiple iterations to find the right splits. However, to identify the right combinations which work for most of the ears and shows considerable improvements in capturing the variations in Concha, Conchal Ridge, and Triangular Fossa took us many runs and was a very challenging task.

7.2 Future Work

The future works for this thesis include studying the effects of affine transformations on the head and torso shapes as well. In our study, we only used the rotation data coming from the affine matching of the ear shapes, which is shown to harm some subjects for the frequency range below 5 kHz. This demands the understanding of how the corrections for rotation are to be performed for these frequency ranges. For the study of HRTF personalization and modeling, there is a need to perform psycho-perceptual tests, which not just only evaluate the performance of the SPCA based modeling but also aid in the evaluation and refinement of the personalization methods.

In the study of a weighted KPCA based morphable model, this study demands a way to find the optimal weighting for all the regions, which can only be assigned once the contributions of each of the regions are understood. Here again, the psycho-perceptual tests can be of great aid. Once the optimal scalings are chosen, this can provide good personalization results even when only a small number of parameters are used in modeling the ear shapes. Another thing that needs to be explored here is the use of weighted LDDMM, which lets one different match surface focussing more on a given region of the space while calculating the matching.

Bibliography

[1] R. Zolfaghari. *Large Deformation Diffeomorphic Metric Mapping Provides New Insights Into the Link between Human Ear Morphology and the Head-related Transfer Functions*. PhD thesis, School of Electrical and Information Engineering, The University of Sydney, Australia, 2016.

[2] V. Algazi, C. Avendano, and R. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122, 2001.

[3] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida. Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *The Journal of the Acoustical Society of America*, 132(6):3832–3841, 2012.

[4] G. Romigh, D. Brungart, R. Stern, and B. Simpson. Efficient real spherical harmonic representation of head-related transfer functions. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):921–930, 2015.

[5] R. Zolfaghari, N. Epain, C. Jin, A. Tew, and J. Glaunès. A multiscale lddmm template algorithm for studying ear shape variations. In *Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on*, pages 1–6. IEEE, 2014.

[6] V. Algazi, R. Duda, R. Duraiswami, N. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064, 2002.

[7] R. Teranishi and E. Shaw. External-ear acoustic models with simple geometry. *The Journal of the Acoustical Society of America*, 44(1):257–263, 1968.

[8] L. Poveda and R. Meddis. A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, 100(5):3248–3259, 1996.

[9] Y. Tao, A. Tew, and S. Porter. A study on head-shape simplification using spherical harmonics for hrtf computation at low frequencies. *Journal of the Audio Engineering Society*, 51(9):799–805, 2003.

[10] C. Hetherington and A. Tew. Parameterizing human pinna shape for the estimation of head-related transfer functions. In *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

[11] A. Tew, C. Hetherington, and J. Thorpe. Morphoacoustic perturbation analysis: principles and validation. In *Acoustics 2012*, 2012.

[12] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Acoustic sensitivity to micro-perturbations of kemar's pinna surface geometry. In *Proc. Int. Congress on Acoustics*, volume 8, 2010.

[13] V. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, 118(1):364–374, 2005.

[14] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The cipic hrtf database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, 2001*, pages 99–102. IEEE, 2001.

[15] E. Gallegos, F. Bustamante, and F. Arámbula-Cosío. Personalization of head-related transfer functions (hrtf) based on automatic photo-anthropometry and inference from a database. *Applied Acoustics*, 97:84–95, 2015.

[16] C. Jin, A. Corderoy, S. Carlile, and A. Schaik. Contrasting monaural and interaural spectral cues for human sound localization. *J. of the Acoust. Soc. of Am.*, 115(6):3124–3141, 2004.

[17] R. Zolfaghari, N. Epain, C. Jin, J. Glaunès, and A. Tew. Generating a morphable model of ears. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 1771–1775. IEEE, 2016.

[18] M. Gardner and R. Gardner. Problem of localization in the median plane: effect of pinnae cavity occlusion. *The Journal of the Acoustical Society of America*, 53(2):400–408, 1973.

[19] B. Cunningham. Applications of virtual auditory displays. In *Proceedings of the 20th international Conference of the IEEE Engineering in Biology and Medicine Society*, volume 20, pages 1105–1108. Citeseer, 1998.

[20] E. Wenzel, F. Wightman, and S. Foster. A virtual display system for conveying three-dimensional acoustic information. In *Proceedings of the Human Factors Society Annual Meeting*, volume 32, pages 86–90. SAGE Publications Sage CA: Los Angeles, CA, 1988.

[21] H. Moller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3):171 – 218, 1992.

[22] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2408–2411. IEEE, 2011.

[23] S. Spagnol, M. Geronazzo, and F. Avanzini. On the relation between pinna reflection patterns and head-related transfer function features. *IEEE transactions on audio, speech, and language processing*, 21(3):508–519, 2013.

[24] J. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3):1493–1510, 1999.

[25] J. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3):1480–1492, 1999.

[26] J. Middlebrooks, E. Macpherson, and Z. Onsan. Psychophysical customization of directional transfer functions for virtual sound localization. *The Journal of the Acoustical Society of America*, 108(6):3088–3091, 2000.

[27] M. Shahnawaz, L. Bianchi, A. Sarti, and S. Tubaro. Analyzing notch patterns of head related transfer functions in cipic and symare databases. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 101–105. IEEE, 2016.

[28] P. Bilinskir, J. Ahrens, M. Thomas, I. Tashev, and J. Platt. Hrtf magnitude synthesis via sparse representation of anthropometric features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4468–4472. IEEE, 2014.

[29] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti. HRTF personalization based on weighted sparse representation of anthropometric features. In *2017 International Conference on 3D Immersion (IC3D)*, pages 1–7. IEEE, 2017.

[30] C. Jin, R. Zofaghari, X. Long, A. Sebastian, S. Hossain, J. Glaunés, A. Tew, M. Shahnawaz, and A. Sarti. Considerations regarding individualization of head-related transfer functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*. IEEE, 2018.

[31] M. Shahnawaz, C. Jin, J. Glaunès, A. Sarti, and A. Tew. Morphological weighting improves individualized prediction of hrtf directivity patterns. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 75–79. IEEE, 2019.

[32] R. Zolfaghari, N. Epain, C. T. Jin, J. Glaunès, and A. Tew. Kernel principal component analysis of the ear morphology. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485, March 2017.

[33] J. Blauert. *The technology of binaural listening*. Springer, 2013.

[34] A. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[35] D. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.

[36] B. Xie. *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

[37] L. Rayleigh. XII. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[38] F. Wightman and D. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

[39] E. Macpherson and J. Middlebrooks. Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002.

[40] R. Algazi, R. Duda, and P. Satarzadeh. Physical and filter pinna models based on anthropometry. In *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

[41] A. Hunter, J. Frias, G. Kaesbach, H. Hughes, K. Jones, and L. Wilson. Elements of morphology: Standard terminology for the ear. *American Journal of Medical Genetics Part A*, 149(1):40–60, 2009.

[42] J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974.

[43] R. Butler and K. Belendiuk. Spectral cues utilized in the localization of sound in the median sagittal plane. *The Journal of the Acoustical Society of America*, 61(5):1264–1269, 1977.

[44] S. Spagnol, M. Geronazzo, and F. Avanzini. Structural modeling of pinna-related transfer functions. In *In Proc. Int. Conf. on Sound and Music Computing (SMC 2010)*, volume 34, 2010.

[45] F. Asano, Y. Suzuki, and T. Sone. Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America*, 88(1):159–168, 1990.

[46] E. Langendijk and A. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, 2002.

[47] C. Cheng and G. Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. In *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.

[48] J. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5):2607–2624, 1992.

[49] C. Cheng and G. Wakefield. Spatial frequency response surfaces: an alternative visualization tool for head-related transfer functions (HRTFs). In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 2, pages 961–964 vol.2, Mar 1999.

[50] P. Majdak, P. Balazs, and B. Laback. Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8):623–637, 2007.

[51] C. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. Schaik, A. Tew, C. Hetherington, and J. Thorpe. Creating the sydney york morphological and acoustic recordings of ears database. *IEEE Transactions on Multimedia*, 16(1):37–46, 2014.

[52] N. Gumerov, R. Duraiswami, and D. Zotkin. Fast multipole accelerated boundary elements for numerical computation of the head related transfer function. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–165. IEEE, 2007.

[53] Y. Kahana and P. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of sound and vibration*, 300(3-5):552–579, 2007.

[54] G. Chertock. Sound radiation from vibrating surfaces. *The Journal of the Acoustical Society of America*, 36(7):1305–1313, 1964.

[55] J. Hallett. Climate change 2001: The scientific basis. edited by jt houghton, y. ding, dj griggs, n. noguer, pj van der linden, d. xiaosu, k. maskell and ca johnson. contribution of working group i to the third assessment report of the intergovernmental panel on climate change, cambridge university press, cambridge. 2001. 881 pp. isbn 0521 01495 6. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 128(581):1038–1039, 2002.

[56] M. Joshi, N. Gupta, and L. Hmurcik. Modeling of pinna related transfer functions (prtf) using the finite element method (fem). In *COMSOL conference*, 2013.

[57] F. Ma, J. Wu, M. Huang, W. Zhang, W. Hou, and C. Bai. Finite element determination of the head-related transfer function. *Journal of Mechanics in Medicine and Biology*, 15(05):1550066, 2015.

[58] P. Fiala, J. Huijssen, B. Pluymers, R. Hallez, and W. Desmet. Fast multipole bem modeling of head related transfer functions of a dummy head and torso. In *ISMA 2010 Conference, Leuven, Belgium*, 2010.

[59] N. Gumerov, A. O'Donovan, R. Duraiswami, and D. Zotkin. Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. *The Journal of the Acoustical Society of America*, 127(1):370–386, 2010.

[60] ACVD. http://www.creatis.insa-lyon.fr/site/en/acvd.

[61] H. Ziegelwanger, P. Majdak, and W. Kreuzer. Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization. *The Journal of the Acoustical Society of America*, 138(1):208–222, 2015.

[62] MeshLab. http://www.meshlab.net/.

[63] Coustyx Ansol. http://ansol.us/products/coustyx/.

[64] M. Wenninger. *Polyhedron models*. Cambridge University Press, 1974.

[65] W. Gardner. *3-D audio using loudspeakers*, volume 444. Springer Science & Business Media, 1998.

[66] S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hearing research*, 114:179–196, 1997.

[67] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on multimedia*, 6(4):553–564, 2004.

[68] D. Begault, E. Wenzel, and M. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, 2001.

[69] S. Carlile. *Virtual auditory space: Generation and applications*. Springer Science & Business Media, 2013.

[70] R. Häusler, S. Colburn, and E. Marr. Sound localization in subjects with impaired hearing: spatial-discrimination and interaural-discrimination tests. *Acta Oto-Laryngologica*, 96(sup400):1–62, 1983.

[71] F. Wightman and D. Kistler. Headphone simulation of free-field listening. i: stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867, 1989.

[72] Eric Lindemann and John L Melanson. Binaural hearing aid, December 26 1995. US Patent 5,479,522.

[73] G. Bienvenue and B. Siegenthaler. A clinical procedure for evaluating auditory localization. *Journal of Speech and Hearing Disorders*, 39(4):469–477, 1974.

[74] W. Noble, D. Byrne, and B. Lepage. Effects on sound localization of configuration and type of hearing impairment. *The Journal of the Acoustical Society of America*, 95(2):992–1005, 1994.

[75] V. Algazi and R. Duda. Headphone-based spatial sound. *IEEE Signal Processing Magazine*, 28(1):33–42, 2011.

[76] U. Grenander and M. Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.

[77] M. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1):61–84, 2001.

[78] M. Vaillant and J. Glaunès. Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 381–392. Springer, 2005.

[79] V. Camion and L. Younes. Geodesic interpolating splines. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 513–527. Springer, 2001.

[80] K. Atkinson, W. Han, and D. Stewart. *Numerical solution of ordinary differential equations*, volume 108. John Wiley & Sons, 2011.

[81] L. Younes. *Shapes and diffeomorphisms*, volume 171. Springer Science & Business Media, 2010.

[82] M. Vaillant, M. Miller, L. Younes, A. Trouvé., et al. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23(1):161, 2004.

[83] M. Miller., A. Trouva, and L. Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, 2006.

[84] J. Glaunés, A. Trouvé, and L. Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. Citeseer, 2004.

[85] G. Kuhn. Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1):157–167, 1977.

[86] C. Müller. *Spherical harmonics*, volume 17. Springer, 1996.

[87] D. Saupe and D. Vranić. 3d model retrieval with spherical harmonics and moments. In *Joint Pattern Recognition Symposium*, pages 392–397. Springer, 2001.

[88] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003.

[89] M. Blanco, M. Flórezl, and M. Bermejo. Evaluation of the rotation matrices in the basis of real spherical harmonics. *Journal of Molecular Structure: THEOCHEM*, 419(1-3):19–27, 1997.

[90] K. Park and N. Lee. A three-dimensional fourier descriptor for human body representation/reconstruction from serial cross sections. *Computers and biomedical research*, 20(2):125–140, 1987.

[91] C. Hetherington, A. Tew, and Y. Tao. Three-dimensional elliptic fourier methods for the parameterization of human pinna shape. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–612. IEEE, 2003.

[92] J. Glaunés and S. Joshi. Template estimation form unlabeled point set data and surfaces for computational anatomy. In *1st MICCAI workshop on mathematical foundations of computational anatomy: geometrical, statistical and registration methods for modeling biological shape variability*, 2006.

[93] C. Cury, J. Glaunés, and O. Colliot. Template estimation for large database: a diffeomorphic iterative centroid method using currents. In *Geometric Science of Information*, pages 103–111. Springer, 2013.

[94] C. Cury, J. Glaunès, and O. Colliot. Diffeomorphic iterative centroid methods for template estimation on large datasets. In *Geometric Theory of Information*, pages 273–299. Springer, 2014.

[95] Y. Tao, A. Tew, and S. Porter. The differential pressure synthesis method for efficient acoustic pressure estimation. *Journal of the Audio Engineering Society*, 51(7/8):647–656, 2003.

[96] M. Burkhard and R. Sachs. Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1):214–222, 1975.

[97] C. Guezenoc and R. Seguier. HRTF individualization: A survey. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[98] J. Middlebrooks. Viritual localisation improved by scaling nonindividualized external-ear transfer functions in frequency. *J. Acoust. Soc. of Am.*, 106(3):1493–1510, 1999.

[99] K. Maki and S. Furukawa. Reducing individual differences in the external-ear transfer functions of the mongolian gerbil. *The Journal of the Acoustical Society of America*, 118(4):2392–2404, 2005.

[100] P. Guillon, R. Nicol, and L. Simon. Head-related transfer functions reconstruction from sparse measurements considering a priori knowledge from database analysis: A pattern recognition approach. In *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[101] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis. Hrtf personalization using anthropometric measurements. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pages 157–160. Ieee, 2003.

[102] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4463–4467. IEEE, 2014.

[103] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile. Enabling individualized virtual auditory space using morphological measurements. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing)*, pages 235–238. Citeseer, 2000.

[104] Q. Huang and Q. Zhuang. HRIR personalisation using support vector regression in independent feature space. *Electronics letters*, 45(19):1002–1003, 2009.

[105] H. Hu, L. Zhou, H. Ma, and Z. Wu. Hrtf personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics*, 69(2):163–172, 2008.

[106] B. Seeber and H. Fastl. Subjective selection of non-individual head-related transfer functions. Georgia Institute of Technology, 2003.

[107] B. Katz and G. Parseihian. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, 131(2):EL99–EL105, 2012.

[108] B. Xie, X. Zhong, and N. He. Typical data and cluster analysis on head-related transfer functions from chinese subjects. *Applied Acoustics*, 94:1–13, 2015.

[109] C. Tan and W. Gan. User-defined spectral manipulation of HRTF for improved localisation in 3d sound systems. *Electronics letters*, 34(25):2387–2389, 1998.

[110] P. Runkle, A. Yendiki, and G. Wakefield. Active sensory tuning for immersive spatialized audio. Georgia Institute of Technology, 2000.

[111] S. Hwang, Y. Park, and Y. Park. Customization of spatially continuous head-related impulse responses in the median plane. *Acta Acustica united with Acustica*, 96:351–363, 03 2010.

[112] K. Fink and L. Ray. Individualization of head related transfer functions using principal component analysis. *Applied Acoustics*, 87:162–173, 2015.

[113] J. Holzl. *A Global Model for HRTF Individualization by Adjustment of Principal Component Weights*. PhD thesis, Diploma Thesis, 2014.

[114] P. Yan and K. Bowyer. Ear biometrics using 2d and 3d images. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 121–121. IEEE, 2005.

[115] W. Hugeng and D. Gunawan. Effective preprocessing in modeling head-related impulse responses based on principal components analysis. *Signal Processing: An International Journal (SPIJ)*, 4(4):201–212, 2010.

[116] ITU-R BS.1534-1:2003. *Method for the subjective assessment of intermediate quality level of coding systems*. ITU-R, 2003.

[117] P. Majdak, T. Walder, and B. Laback. Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *The Journal of the Acoustical Society of America*, 134(3):2148–2159, 2013.

[118] P. Majdak, R. Baumgartner, and B. Laback. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Frontiers in psychology*, 5:319, 2014.

[119] R. Baumgartner, P. Majdak, and B. Laback. Assessment of sagittal-plane sound localization performance in spatial-audio applications. In *The technology of binaural listening*, pages 93–119. Springer, 2013.

[120] D. Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 168(1011):158–180, 1967.

[121] J. Hartigan and M. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[122] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100, 1990.

[123] J. He, W. Gan, and E. Tan. On the preprocessing and postprocessing of hrtf individualization based on sparse representation of anthropometric features. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 639–643. IEEE, 2015.

[124] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[125] X. Zeng, S. Wang, and L. Gao. A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures. *Journal of Sound and Vibration*, 329(19):4093–4106, 2010.

[126] V. Algazi, C. Avendano, and R. Duda. Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society*, 49(6):472–479, 2001.

[127] E. Rasumow, M. Blau, M. Hansen, P. V. Steven, S. Doclo, V. Mellert, and D. Püschel. Smoothing individual head-related transfer functions in the frequency and spatial domains. *The Journal of the Acoustical Society of America*, 135(4):2012–2025, 2014.

[128] G. Grindlay and M. Vasilescu. A multilinear approach to hrtf personalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[129] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang. Statistical method to identify key anthropometric parameters in HRTF individualization. In *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pages 213–218, May 2011.

[130] D. Kistler and F. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91:1637–1647, 1992.

[131] J. Chen, B. Veen, and K. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *J. Acoust. Soc. of Am.*, 97(1):1493–1510, 1995.

[132] O. Ramos and F. Tommasini. Magnitude modelling of hrtf using principal component analysis applied to complex values. *Archives of Acoustics*, 39(4):477–482, 2014.

ANALYZING NOTCH PATTERNS OF HEAD RELATED TRANSFER FUNCTIONS IN CIPIC AND SYMARE DATABASES

M. Shahnawaz, L. Bianchi, A. Sarti, S. Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

ABSTRACT

The sensation of elevation in binaural audio is known to be strongly correlated to spectral peaks and notches in HRTFs, introduced by pinna reflections. In this work we provide an analysis methodology that helps us to explore the relationship between notch frequencies and elevation angles in the median plane. In particular, we extract the portion of the HRTF due to the presence of the pinna and we use it to extract the notch frequencies for all the subjects and for all the considered directions. The extracted notch frequencies are then clustered using the K-means algorithm to reveal the relationship between notch frequencies and elevation angles. We present the results of the proposed analysis methodology for all the subjects in the CIPIC and SYMARE HRTFs databases.

Index Terms— Binaural audio, Elevation perception, Head Related Transfer Function (HRTF), k-means.

1. INTRODUCTION

Sound perception is the result of the interaction between the acoustic wavefield and the listener's body, which causes wave scattering, reflection and diffraction. These phenomena alter the spectral content of the sound signal in a direction-dependent fashion, and introduce a wide variety of cues that enable sound localization. The interaction between soundfield and listener's body is encoded by a complex-valued transfer function, usually known as *Head Related Transfer Function (HRTF)*, which describes the spectral modifications that are characteristics of a source in a given location with respect to the listener [1]. The time-domain equivalent of this transfer function is known as *Head Related Impulse Response (HRIR)*.

Knowing the HRTF of a person is what enables spatial sound reproduction using headphones. However, as confirmed by many studies, HRTFs are strongly dependent on the listener's anatomy. This means that, in order to guarantee the best performance in terms of sound localization, individualized HRTFs need to be adopted [2, 3]. Unfortunately, the measurement of HRTFs is so expensive and time-consuming to prevent its use in consumer applications.

We would like to thank Prof. Craig Jin and Dr. Nicolas Epain and CAR-Lab research team for all the help and support, as well as for the permission to use the SYMARE database for our work.

A great deal of effort has been put into the personalization of HRTFs. In [4, 5], for example, suggest to estimate individualized HRTFs from 3D models of the user's pinnas. Some techniques based on low-cost capturing devices [6] have been proposed for this purpose, though the acquisition of a sufficiently accurate 3D model is still not an easy task for the average user. An alternate solution consists of synthesizing individualized HRTFs from a structural model of the listener's body [7–9]. Using parametric filters that rely on a given mapping between parameters and anthropometric data, the authors obtain computationally efficient and customizable solutions that can be used to approximate individualized HRTFs.

Notches in the HRTF caused by the pinna, are known to have significant perceptual relevance for sound localization, particularly in the frontal region [10–13]. Some studies, e.g. [14, 15], reported that the frequencies of the notches greatly depend on the elevation angle of the sound source, and they are almost independent of azimuth and distance. Recently, an important observation has been made in [4], where the authors related the notches in the HRTF with the three main pinna contours.

In this manuscript we study the relation between the notch frequencies and the elevation angles for a large number of subjects, whose HRTFs have been acoustically measured and stored into two databases: the CIPIC database [16] and the SYMARE database [5]. Notch frequencies are extracted from the collected HRTFs after removing all contributions of head, torso, and shoulders, while retaining only the contribution of the pinna, as described in [17]. We group the notch frequencies for all the subjects under consideration into three clusters, each corresponding to one of the three main pinna contours identified in [4]. In this setting, we analyze the evolution of the notch frequencies as a function of sound source elevation in the median plane. Moreover, we analyze the correlation between notch frequencies in the left and right ears.

2. ANALYSIS METHODOLOGY

This section introduces the methodology used in this work. The overall methodology can be divided into two conceptual steps: notch frequency extraction and clustering. Figure 1 shows the block diagram of the overall analysis methodology while Fig. 2 explains the steps involved in notch extraction.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

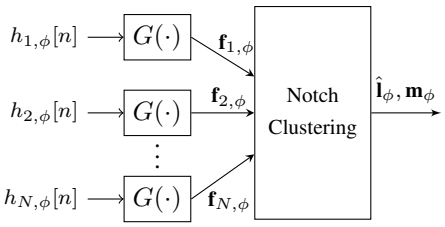


Fig. 1 Block diagram of the proposed analysis methodology.

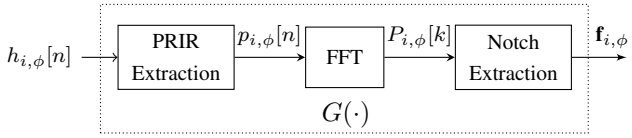


Fig. 2 Detail of the notch extraction procedure.

The detailed description of each step is given below.

2.1. PRTF Extraction

The deep spectral notches are produced in HRTF due to reflections caused by different body parts including pinna cavities, head, torso and knees. In this study we aim to analyze the spectral notches caused by pinna, so the first step is removing all unnecessary components of HRIR namely the contributions of head, shoulders and knees preserving the contributions of pinna. In [17] it was reported that the delays of pinna, torso and knee reflections are typically around 0.1 to 0.3, 1.6 and $3.2\,\mathrm{ms}$ [10, 17] respectively.

To get rid of shoulders, torso and knees reflection components we shorten our HRIR by applying a half Hanning window [17] of length $1\,\mathrm{ms}$, starting from onset of HRIR. This removes the reflective components due to shoulders, torso and knees, while preserving the reflection caused by pinna.

Given the HRIR $h_{i,\phi}[n]$ for the user i, and elevation angle ϕ, the PRIR $p_{i,\phi}[n]$ can be extracted by applying a half Hanning window $w[n]$ starting from onset of HRIR n_o, i.e. $p_{i,\phi}[n] = h_{i,\phi}[n]w[n - n_0]$. Figure 3 illustrates the windowing operation. The value of n_0 can be found by taking the slope of unwrapped phase function of HRTF [17].

Once the PRIR $p_{i,\phi}[n]$ is obtained, the PRTF (Pinna related transfer function) $P_{i,\phi}[f]$ can be obtained by evaluating

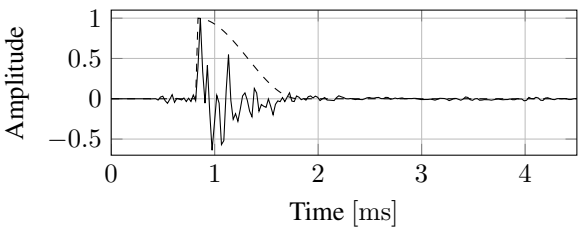


Fig. 3 HRIR windowing for PRIR extraction.

its Fourier transform, where f denotes the frequency of the signal. Next we describe the notch frequency extraction procedure from the PRTFs $P_{i,\phi}[f]$, $i = 1, \ldots, N$, relative to all N users.

2.2. Notch Extraction

As reported in [11], the frequency content in the range $4\,\mathrm{kHz}$ to $16\,\mathrm{kHz}$ is the main cause of median plane localization. For this reason, we restrict the frequency bandwidth of our analysis to this range.

To extract the notches we use the negative of log-scale magnitude function of the PRTFs, i.e.

$$\mathring{P}_{i,\phi}[f] = -20\log_{10}(|P_{i,\phi}[f]|). \tag{1}$$

The purpose of this step is to turn the notches into peaks, so that they can be effectively extracted by finding the local maxima in $\mathring{P}_{i,\phi}[f]$. In order to get meaningful results, we also have to make sure that we are considering just the significant and prominent notches, while discarding all those which are not relevant. For this purpose, we consider the prominence of the local maxima. The prominence describes how much the peak stands out from the neighboring peaks. For instance, a low isolated peak can be more prominent than one that is higher but is next to an other higher peak and vice-versa.

In the following, we considered those peaks in $\mathring{P}_{i,\phi}[f]$ that have a prominence greater than $3\,\mathrm{dB}$. These values are stored in vectors $\mathbf{f}_{i,\phi}$ for each subject i and elevation ϕ as

$$\mathbf{f}_{i,\phi} = [f_{i,\phi,1}, \ldots, f_{i,\phi,M_{i,\phi}}], \tag{2}$$

being $M_{i,\phi}$ the number of relevant peaks in PRTF of i^{th} user for elevation angle ϕ.

Once we have notch frequency vectors, $\mathbf{f}_{i,\phi} \in \mathbb{R}^{1\times M_{i,\phi}}$ for all the users and elevations, we arrange them into the vector $\mathbf{f}_\phi$, which contains the notch frequencies for all the users for elevation ϕ, i.e.

$$\mathbf{f}_\phi = \left[\mathbf{f}_{1,\phi}\mathbf{f}_{2,\phi}\ldots\mathbf{f}_{N,\phi}\right] \in \mathbb{R}^{1\times M_\phi}, \text{ with } M_\phi = \sum_{i=1}^{N} M_{i,\phi}. \tag{3}$$

2.3. Clustering of Notches

The next step of the analysis is to find the meaningful information from the frequency vectors $\mathbf{f}_\phi$. In a recent study [4], the authors reported that in each PRTF in CIPIC database up to three main spectral notches can be extracted, and mapped to three distinctive and prominent pinna contours: the helix, anti helix and outer wall of the concha.

Based on these findings, in this study we clustered the notch frequency vector $\mathbf{f}_\phi$ consisting of M_ϕ elements into $K = 3$ groups, using a well known clustering algorithm K-means [18]. At the end of the process, each element in $\mathbf{f}_\phi$ will

be assigned to a single cluster, whose centroid is the closest to the actual value of the element.

We evaluate the distance between each element $f_{i,\phi,j} \in \mathbf{f}_\phi$ and the corresponding centroid $m_{k,\phi}$ as the euclidean distance $D(f_{1,\phi,j}, m_{k,\phi}) = |f_{1,\phi,j} - m_{k,\phi}|$.

The K-means algorithm is initialized by assigning random values to the centroids $m_{k,\phi}$, $k = 1, 2, 3$. The algorithm is defined as an iterative two-step process. The first step is the assignment of each notch frequency to a cluster having closest centroid and label it with the number of that cluster e.g. $1, 2$ or 3 according to

$$\hat{l}_j = \arg\min_k \{D(f_{i,\phi,j}, m_{k,\phi})\} \tag{4}$$

where $j = 1, \ldots, M_\phi$ and $k = 1, 2, 3$. Moreover, a responsibility vector is defined for each cluster as

$$r_{k,j} = \begin{cases} 1, & \text{if } \hat{l}_j = k, \\ 0, & otherwise. \end{cases} \tag{5}$$

The second step is to update the centroid for all the clusters. The updated value for the kth centroid is

$$m_{k,\phi} = \frac{\sum_{j=1}^{M_\phi} r_{k,j} f_{i,\phi,j}}{R_k}, \quad R_k = \sum_{j=1}^{M_\phi} r_{k,j} \tag{6}$$

where R_k is the total responsibility of cluster k, defined as the number of points belonging to cluster k.

The process continues until no further changes occur in the cluster centroids.

After applying the K-means algorithm, we obtain the centroids $\mathbf{m}_\phi = [m_{1,\phi}, m_{2,\phi}, m_{3,\phi}]$, corresponding to helix, anti-helix and outer wall of concha respectively. Moreover, in order to associate a relevance descriptor to the clustered data, we introduce the cluster spread as the standard deviation of their elements, i.e.

$$\sigma_{k,\phi} = \sqrt{\frac{\sum_{j=1}^{M_\phi} \left(f_{i,\phi,j} - m_{k,\phi}\right)^2 r_{k,j}}{R_k}} \tag{7}$$

The results are further analyzed in the next section.

3. RESULTS

In this section we describe the application of the analysis methodology described in Sec. 2 to the CIPIC and SYMARE databases.

3.1. Description of the databases

For this study we used acoustically measured HRTFs from two well known databases of fairly large population set.

3.1.1. CIPIC

CIPIC [16] is a public-domain database of acoustically measured HRIRs with a high spatial resolution. It contains HRIRs for 45 subjects (27 male, 16 female and two KEMAR) measured at 1250 different directions around the head of the subjects. The measurements are done using Golay code as analysis signals, with a sampling frequency of $44.1\,\mathrm{kHz}$. Measurements loudspeakers are mounted on a circular arc of radius $1\,\mathrm{m}$, which is rotated around a fixed listener. The length of each HRIR stored in the database is 200 samples. For the purpose of this work, we consider all the HRIRs at azimuth $0°$ and elevations ϕ between $-45°$ and $45°$, with a uniform spacing equal to $5.625°$.

3.1.2. SYMARE

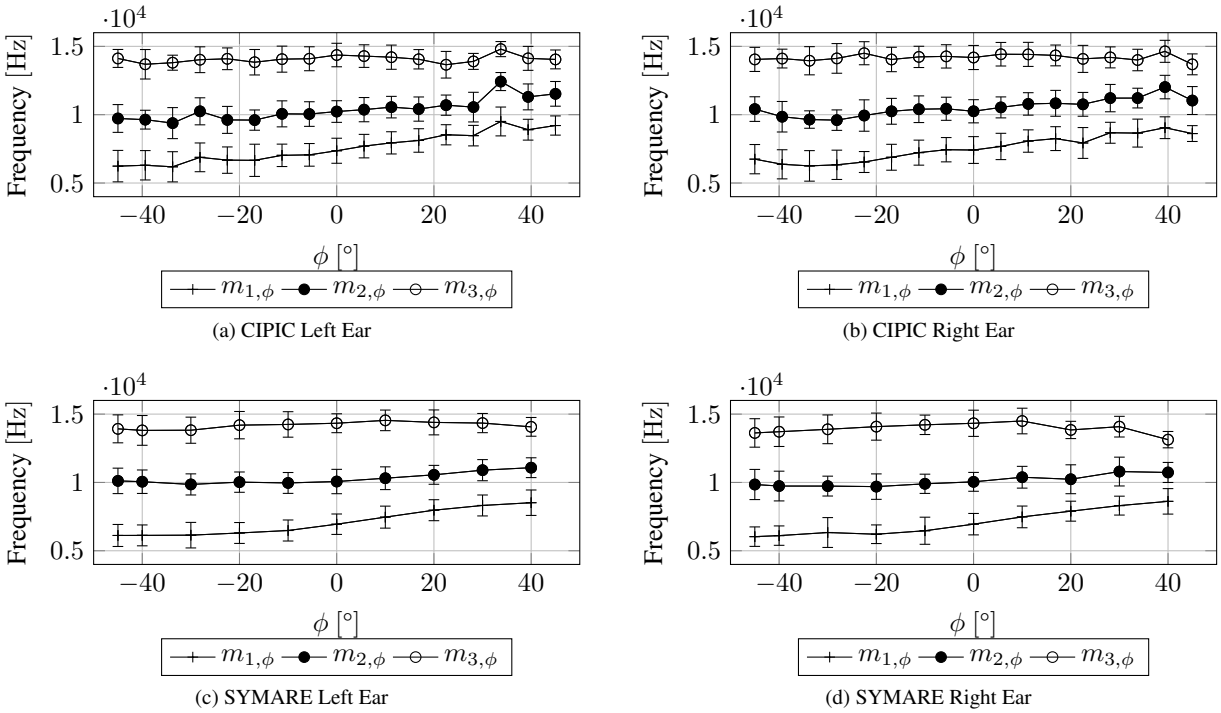
SYMARE [5] database was created by a collaborative team of Sydney University Australia and University of York England. This database contains acoustically measured HRTFs for 61 users (45 males and 16 females) measured in 393 directions around the head at a distance of $1\,\mathrm{m}$, with a non-uniform angular spacing in elevation for different azimuth angles. Impulse responses are recorded using Golay codes with a sampling frequency equal to $48\,\mathrm{kHz}$. The length of each HRIR is 256 samples. For the purpose of this work, we consider all the HRIRs at azimuth $0°$ and elevations ϕ between $-45°$ and $40°$.

3.2. Analysis 1

The steps defined in section 2 were applied to all the HRIR sets in both databases. HRIRs for the mentioned elevations were retrieved from the databases and PRIRs were extracted from each HRIR. The PRIRs were then transformed in the frequency domain by a zero-padded 512-point FFT.

Notches vectors $\mathbf{f}_\phi$ are estimated for each direction ϕ according to the angular grid adopted by the database, and notch frequencies are grouped into 3 clusters $m_{k,\phi}$, $k = 1, 2, 3$, along with their corresponding spread $\sigma_{k,\phi}$. Figure 4 shows the cluster centroids and spreads as a function of the elevation angle ϕ for the left and right ears of all the subjects in CIPIC and SYMARE databases.

We notice that for $\phi = -45°$ the cluster mean for all four cases (CIPIC and SYMARE databases, left and right ears) has almost the same value. Another observation that we want to point out is that all the cluster means $m_{k,\phi}$, $k = 1, 2$, exhibit a monotonically increasing behavior as a function of ϕ, despite of some slight irregularities. These irregularities are more prominent in the CIPIC database. On the other hand, $m_{3,\phi}$ results to be almost constant in all the four considered cases. In a more general way, we observe that the slope of the clusters $m_{k,\phi}$, $k = 1, 2, 3$, is the highest for $m_{1,\phi}$ and almost null for $m_{3,\phi}$. This behavior suggests that the pinna reflection causing a notch in the range of $m_{1,\phi}$ might be the most



(a) CIPIC Left Ear

(b) CIPIC Right Ear

(c) SYMARE Left Ear

(d) SYMARE Right Ear

Fig. 4 Cluster centroids and spreads as a function of elevation angle ϕ.

informative one for elevation perception.

In the case of data extracted from the CIPIC database, we observe a peak around $\phi = 30°$ for the left ear, while the right ear exhibit a peak around $\phi = 40°$. In the SYMARE database these irregularities are very mild and are present in just right ear, while the tracks for left ear are very smooth.

3.3. Analysis 2

Further, we compare the results obtained for left and right ears in both databases. First, we convert the frequency centroids $m_{k,\phi}$ to the Bark scale [19] and then we compute their Euclidean distance. In the following we denote by $d_{k,\phi}$ the distance between the centroid of left and right ears for the kth cluster and elevation ϕ. Results are reported in Fig. 5.

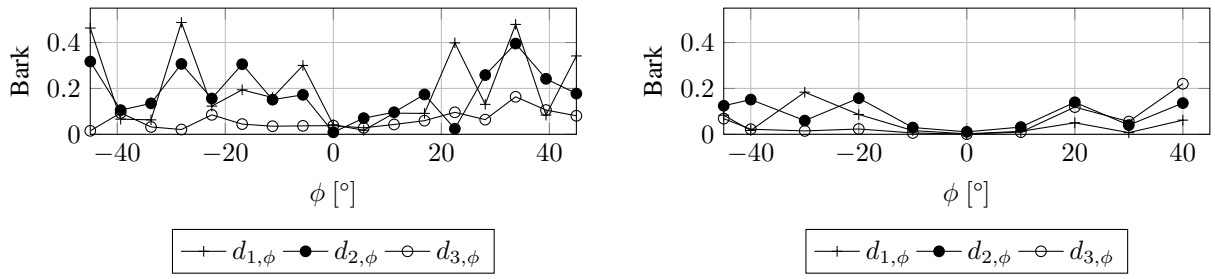
We observe that, in both CIPIC and SYMARE, the maximum value for the distance between clusters is less than 0.5 Bark for all the considered cases and for all the elevations. In case of SYMARE database distances have smaller values and a smoother distribution, while in the CIPIC database distances are, in general, greater and less regular with respect to ϕ. We would like to point out that the differences exhibit minima in the horizontal plane ($\phi = 0°$) in all the considered cases and for all the clusters, suggesting that binaural cues are not relevant in the frontal direction. On the other hand, it can be observed that the distances are greater moving away from the horizontal plane; this behavior suggests that both monaural and binaural cues are relevant for elevation perception in the median plane.

4. CONCLUSIONS

In this manuscript we provide a methodology to analyze HRTFs in publicly available databases. In particular, we describe a technique to extract notch frequencies from HRTF data and to classify them into three clusters, each corresponding to a specific contour in the pinna namely the helix, anti helix and outer wall of the concha. We validated the proposed methodology with acoustically measured HRTFs from the CIPIC and SYMARE databases. We performed a comparative study on the evolution of notch frequencies in median plane in CIPIC and SYMARE databases. Results show the strong dependency between notches in the HRTFs and elevation angles in the median plane. Moreover, we also studied the binaural differences between noth frequencies which revealed that not only monaural but also binaural cues are important for elevation perception. We envision our approach to be applied in combination with the techniques mentioned in [20–22] for the auralization of virtual and real sound environments.

REFERENCES

[1] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs



(a) CIPIC database

(b) SYMARE database

Fig. 5 Distance between the centroids for the left and right ears as a function of elevation ϕ.

in time, frequency, and space," in *Proc. AES 107th Conv.* AES, 1999.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[3] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.

[4] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and Head-Related Transfer Function features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 508–519, 2013.

[5] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the Sydney York morphological and acoustic recordings of ears database," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 37–46, 2014.

[6] L. Bonacina, A. Canclini, F. Antonacci, M. Marcon, A. Sarti, and S. Tubaro, "A low-cost solution to 3D pinna modeling for HRTF prediction," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, 2016.

[7] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, 1998.

[8] V. R. Algazi, R. O. Duda, and P. Satarzadeh, "Physical and filter pinna models based on anthropometry," in *Proc. AES 122nd Conv.* AES, 2007.

[9] I. Faller, K. John, A. Barreto, and M. Adjouadi, "Augmented Hankel total least-squares decomposition of head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 58, no. 1/2, pp. 3–21, 2010.

[10] D. W. Batteau, "The role of the pinna in human localization," *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 168, no. 1011, pp. 158–180, 1967.

[11] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, 1974.

[12] D. Wright, J. H. Hebrank, and B. Wilson, "Pinna reflections as cues for localization," *Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 957–962, 1974.

[13] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, 2007.

[14] E. A. G. Shaw, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter Acoustical features of the human external ear, Lawrence Erlbaum, Mahwah, NJ, US, 1997.

[15] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA)*, 2001.

[17] V. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, 2005.

[18] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[19] H. Traunmüller, "Analytical expression for the tonotopic sensory scale," *Journal of the Acoustical Society of America*, vol. 88, pp. 97–100, 1990.

[20] M. Foco, P. Polotti, A. Sarti, and S. Tubaro, "Sound spatialization based on fast beam tracing in dual space," in *Proc. of the 6^{th} Int. Conference on Digital Audio Effects, (DAFx-03)*, 2003.

[21] F. Antonacci, M. Foco, A. Sarti, and S. Tubaro, "Real time modeling of acoustic propagation in complex environments," in *Proc. of the 7^{th} Int. Conference on Digital Audio Effects DAFx-04*, 2004.

[22] M. Vorländer, *Auralization, Fundamentals of acoustics, modemodel, simulations, algorithms and acoustic virtual reality*, Springer-Verlag Berlin Heidelberg, 2008.

HRTF PERSONALIZATION BASED ON WEIGHTED SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES

Mo Zhu, Muhammad Shahnawaz, Stefano Tubaro, Augusto Sarti

Sound and Music Computing Lab,
Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy

ABSTRACT

Personalized head-related transfer functions (HRTFs) are essential for presenting authentic spatial audio through binaural rendering. However, measuring personalized HRTFs for every user is a tedious task and requires specialized equipment. This paper presents an easy and efficient method for obtaining personalized magnitude response of HRTFs. It treats the problem of HRTF synthesis as finding the sparse representation of the anthropometric features of the new listener with respect to anthropometric features of the user set in the CIPIC database. Unlike the previous sparse representation methods, our method assigns different weights to different anthropometric features depending on their relevance. We compared our approach with state of the art sparse representation and closest-match based approaches. The results show that our approach outperforms the previous approaches resulting an average spectral distortion value of 5.53 dBs between synthesized and actual HRTFs for all users present in the CIPIC database.

Index Terms— Head-related transfer functions, HRTF personalization, anthropometric features, weighted sparse representation

1. INTRODUCTION

Spatial hearing is the result of the interaction between the acoustic wavefield and the listener's anatomy, which causes wave scattering, reflection and diffraction. These phenomena modify the spectral content of the sound signal in a direction dependent fashion and introduce a wide variety of cues which enable the listener to localize the sound sources in 3D space. Interactions between the soundfield and listener's body can be encoded by a direction-dependent, complex-valued transfer function, known as head-related transfer function (HRTF). An HRTF describes the spectral colorations that are imposed on a source in a given location with respect to the listener [1]. The time-domain equivalent of this transfer function is known as the head-related impulse response (HRIR).

Availability of the HRTFs, enables spatial audio reproduction over headphones. However, as confirmed by many studies, HRTFs are highly idiosyncratic due to their strong dependence on the listener's anatomy. This means the best performance can only be guaranteed by using individualized HRTFs [2, 3]. However, the measurement of HRTFs is expensive and time-consuming and requires specialist equipment and trained operators. Thus it is limited to few laboratories in the world. This prevents its use in consumer applications [4, 5, 6]. As a result, a simple and efficient solution is needed that can provide the personalized HRTFs for the listener without going through the currently used lengthy and inconvenient process of HRTF measurement.

Considering the dependence of HRTFs on anatomy, much efforts has been concentrated on personalizing HRTFs based on a selected set of anthropometric features, such as head width, height and depth, height and width of the pinna (outer ears), etc. The simplest possible approach, as proposed in [7, 8], is to use these anthropometric features to select the closest match from a database of non-individualized HRTFs. However, the closest match does not guarantee good performance in all cases because it simply returns the closest-match non individualized set of HRTFs in the database and does not let the user adjust the HRTF magnitudes.

Moreover, studies in [9, 10, 11, 12], attempt to identify the relationship between the anthropometric and the HRTF features by directly relating them. Other approaches [13, 14], investigate the complex relationship using PCA and neural networks. However, the performance of all these approaches strongly depends on the choice of the selected features.

Recently, authors in [15] proposed a new HRTF personalization method based on sparse representation. They assumed that the magnitude of an HRTF can be described by the same sparse representation as the anthropometric features in the training data. Based on this assumption, HRTFs for a new subject, not present in the database, can be synthesized by sparse representation of its anthropometric features and HRTFs in a database of non-individualized HRTFs. The results show that this method can improve personalization. Refinements in [16] in which, post and preprocessing methods are incorporated improve the performance further.

After studying the work of [15, 16], we introduce an HRTF

personalization method based on weighted anthropometric sparse representation and combine it with the preprocessing and post-processing methods described in [16]. In our work, we used only 17 anthropometric parameters (10 for the head and torso and 7 for each ear), all of which can be measured from three scaled pictures of a subject, as in [17].

2. METHODOLOGY

Sparse representation based HRTF personalization schemes begin by finding a sparse representation of the subject's anthropometric features, i.e finding the linear combinations of the given anthropometric features which can generate the anthropometric features of the new subject. The approach is based on a strong assumption, that the HRTFs can be represented in the same sparse representation as the anthropometric features. The second assumption is that the given training set is sufficiently rich to encompasses the anthropometric features of any new subject.

In previous sparse representation based HRTF personalization techniques [15, 16], the anthropometric parameters are considered equally important. However, previous studies suggest that this is not the case and some features are more relevant than the others. For example, ear features are the more relevant than torso features [9].

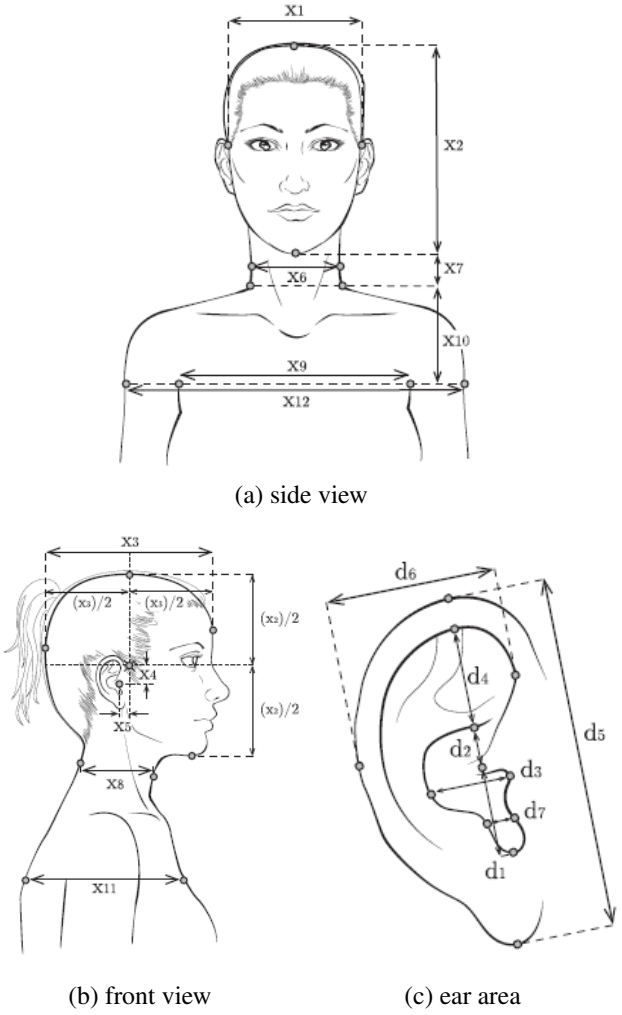
Moreover, previous approaches combine the anthropometric measurements of left and right ears into a single vector for determining a new subject's sparse representation. As a result, each subject only has one sparse representation of anthropometric features and this is used for both left and right HRTFs. This may be satisfactory in the cases, where a subject has perfectly symmetrical ears. However, many users have asymmetric ears and use of single vector may lead to a poor listening experience.

In this study, our contributions are twofold. First, we assign weights to the anthropometric features using a partially on-off strategy approach, described in [17]. Later, these weights are used to devise the relevance of the anthropometric feature while calculating a sparse representation. Furthermore, we compute separate sparse representations for personalization for both the left and the right ear.

Based on studies and experiments conducted in [16], we applied pre- and post-processing to our anthropometric features and HRTFs using the best combination of methods reported by authors. Figure 2 shows the block diagram of the methodology. The details of each block are provided in the following section.

2.1. Database

All the experiments in our study are conducted on the CIPIC (Center for Image Processing and Integrated Computing) database [4]. This is a publicly available database of HRIRs that contains measured HRIRs for 45 subjects in 1250 different directions. The database also includes the measurements of 27 anthropometric parameters as shown in figure 1. As only 35 subjects have all 27 anthropometric measurements, we confined our study to just this subset.



(a) side view

(b) front view

(c) ear area

Fig. 1. Anthropometric parameters that can be measured from (a) the side view, (b) front view, and (c) the pinna closeup view respectively (picture taken from [17]).

2.2. Anthropometric feature selection

The study in [17] reported that 19 anthropometric feature (12 for the head and torso and 7 for the pinna) can be directly measured using only the three scaled pictures, as illustrated in Figure 1. All 19 of these anthropometric features are listed in Table 1. Usually, x_5 (pinna offset back) is not easy to measure from pictures, because it highly depends on the flair angle of the pinna which can varies a lot across different users.

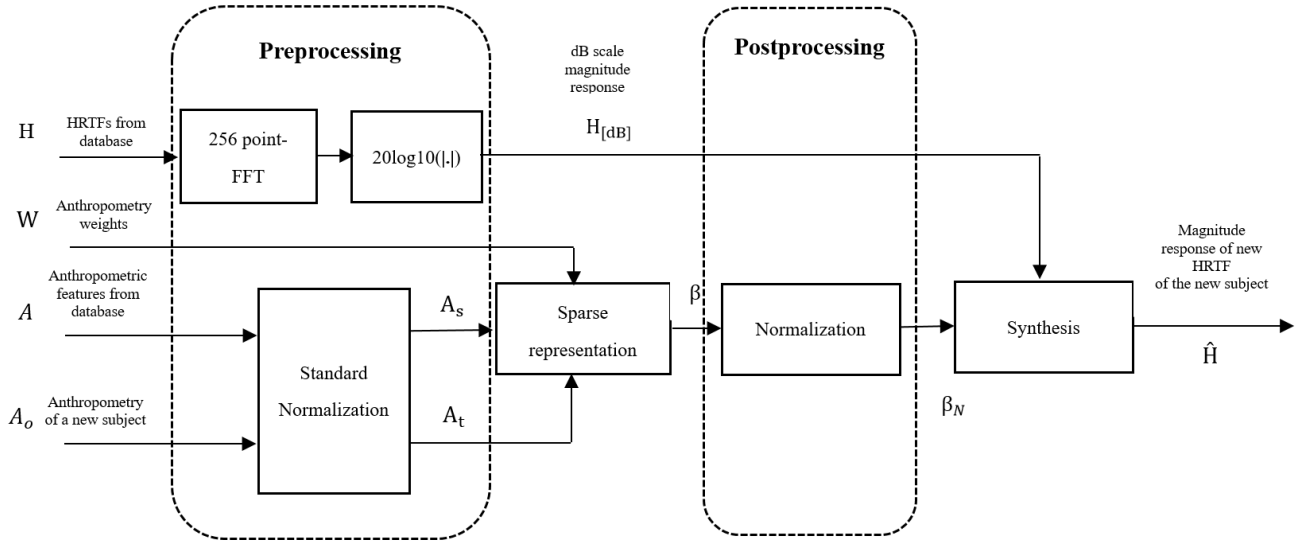


Fig. 2. Block diagram of HRTFs personalization using weighted sparse representation of anthropometric features

Var	Measurement	Var	Measurement
x_1	head width	d_1	cavum concha height
x_2	head height	d_2	cymba concha height
x_3	head depth	d_3	cavum concha width
x_4	pinna offset down	d_4	fossa height
x_5	pinna offset back	d_5	pinna height
x_6	neck width	d_6	pinna width
x_7	neck height	d_7	intertragal incisure width
x_8	neck depth		
x_9	torso top width		
x_{10}	torso top height		
x_{11}	torso top depth		
x_{12}	shoulder width		

Table 1. 19 Anthropometric parameters can be measured from scaled picture.

Similarly, x_7 (neck height) strongly depends on the posture of the subject while the photograph is captured so can not be measured. For this reason we suggest to use only the remaining 17 anthropometric parameters. Further details of obtaining the anthrpometric features from pictures are outside of the scope of the study and can be found in [17]. In this study, we directly used the anthropometric features provided in CIPIC database rather than measuring through described method.

2.3. Calculation of weights for anthropometric features

In previous sparse representation based approaches, such as [15, 16], all the anthropometric parameters are considered to be equally relevant. This has been shown incorrect in past studies and we use the approach described in [17] to calculate the weights of each anthropometric feature. Figure 3 shows the block diagram of the scheme. Only 17 features are ultimately required at the end, but in the CIPIC database 27 features are provided for every subject. For the relevance calculation, we used 25 of these 27 anthropometric features rather than 17 only omitting x_{14} (height) and x_{15} (seated height), as these features are not relevant when measuring the HRIRs. We used 25 features instead of 17 to determine the relevance metric for all 25 features which might be useful for future studies. As different anthropometric features lie in different ranges and scales, to bring them to a notionally common scale we normalized them using a min-max method. Given the anthropometric feature set $A = [a_1, a_2, \cdots, a_{25}]$, containing the 25 anthropometric features of a given subject in the CIPIC database the min-max normalization is expressed by:

$$A_N^{(i)} = \frac{A^{(i)} - min[A^{(i)}]}{max[A^{(i)}] - min[A^{(i)}]} \quad \forall i = 1, 2, \cdots, 25, \quad (1)$$

where $A^{(i)}$ and $A_N^{(i)}$ represent the actual and normalized i-th anthropometric features in the feature set, respectively.

To obtain all possible combinations of 25 anthropometric parameters, we used a partially on-off scheme. For every anthropometric feature, we can either include or exclude it in the calculation which results in: $2^{25} - 1 = 33,554,431$ different possible combinations, (-1 because we do not consider the case where all features are excluded). Next, we compared the subjects in pairs by calculating the distance between them as follows:

$$D^{(i,j,k)} = \|\sum_{a=1}^{25} A_N^{(i,k)} - \sum_{a=1}^{25} A_N^{(j,k)}\|, \quad \forall k = 1, 2, ..., 2^{25}-1, \tag{2}$$

where $D^{(i,j,k)}$ corresponds to the difference between the sum of the anthropometric parameters of i-th and of the j-th subject in the k-th combination.

Next, we calculated the average spectral distortions (SD) of HRTFs between all subject pairs according to:

$$SD(H^{(i)}, H^{(j)}) = \sqrt{\frac{1}{D}\frac{1}{F}\sum_{d=1}^{D}\sum_{f=1}^{F}(20\log_{10}\frac{\|H^{(i,d)}(f)\|}{\|H^{(j,d)}(f)\|})^2}, \tag{3}$$

where $H^{(i,d)}$ and $H^{(j,d)}$ correspond to the HRTF of i-th and the j-th subject in direction d. F is the number of frequency bins and is equal to 128 in our case as we took a 256 point FFT. D is the number of directions for which HRTFs are available and is equal to 1250 here. This process resulted in an $S \times S$ matrix of SD, where S is the number of users and is equal to 35 for our study.

We calculated the correlation between D and SD for all possible combinations of anthropometric parameters as follows:

$$\rho^{(i,k)} = corr(D^{(i\times 35,k)}, SD^{i\times 35})) \tag{4}$$

where $\rho^{(i,k)}$ corresponds to the Pearson's correlation coefficient of the i-th subject in the k-th combination. The combination which resulted in the maximum value of the correlation was selected as the best combination for the given subject. We obtained the best combinations for all subjects. The weights for different anthropometric features are obtained by dividing the number of occurrences of any given anthropometric feature with the total number of subjects:

$$W^{(i)} = \frac{t^{(i)}}{S}, \tag{5}$$

where $W^{(i)}$ corresponds to weight of i-th anthropometric feature. $t^{(i)}$ is number of times of i-th anthropometric parameter occurred in all best combinations. S is the number of best combinations and is equal to 35.

2.4. Preprocessing for anthropometric features and HRTFs for Sparse Representation

The authors in [16], suggest that using the standard normalized anthropometric feature vectors instead of the scalar magnitude vectors results in an improved sparse representation. For this purpose, we normalized the anthropometric feature sets calculating their standard scores as:

$$A_s = \frac{A - mean[A_d]}{std[A_d]}, \tag{6}$$

$$A_t = \frac{A_o - mean[A_d]}{std[A_d]}, \tag{7}$$

where A denotes the anthropometric features of all the subjects in the database, A_o denotes the anthropometric features of the new user, A_d is the superset of both and is given by $A_d = [A \quad A_o]$, A_s and A_t denotes the standard normalized vectors of anthropometric features of the users in the database and for a new subject, respectively.

The HRTFs are obtained from the HRIRs by computing a 256 point FFT. In[16], the authors state that using log-scale magnitude can result in an improved performance. For this reason in our work we used HRTFs on a dB scale instead of complex amplitudes.

$$H_{[dB]} = 20\log_{10}|H| \tag{8}$$

2.5. Sparse representation of anthropometric features

We used sparse representation to estimate the standard score of the new subject's anthropometric parameters A_t, as the linear superposition [15]:

$$A_t \approx \beta A_s \tag{9}$$

where A_s is the standard score of the anthropometric parameters A in the database.

In the sparse vector $\beta = [\beta^{(1)}, \beta^{(2)}, ..., \beta^{(S)}]^T$, each element corresponds to the weight of a subject in the linear superposition.

Thus, the problem of looking for an optimal sparse vector can be considered as a minimization problem:

$$\beta = \arg\min_{\beta}(\|W(A_t - \beta A_s)\|_2^2 + \lambda\|\beta\|), \quad \text{s.t.}\beta^{(i)} \geq 0, \tag{10}$$

where W represents the weights of different anthropometric parameters. In line with [16], we added a non-negative constraint on β, e.g. $\beta^{(i)} \geq 0$. Where the regularization parameter λ of this minimization problem is a non-negative parameter.

2.6. Postprocessing for sparse vectors

To ensure that the synthesizing process has consistent amplitudes at the output, as in the database, we normalized the values of β vector such that the sum of beta vector is equal to 1:

$$\beta_N = \frac{\beta}{\sum_{s=1}^{25}\beta(s)}. \tag{11}$$

2.7. HRTF synthesis

As in [15], we assume that the HRTFs can be represented using the same sparse representations as the anthropometric

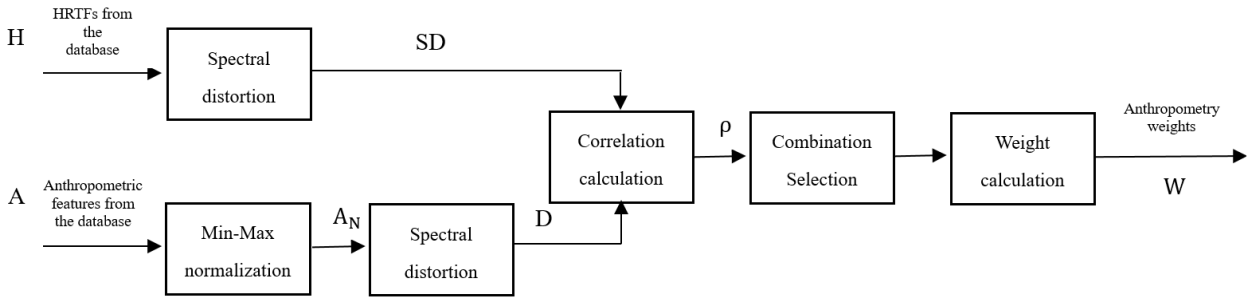


Fig. 3. Block diagram of weight calculation

features. Once we get the normalized sparse vector β_m, we can directly apply it to the log-scale HRTF data $H_{[dB]}$ in the database

$$\hat{H}_{[dB]} = \beta_N H_{[dB]}. \tag{12}$$

However, the new synthesized HRTF $\hat{H}_{[dB]}$ is expressed in dBs, so re-expressing the synthesized result as a scalar magnitude gives:

$$\hat{H} = 10^{\frac{\hat{H}_{[dB]}}{20}}. \tag{13}$$

2.8. Regularization parameter

The authors in [16] suggest that by adding only a single parameter λ into the minimization problem, one can prevent over-fitting. Several values for λ were tested using the anthropometric measurements and measured HRTFs in the database and one was selected as the optimal value.

To find the optimum value of λ, we used the "leave one person out" cross-validation approach [18] in the CIPIC database, and chose the value for λ that results in the smallest cross-validation error. We used the root-mean-square error as a cross validation measure as in eq 10.

To fit the scale of λ to the preprocessed anthropometric parameters and tune the value of λ easily, we normalized λ as suggested in [16]:

$$\lambda = \frac{\lambda_0}{1-\lambda_0}\|A_t\|_2^2, \tag{14}$$

where A_t corresponds to the preprocessed anthropometric parameters of the new subject. In this case, by tuning the value of λ_0 from 0 to 1, we can generate any nonnegative value for λ.

3. EXPERIMENTS

To evaluate the performance of our proposed approach, we applied the "leave one person out cross-validation" approach [18]. Each of the 35 subjects is taken out one-by-one as the test subject and the remaining 34 subjects are regarded as the training subjects.

Using the spectral distortion SD as our evaluation metric as described in Equation 15. We compared the results of our scheme with previously available sparse representation techniques and some closest-match based personalization schemes.

3.1. Evaluation Criteria

For evaluating the difference between synthesized HRTFs $\hat{H}$ and the original HRTFs H of the test subject, we employed a widely used error metric spectral distortion as our evaluation criteria [9, 15, 19].

$$SD^{(d)}(H,\hat{H}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(20\log_{10}\frac{\|H^{(d)}(n)\|}{\|\hat{H}^{(d)}(n)\|})^2} \quad [dB] \tag{15}$$

where $H^{(d)}$ is the original HRTF in the d-th direction, and $\hat{H}^{(d)}$ is the synthesized HRTF in same direction. N is the number of frequency bins ($N = 128$).

We used the root-mean-square error (RMSE) to compare the two sets of HRTFs for all 1250 directions:

$$SD(H,\hat{H}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(SD^{(d)}(H,\hat{H}))^2} \quad [dB] \tag{16}$$

where D is 1250.

3.2. Creating performance baselines

We also compared the performance of our proposed approach with three different closest-match methods introduced by [17, 20, 21]. For each method, we calculated average spectral distortion baselines for all 35 subjects in the CIPIC database.

The best-matched HRTF for a new subject was selected based on finding the minimum average spectral distortion between the actual and matched HRTF. The worst-matched HRTF, on the other hand, was selected from the result with the maximum average spectral distortion.

The spectral distortion values given in Table 3, for the best and the worst matching cases defines the boundaries for the spectral distortion values obtainable when using any of the possible closest match based scheme.

3.3. Results and discussion

The results of our experiments are presented in Table 2 and Table 3. The results presented in Table 2 show that the average spectral distortion using the weighted sparse representation, with 17 anthropometric parameters is 5.53dB. This is better than value obtained by using the unweighed sparse representation (5.57 dBs), even when 27 anthropometric parameters are used and also better than the SD score of (5.63dB), when unweighed sparse representation is used with only 17 anthropometric parameters. Even though we have used fewer anthropometric parameters, the weighted sparse representation still provides better results.

Results presented in Table 3 show that our proposed approach outperforms the three closest-match methods.

Considering "The Best" baseline, we find that the average spectral distortion of weighted sparse representation using 17 anthropometric parameters (5.53dB) is lower than "The Best" baseline (6.13dB), which means that our approach will perform better than any other closest-match methods under this evaluation criteria.

4. CONCLUSION AND FUTURE WORK

Building on the previous sparse representation techniques, we have introduced an easy and effective HRTF personalization method based on a weighted sparse representation with preprocessing and postprocessing. All anthropometric parameters used in our approach can be measured from scaled pictures of the subject. To reflect their relative influence in sparse representation, we have assigned weights to these anthropometric parameters, using spectral distortion as the experimental evaluation criteria. Our experiments show that the proposed approach has an average spectral distortion lower than that of previous sparse representation and other closest-match based personalization methods, indicating the effectiveness of our approach.

Future work includes validating our approach using perceptual localization tests and using non-linear higher order sparse representations to improve the performance further.

5. REFERENCES

[1] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," in *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F.L Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[3] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.

[4] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*. IEEE, 2001, pp. 99–102.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[6] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghariand A. Van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 37–46, 2014.

[7] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Ieee, 2003, pp. 157–160.

[8] A. Mohan, R. Duraiswami, D. Zotkin, D. DeMenthon, and L. S. Davis, "Using computer vision to generate customized spatial audio," in *Proceedings of International Conference on Multimedia and Expo*. IEEE, 2003, vol. 3, pp. III–57.

[9] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 3, pp. 508–519, 2013.

Sparse representation using	Left Ear	Right Ear	Average
17 parameters (weighted)	5.5235	5.5351	5.5293
17 parameters (unweighed)	5.6298	5.6359	5.6328
27 parameters (unweighed)	5.5770	5.5707	5.5738

Table 2. Average spectral distortion between the synthesized and the ground truth HRTF in [dBs].

Personalization method	Left Ear	Right Ear	Average
Weighted sparse representation with 17 anthropometric features	5.5235	5.5351	5.5293
Closest-match based on pinna contours *[20]	7.3403	7.3403	7.3403
Closest-match based on anthropometry and PCA[21]	7.6287	7.1844	7.4065
Closest-match based on weighted anthropometry[17]	7.5451	7.2239	7.3845
Closest-match "Best" baseline	6.2306	6.0317	6.1311
Closes-match"Worst" baseline	9.5628	9.0821	9.3324

Table 3. Average spectral distortion between the resulted and ground HRTFs in [dBs].

[10] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proceeding 18th International Conference Digital Audio Effects (DAFx-2015)*, 2015, pp. 231–236.

[11] M. Shahnawaz, L. Bianchi, A. Sarti, and S. Tubaro, "Analyzing notch patterns of head related transfer functions in CIPIC and SYMARE databases," in *24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 101–105.

[12] G. Grindlay and M. A. O. Vasilescu, "A multilinear approach to HRTF personalization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[13] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.

[14] L. Li and Q. Huang, "HRTF personalization modeling based on rbf neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3707–3710.

[15] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4468–4472.

[16] J. He, W. S. Gan, and E. L. Tan, "On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometric features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 639–643.

[17] E. A. T. Gallegos, F. O. Bustamante, and F. A. Cosío, "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Applied Acoustics*, vol. 97, pp. 84–95, 2015.

[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*. Stanford, CA, 1995, vol. 14, pp. 1137–1145.

[19] F. C. Tommasini, O. A. Ramos, M. X. Hüg, and F. Bermejo, "Usage of spectral distortion for objective evaluation of personalized HRTF in the median plane," *International Journal of Acoustics & Vibration*, vol. 20, no. 2, 2015.

[20] M. Michele, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 4463–4467.

[21] X. Y. Zeng, S. G. Wang, and L. P. Ga, "A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures," *Journal of Sound and Vibration*, vol. 329, no. 19, pp. 4093–4106, 2010.

CONSIDERATIONS REGARDING INDIVIDUALIZATION OF HEAD-RELATED TRANSFER FUNCTIONS

C. T. Jin, R. Zolfaghari, X. Long, A. Sebastian, S. Hossain, J. Glaun`es, A. Tew, M. Shahnawaz, A. Sarti

ABSTRACT

This paper provides some considerations regarding using individualized head-related transfer functions for rendering binaural spatial audio over headphones. It briefly considers the degree of benefit that individualization may provide. It then examines the degree of variation existing within the ear morphology across listeners within the Sydney-York Morphological and Recording of Ears (SYMARE) database using kernel principal component analysis and the large deformation diffeomorphic metric mapping framework. The degree of variation across listeners in the directivity patterns associated with head-related transfer functions is also analyzed as a function of frequency. The variation in ear morphology is related to the variation in the directivity patterns using simple linear regression.

***Index Terms*—** Morphoacoustics, LDDMM, Kernel principal Component Analysis, Head-related transfer functions, Binaural hearing, Hearables

1. INTRODUCTION

This paper focuses on individualization of head-related transfer functions for rendering binaural spatial audio using headphones - a research area with a long and varied history, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. It is well known that ear acoustics depends on the morphology of the periphery of the outer ear. Indeed, the study of the relationship between ear acoustics and the shape of the outer ear periphery has been termed morphoacoustics [13, 4, 14, 15]. Ear acoustics is often described in terms of 3D audio filter functions, referred to as head-related impulse reponses (HRIRs). HRIRs vary for each listener because each listener has different and uniquely shaped ears. There is an HRIR filter for each ear and each direction in space and these HRIR filters enable the rendering of binaural 3D audio for a listener.

The primary contribution of this work relates to a new study based on our recent work using the large deformation diffeomorphic metric mapping (LDDMM) approach to model ears and the fast-multipole boundary element method (FM-BEM) to numerically simulate ear acoustics. More specifically, we study the morphoacoustics of a simpler synthetic database of ear shapes which have been created from the SYMARE database by rotating, translating and scaling the ears to match a template ear shape. The synthetic database of ear shapes provides interesting viewpoints relating to the relationship between ear morphology and ear acoustics.

In addition to the primary morphoacoustic study which is the real focus of this paper, we also briefly consider a a psychoacoustic experiment contrasting individualized binaural spatial audio versus generic or non-individualized binaural spatial audio both with and without head-tracking enabled. These experiments highlight a few important considerations that are generally well-accepted within the community, but which would be useful to review given the recent, renewed interest in binaural spatial audio related to the rapid uptake of mixed reality and virtual reality technologies [16, 17] as well as hearable devices [18]. With regard to the psychoacoustics of binaural spatial hearing, there have been numerous psychophysical investigations relating to the influence of HRIRs on binaural hearing and localization, e.g., refer to the following books and references therein [19, 20, 21, 22].

2. BINAURAL SPATIAL RENDERING OF MUSIC

2.1. Methods

We recently conducted a binaural music listening test contrasting individualized HRIRs and generic HRIRs. More specifically, there were four listening conditions of relevance to this paper: (1) binaural rendering with individualized HRIR filters and head-tracking; (2) binaural rendering with generic HRIR filters and head-tracking; (3) binaural rendering with individualized HRIR filters and no head-tracking; and (4) normal headphone listening without binaural spatial rendering. We had twenty-three self-reporting normally-hearing listeners participate in the listening test. Listeners were asked to listen to six sound excerpts:

- Mono: drums, Radiohead - Weird Fishes/Arpeggi
- Mono: guitar, Tarrega - Capriccio Arabe
- Stereo: Pop, Radiohead - Jigsaw Falling Into Place
- Stereo: Bossa-Nova, Stan Getz, João Gilberto - Vivo Sonhando
- 5.1 Surround: Rock, Pink Floyd - Money
- 5.1 Surround: Pop Jazz, Norah Jones - Come Away With Me

Sounds were played to the listener using the AKG 1000 open headphones and also a loudspeaker array consisting of 12 loudspeakers: 5 Tannoy System 15 loudspeakers forming a 5.1 arrangement and 7 additional Tannoy V6 loudspeakers forming a circular array spaced every 45 degrees. The loudspeaker playback provided a reference for the headphone listening. Because the headphones are open, the loudspeakers could be heard without distortion. Every listener had HRIRs recorded using a blocked-ear recording method [23] in an anechoic chamber using a semi-circular robotic arm (methods were similar to those presented here [24]). A MUSHRA-like [25] test paradigm was used in which there was no hidden reference, but an anchor was included. The explicit reference was loudspeaker playback and the

C.T. Jin, R. Zolfaghari, X. Long, A. Sebastian, and S. Hossian are with CARLab, School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia. email: craig.jin@sydney.edu.au.

A. Tew is with the Department of Electronics, The University of York, Heslington, York, UK. email: tony.tew@york.ac.uk

J.A. Glaunès is with the MAP5, Université Paris Descartes, Sorbonne Paris Cité 75006 Paris, France. email: alexis.glaunes@mi.parisdescartes.fr

M. Shahnawaz and A. Sarti are with the Dipartimento di Electtronica, Informazione e Bioingegneria, Politecnico di Milano, Italy. email: augusto.sarti@polimi.it

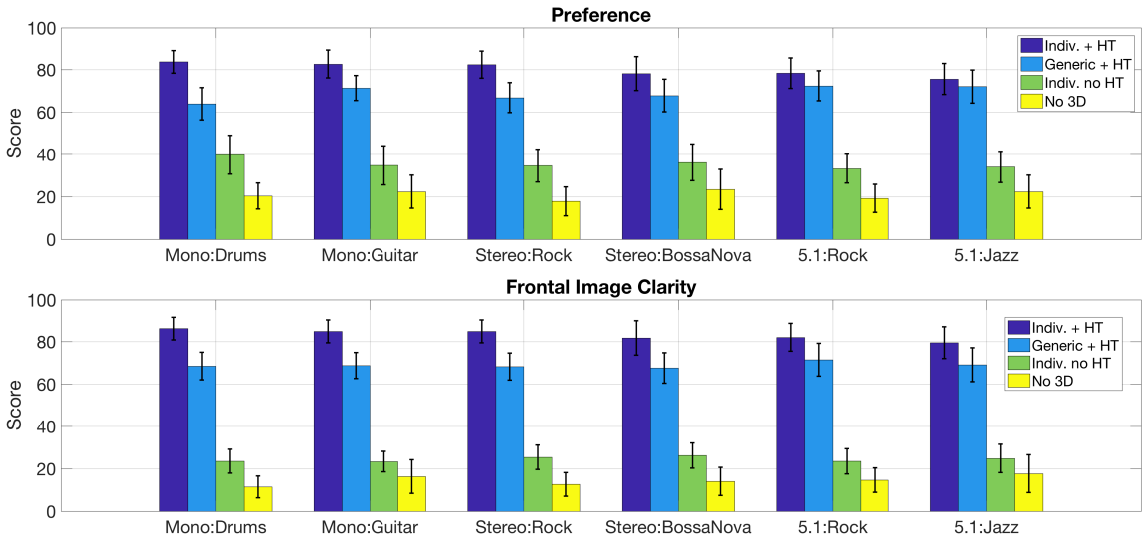


Fig. 1. Results from the binaural listening test are shown for the six sound stimuli. The average population scores for the four listening conditions are shown using a bar plot. The legend labels are as follows: Indiv. + HT – Individualized HRIRs with head-tracking; Generic + HT – generic HRIRs with head-tracking; Indiv. no HT – Individualized HRIRs with no head-tracking; No 3D – no binaural spatial rendering.

anchor was headphone presentation with no spatial audio rendering. Listeners participated in two different trials. In one trial listeners were asked to rate overall preference and in another trial listeners were asked to rate the clarity of the frontal image. Head-tracking was implemented using a Polhemus G4 head-tracking device mounted on the headphones.

2.2. Results

Results of the listening test are shown in Fig. 1. As expected, head-tracking contributed significantly to the listeners' scores because it provides a consistent listening environment in which sound sources are robustly and consistently localized when the head moves. Interestingly, listeners also showed a small, but consistent bias for individualized binaural rendering over generic binaural rendering. The added benefit of individualized binaural rendering is small compared to the benefit of head-tracking. Nevertheless, in listening conditions without a visual reference, there does seem to be a small benefit for individualization in binaural rendering. This would suggest that individualized binaural rendering will play some role when visual stimuli are absent - for example, in augmented spatial hearing conditions using hearables. We hope these data provide some background and motivation for the continued research into morphoacoustics.

3. MORPHOACOUSTICS

We now consider an investigation relating a kernel principal component analysis of ear morphology to a principal component analysis of the directivity of head-related transfer functions (HRTFs) - the spectral representation, i.e., the Fourier transform of HRIRs. We use the SYMARE database [26] but with an interesting twist: we rotated, scaled, and translated all of the ears to match an average, template ear [27]. We then numerically computed the HRIRs for the newly rotated, scaled, and translated ears using FM-BEM. The motivation for such a manipulation is to simplify the morphoacoustic problem. When the ears are mapped via rotation, translation and scaling to the template ear, we expect the acoustics of the ears to be more similar. An additional motivation is that it is well understood that a scaling difference in ear sizes relates to a frequency scaling in the HRTFs as has been well-described by John Middlebrooks [2, 28, 29]. This would indicate that a frequency scaling operation applied to the HRTFs will correct for a scaling of the size of the ear. We have taken a divide-and-conqueror approach to the morphoacoustics problem. We will first consider changes in ear shape that are independent of rotations and scaling. Later on, we will have to account for rotations and scaling, but that is not the focus of this work.

To begin, we briefly review the LDDMM framework. LDDMM [30, 31] is a mathematical framework that can be employed for the registration and morphing of three-dimensional shapes [32, 33]. It is based on theories from functional analysis, variational analysis and reproducible kernel Hilbert spaces. We model a 3D-shape as a mesh with triangular faces, which we refer to as $S(\mathbf{X})$ where $\mathbf{X}$ is the matrix specifying the mesh vertices and S represents the mesh connectivity (the triangular faces). LDDMM models the morphing of $S_1(\mathbf{X})$ to $S_2(\mathbf{Y})$ as a dynamic flow of diffeomorphisms of the ambient space, $\mathbb{R}^3$, in which the surfaces are embedded. This flow of diffeomorphisms, $\phi^{\mathbf{v}}(t,\cdot)$, is defined via the partial differential equation:

$$\frac{\partial \phi^{\mathbf{v}}(t,\mathbf{X})}{\partial t} = \mathbf{v}(t) \circ \phi^{\mathbf{v}}(t,\mathbf{X}) \,, \tag{1}$$

where $\mathbf{v}(t)$ is a time-dependent vector field, $\mathbf{v}(t) : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ for $t \in [0,1]$, which models the infinitesimal efforts of the flow, and $\circ$ denotes function composition. This vector field belongs to a Hilbert space of regular vector fields equipped with a kernel, k_V, and a norm $\|\cdot\|_V$ that models the infinitesimal cost of the flow. In the LDDMM framework, we determine $\mathbf{v}(t)$ by minimizing the cost function, J_{S_1,S_2}:

$$J_{S_1,S_2}\left(\mathbf{v}(t)\right) = \gamma \int_0^1 \|\mathbf{v}(t)\|_V^2 dt + E\left(S_1(\phi^{\mathbf{v}}(1,\mathbf{X})), S_2(\mathbf{Y})\right) \,, \tag{2}$$

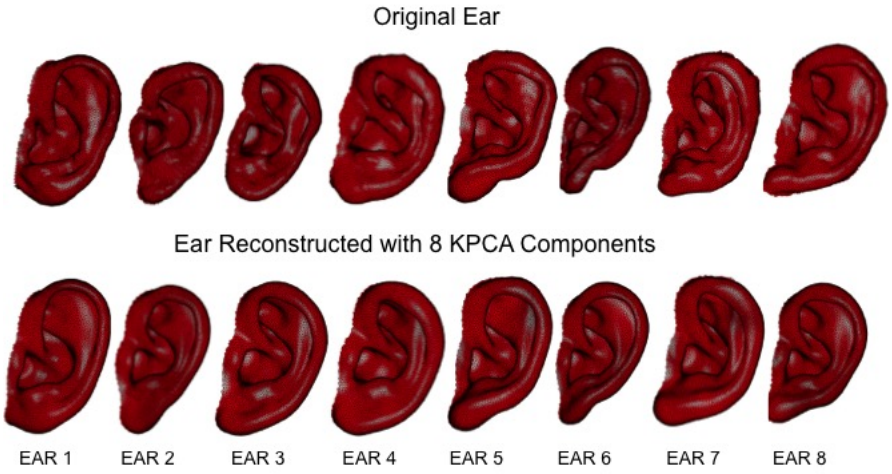


Fig. 2. Top row shows the original ear shapes and the bottom row shows the ear shapes derived from a KPCA representation using 8 principal components.

where E is a norm-squared cost measuring the degree of matching between $S_1(\phi^{\mathbf{v}}(1, \mathbf{X}))$ and $S_2(\mathbf{Y})$. In this work we use the Hilbert space of currents [34, 32] to compute E because it is easier and more natural than using landmarks. The parameter γ is a parameter that sets the relative weight of the two terms in the cost function. In this work $\gamma = 5 \times 10^{-5}$. The optimal $\mathbf{v}(t)$ can be expressed as a sum of momentum vectors, $\boldsymbol{\alpha}_n(t)$, with one momentum vector defined for each of the N vertices in $\mathbf{X}$:

$$\mathbf{v}(t) = \frac{d\mathbf{x}(t)}{dt} = \sum_{n=1}^{N} k_V(\mathbf{x}_n(t), \mathbf{x}(t))\, \boldsymbol{\alpha}_n(t)\,, \qquad (3)$$

where in this work we use the Cauchy kernel.

3.1. Kernel Based Principal Component Analysis (KPCA)

We have previously described the details of a kernel principal component analysis (KPCA) using the LDDMM framework [35]. The KPCA is based on the initial momentum vectors describing the diffeomorphic deformation of the template ear to each ear in the dataset. These initial momentum vectors are taken as a numerical representation of the diffeomorphic deformation. In this paper, we focus on the interpretation of the KPCA applied to the ear morphology. To begin, we use eight principal components to represent ear shape. As we have a dataset of 62 ears, the eight principal components likely form a reasonable subspace. The ability of eight numbers to characterize ear shape is shown in Fig. 3 and works surprisingly well. Recall that the ears have been rotated and scaled to match the template ear so we are only considering changes in ear shape.

3.2. Results

Let us now consider how the eight principal components from the KPCA relate to the changes in ear acoustics. We shall represent ear acoustics based on the HRTF directivity patterns using the spatial frequency response surface [36]. We use standard principal component analysis to analyse the HRTF directivity patterns. Three principal components provides a reasonable representation of the HRTF directivity patterns. We then use simple linear regression to relate the eight principal components from ear morphology to the three principal components related to the HRTF directivity patterns. In Fig. 3, we show the results for three frequencies: 6000 Hz, 8063 Hz, and 9938 Hz. These results seem surprisingly good given the simplicity of the modelling. We have kept the modelling simple to avoid over-fitting and to provide realistic expectations.

4. CONCLUSION

This paper shows variations in ear morphology that commonly occur across a population of ears and the associated changes in the ear acoustics. All of the ears in the dataset have been rotated and scaled to match a template ear. Given the simplified morphological conditions, linear regression between the morphological and acoustic principal components seems to model the data reasonably well.

5. REFERENCES

[1] J. Chen, B. D. van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. of Am.*, vol. 97, no. 1, pp. 1493–1510, 1995.

[2] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. of Am.*, vol. 106, no. 3, pp. 1480–1492, 1999.

[3] C. T. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," in *Proc. of the First IEEE Pacific-Rim Conf. on Multimedia - 2000 Intl. Symposium on Multimedia Inf. Proc.*, December 2000, pp. 235–238.

[4] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* Ieee, 2003, pp. 157–160.

[5] M. A. Ramirez and S. G. Rodriguez, "Hrtf individualization by solving the least squares problem," in *Audio Engineering Society Convention 118*, May 2005.

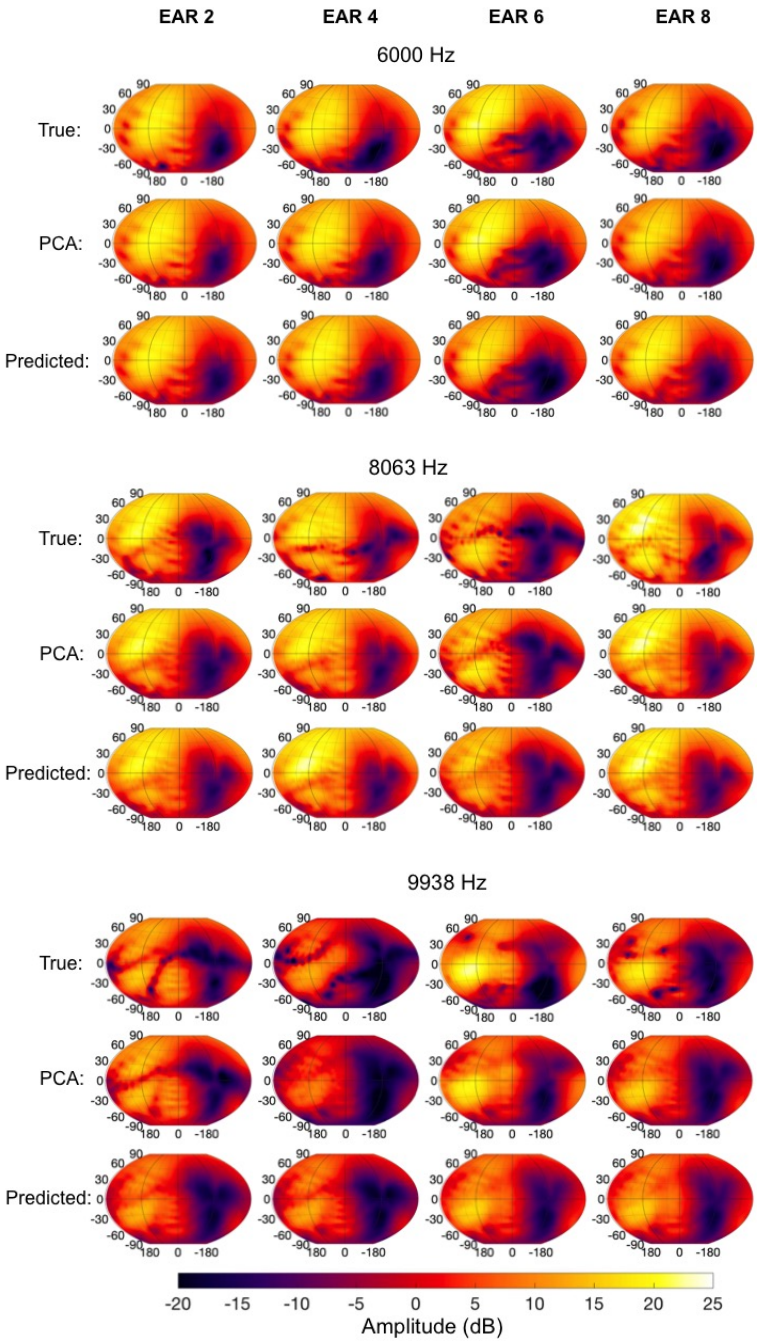


Fig. 3. The spatial frequency response surface for HRTFs are shown for three different frequencies. The true SFRS is shown; followed by the SFRS obtained used three principle components; followed by the predicted SFRS obtained using linear regression.

[6] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: A review," in *Virtual Reality, Second Intl. Conf., ICVR 2007*, 07 2007, pp. 397–407.

[7] S. Hwang, Y. Park, and Y.-s. Park, "Modeling and customization of head-related impulse responses based on general basis functions in time domain," *Acta Acustica united with Acustica*, vol. 94, pp. 965–980, 11 2008.

[8] M. Akagi and H. Hisatsune, "Admissible range for individualization of head-related transfer function in median plane," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Oct 2013, pp. 326–329.

[9] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Virtual autoencoder

based recommendation system for individualizing head-related transfer functions," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2013, pp. 1–4.

[10] J. He, W. S. Gan, and E. L. Tan, "On the preprocessing and postprocessing of hrtf individualization based on sparse representation of anthropometric features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 639–643.

[11] W. Lei and Z. Xiangyang, "New method for synthesizing personalized head-related transfer function," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[12] M. Buerger, S. Meier, C. Hofmann, W. Kellermann, E. Fischer, and H. Puder, "Retrieval of individualized head-related transfer functions for hearing aid applications," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 6–10.

[13] A. Tew, C. Hetherington, and J. Thorpe, "Morphoacoustic perturbation analysis," in *Proceedings of the Joint meeting of the 11th Congrès Français d'Acoustique and the 2012 Annual Meeting of the Institute of Acoustics from UK*, Nantes, France, 2012, pp. 867–872.

[14] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," in *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing)*. Citeseer, 2000, pp. 235–238.

[15] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2408–2411.

[16] R. L. Adams, "Five reasons why virtual reality is a game-changer," March 2016. [Online]. Available: https://www.forbes.com/sites/robertadams/2016/03/21/5-reasons-why-virtual-reality-is-a-game-changer/#5fe88b4a41be

[17] ——, "Virtual reality is about to revolutionize these three industries," Sept. 2016, Forbes, posted 7-September-2016. [Online]. Available: https://www.forbes.com/sites/robertadams/2016/09/07/virtual-reality-is-about-to-revolutionize-these-three-industries/#225c07353035

[18] L. Banks, "The complete guid to hearable technology in 2017," August 2017, Everyday Hearing, posted 18-August-2017. [Online]. Available: https://www.everydayhearing.com/hearing-technology/articles/hearables/

[19] S. Carlile, *Virtual Auditory Space: Generation and Applications*, ser. Neuroscience Intelligence Unit. Austin, Texas, USA.: R.G. Landes Company, 1996.

[20] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*. Cambridge, MA: MIT Press, 1997.

[21] R. H. Gilkey and T. R. Anderson, *Binaural and Spatial Hearing in real and virtual environments*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers, 1997.

[22] Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and K. Kato, *Principles and applications of spatial hearing*. Singapore: World Scientific, 2011.

[23] H. Moller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3, pp. 171 – 218, 1992.

[24] C. T. Jin, A. Corderoy, S. Carlile, and A. van Schaik, "Contrasting monaural and interaural spectral cues for human sound localization," *J. of the Acoust. Soc. of Am.*, vol. 115, no. 6, pp. 3124–3141, 2004.

[25] ITU-R BS.1534-1:2003, *Method for the subjective assessment of intermediate quality level of coding systems*. ITU-R, 2003.

[26] C. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. Tew, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *Multimedia, IEEE Transactions on*, vol. 16, no. 1, pp. 37–46, Jan 2014.

[27] R. Zolfaghari, N. Epain, C. Jin, A. Tew, and J. Glaunes, "A multiscale lddmm template algorithm for studying ear shape variations," in *Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on*, Dec 2014, pp. 1–6.

[28] J. C. Middlebrooks, "Viritual localisation improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. of Am.*, vol. 106, no. 3, pp. 1493–1510, 1999.

[29] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical customization of direction transfer functions for virtual sound localization," *J. Acoust. Soc. of Am.*, vol. 108, no. 6, pp. 3088–3091, 2000.

[30] U. Grenander and M. I. Miller, "Computational anatomy: An emerging discipline," *Quarterly of applied mathematics*, vol. 56, no. 4, pp. 617–694, 1998.

[31] M. Miller and L. Younes, "Group actions, homeomorphisms, and matching: A general framework," *International Journal of Computer Vision*, vol. 41, no. 1, pp. 61–84, 2001.

[32] M. Vaillant and J. A. Glaunés, "Surface matching via currents," in *Information Processing in Medical Imaging*, ser. Lecture Notes in Computer Science, G. E. Christensen and M. Sonka, Eds. Springer Berlin Heidelberg, 2005, vol. 3565, pp. 381–392.

[33] M. Vaillant, A. Qiu, J. A. Glaunés, and M. I. Miller, "Diffeomorphic metric surface mapping in subregion of the superior temporal gyrus," *NeuroImage*, vol. 34, no. 3, pp. 1149 – 1159, 2007.

[34] J. Glaunés, A. Qiu, M. I. Miller, and L. Younes, "Large deformation diffeomorphic metric curve mapping," *International Journal of Computer Vision*, vol. 80, pp. 317–336, 2008.

[35] R. Zolfaghari, N. Epain, C. T. Jin, J. Glauns, and A. Tew, "Kernel principal component analysis of the ear morphology," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 481–485.

[36] C. I. Cheng and G. H. Wakefield, "Spatial frequency response surfaces: an alternative visualization tool for head-related transfer functions (hrtfs)," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 2, Mar 1999, pp. 961–964 vol.2.

MORPHOLOGICAL WEIGHTING IMPROVES INDIVIDUALIZED PREDICTION OF HRTF DIRECTIVITY PATTERNS

Muhammad Shahnawaz,[1,2] *Craig Jin,*[2] *Joan Glaunès,*[3] *Augusto Sarti*[1]*, Anthony Tew*[4]

[1] Dipartimento di Electtronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
[2] School of Electrical and Information Engineering, The University of Sydney, Australia
[3] MAP5, Université Paris Descartes, Sorbonne Paris Cité 75006 Paris, France
[4] Department of Electronics, The University of York, Heslington, York, UK
{first.last}@{sydney.edu.au,polimi.it}
alexis.glaunes@mi.parisdescartes.fr,tony.tew@york.ac.uk

ABSTRACT

In this work, we explore the potential for morphological weighting of different regions of the pinna (outer ear) to improve the prediction of acoustic directivity patterns associated with head-related transfer functions. Using a large deformation diffeomorphic metric mapping framework, we apply kernel principal component analysis to model the pinna morphology. Different regions of the pinna can be weighted differently prior to the kernel principal component analysis. By varying the weights applied to the various regions of the pinna, we begin to learn the relative importance of the various regions to the acoustic directivity of the ear as a function of frequency. The pinna is divided into nine parts comprising the helix, scaphoid fossa, triangular fossa, concha rim, cymba concha, cavum concha, conchal ridge, ear lobe, and back of the ear. Results indicate that weighting the conchal region (concha rim, cavum and cymba concha) improves the predicted acoustic directivity for frequency bands centered around 3 kHz, 7 kHz, 10 kHz and 13 kHz. Similarly, weighting the triangular and scaphoid fossa improves the prediction of acoustic directivity in frequency bands centered around 7 kHz, 13 kHz and 15.5 kHz.

Index Terms— Morphological weighting, Morphoacoustics, LDDMM, Acoustic directivity patterns, Binaural hearing, HRTFs, Kernel principal component analysis, Principal component analysis

1. INTRODUCTION

The acoustic directivity of the human outer ear varies with frequency in an individualized manner depending on the morphological characteristics of the outer ear, head and torso. The directional characteristics of human outer ears are measured in the laboratory as head-related impulse responses (HRIRs) and are described in the frequency domain as head-related transfer functions (HRTFs) [1]. The HRIRs play an important role in developing filters for the synthesis of binaural spatial audio over headphones and have gained renewed interest with recent developments in mixed-reality systems. The relationship between physical morphology and acoustic properties is more generally referred to as morphoacoustics [2–5] and the individualization of binaural spatial audio has a long research history [4, 6–18].

In this work, we continue our exploration of morphable models of both outer ear shapes and outer ear acoustic directivity patterns [19–22]. Our research uses the SYMARE database [23] and is fairly unique in that we have developed a set of 61 affined-matched ears [18]. These ears are matched in scale, rotation, and translation and in the first instance simplify the study of outer ear morphoacoustics. For each of these ears we have numerically computed HRTFs using the fast-multiple boundary element method. The primary new contributions of this work are our demonstrations: (1) that morphological weighting can improve linear regression between model parameters of outer ear shape and model parameters of acoustic directivity patterns and (2) that morphological weighting provides a means to explore the acoustic significance and impact of individual morphological regions of the outer ear.

2. METHODS

Our morphable model for the outer ear shapes is based on the large deformation diffeomorphic metric mapping (LDDMM) framework. Using the LDDMM framework, we have derived a kernel principal component model of ear morphology. With regard to outer ear acoustics, we focus on acoustic directivity patterns and not on HRTFs per se. We derive a standard principal component model of the acoustic directivity patterns. The focus of this work is the application of morphological weighting to our morphable model of ear shape in order to obtain a better understanding and prediction of acoustic properties.

2.1. LDDMM Framework

LDDMM [24, 25] is a mathematical framework that can be employed for the registration and morphing of three-dimensional shapes [26, 27]. In the LDDMM framework we model a 3D-shape as a mesh with triangular faces, which we refer to as $S(\mathbf{X})$, where $\mathbf{X}$ is the matrix specifying the mesh vertices and S represents the mesh connectivity (the triangular faces). Core to this work is the operation of LDDMM mapping which consists in determining the diffeomorphic transformation that morphs an initial shape $S_1(\mathbf{X})$, with $\mathbf{X} \in \mathbb{R}^{N\times 3}$, into a target shape $S_2(\mathbf{Y})$ with $\mathbf{Y} \in \mathbb{R}^{M\times 3}$. The result of this operation is a set of vectors, $\{\boldsymbol{\alpha}_n(0)\}_{1\leq n\leq N}$, defined at the vertices $\mathbf{X}$ and known as the *initial momentum* vectors, that characterize the diffeomorphic transformation in its entirety. We have described the application of the LDDMM framework to the study of ear morphology in detail previously [19–22]. Because the exact details of the LDDMM framework do not play a significant role in this work, we refer the interested reader to these previous studies, rather than provide a full exposition here.

We provide a brief summary of the aspects of the LDDMM

framework that are important to this work. LDDMM models the mapping or morphing of $S_1(\mathbf{X})$ to $S_2(\mathbf{Y})$ as a dynamic flow of diffeomorphisms of the ambient space, $\mathbb{R}^3$, in which the surfaces are embedded. The flow of differomorphisms is characterized by a time-dependent vector field, $\mathbf{v}(t)$ that is determined by minimizing a cost function. Significantly, the time-dependent vector field, $\mathbf{v}(t)$, can be expressed as a sum of momentum vectors, $\boldsymbol{\alpha}_n(t)$, with a momentum vector defined for each of the N vertices in $\mathbf{X}$:

$$\mathbf{v}(t) = \frac{d\mathbf{x}(t)}{dt} = \sum_{n=1}^{N} k_V(\mathbf{x}_n(t), \mathbf{x}(t))\, \boldsymbol{\alpha}_n(t)\,, \tag{1}$$

where in this work we use the Cauchy kernel defined by:

$$k_V(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma_V^2}}\,, \tag{2}$$

for $\mathbf{x}$ and $\mathbf{y}$ in $\in \mathbb{R}^3$. The σ_V parameter is a scale parameter that determines through the kernel, k_V, the range of influence of the momentum vectors $\boldsymbol{\alpha}_n(t)$. Setting σ_V to a larger value increases the coupling in the motion of vertices that are further apart. In this work, $\sigma_V = 10$ mm. As emphasized previously, the initial momentum vectors, $\boldsymbol{\alpha}_n(0)$, determine the diffeomorphic mapping of S_1 to S_2 in its entirety [28].

2.2. Kernel Based Principal Component Analysis (KPCA)

The LDDMM framework can be used to create a morphable model [21] of ears, the essence of which is a template or average ear shape (the details of the calculation are described in [20]), and a set of initial momentum vectors that describe the deformation of the template shape to other shapes in the database. A cornerstone to analysis using the morphable model is kernel-based Principal Component Analysis (KPCA). We use the kernel version of PCA because the space of deformations is Riemannian.

In order to calculate the principal components, we calculate the covariance matrix, $\mathbf{C}$, which expresses the mutual correlation of the different ear shapes in the space of deformations. To compute this matrix we first construct a data matrix $\mathbf{A} \in \mathbb{R}^{3N\times L}$ which contains the initial momentum vectors for the entire population of ears:

$$\mathbf{A} = [\mathbf{a}_1,\ \mathbf{a}_2,\ \ldots,\ \mathbf{a}_L]_{3N\times L} \tag{3}$$

where $\mathbf{a}_l$ denotes the column vector containing all the initial momentum vector coefficients for shape S_l, and L denotes the total number of shapes. We then center the data by subtracting the population average momentum vectors. The centered data matrix, $\hat{\mathbf{A}}$, is given by:

$$\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1,\ \hat{\mathbf{a}}_2,\ \ldots,\ \hat{\mathbf{a}}_L]_{3N\times L} \tag{4}$$

where $\hat{\mathbf{a}}_l$ is the vector of the centered momentum vectors for the l-th shape.

We also form the kernel matrix, $\mathbf{K}$, which contains the values of the kernel function for every pair of vertex positions that comprise the vertices, $\mathbf{X}$, of the template shape T :

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \ldots & \mathbf{K}_{1N} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{K}_{N1} & \ldots & \ldots & \mathbf{K}_{NN} \end{bmatrix},$$

$$\mathbf{K}_{mn} = k_V(\mathbf{x}_m, \mathbf{x}_n)\, \mathbf{I}_{3\times 3}\,, \tag{5}$$

where $\mathbf{I}_{3\times 3}$ denotes the 3×3 identity matrix.

The correlation between two shapes is calculated as the inner product of the initial momentum vectors in the Hilbert space of deformations, V. The correlation between shapes S_i and S_j is given by:

$$c_{ij} = \left\langle \{\boldsymbol{\alpha_n}^{(i)}(0)\}, \{\boldsymbol{\alpha_n}^{(j)}(0)\} \right\rangle_V = \hat{\mathbf{a}}_i^{\mathsf{T}} \mathbf{K}\, \hat{\mathbf{a}}_j\,, \tag{6}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose of a vector or matrix. Thus, the covariance matrix for the entire population of ears, $\mathbf{C}$, is given by:

$$\mathbf{C} = \hat{\mathbf{A}}^{\mathsf{T}} \mathbf{K} \hat{\mathbf{A}} \tag{7}$$

In order to calculate the principal components, as well as the coordinates of the ears in the basis of the principal components, we perform the singular value decomposition of the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathsf{T}}\,. \tag{8}$$

The matrix of the principal components, $\mathbf{U}$, can be then calculated as:

$$\mathbf{U} = \hat{\mathbf{A}}\mathbf{V}\mathbf{D}^{-\frac{1}{2}}\,. \tag{9}$$

Note that the principal components are orthogonal in the Hilbert space of deformations, *i.e.*, $\mathbf{U}^{\mathsf{T}}\mathbf{K}\mathbf{U} = \mathbf{I}$. It follows from Equation (9) that $\hat{\mathbf{A}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{V}^{\mathsf{T}}$ and therefore $\mathbf{D}^{\frac{1}{2}}\mathbf{V}^{\mathsf{T}}$ provides the coordinates of the different ear shapes in the basis of the principal components.

2.3. Weighted KPCA (WKPCA)

The KPCA described above has a limitation. The calculations are such that each mesh vertex contributes equally in the analysis. Nonetheless, the relative areas of various regions of the ear do not necessarily accurately represent the importance of their contribution to the acoustic properties of the ear. For example, the back of the ear likely plays a much less important role acoustically than the concha. In order to explore this issue in more detail, we have apportioned the ear into various sections (refer to Fig. 1) which enables a weighting to be applied during the KPCA. In this case, the kernel function is modified as shown below:

$$k'_V(x, y) = \frac{w(x)w(y)}{1 + \frac{1+\|x-y\|^2}{\sigma_v^2}}, \tag{10}$$

where $w(x)$ and $w(y)$ denote the weights for vertices x and y respectively. This way the kernel function does not just depend on the distance between two vertices but also on the weights associated to them.

2.4. Directivity Patterns and Principal Component Analysis

In this work, we focus on the directivity of the ear, i.e., the pattern of acoustic gain and attenuation across space for a given frequency. In general terms, the directivity patterns become sharper with more features as frequency increases. For a given frequency and ear, we are most interested in the ipsilateral hemisphere of space where the ear has high signal-to-noise ratio and does not suffer from head shadow. Because the directivity pattern on the contralateral side can be varied and noisy, but is likely not significant [29], we have applied gentle spherical Gaussian smoothing (std.: 5.7 degrees of spherical angle) to the directivity pattern on the contralateral side. The directivity pattern data is then treated mathematically as a vector and standard principal component analysis is applied across subjects for a given frequency.

Table 1: Relative vertex weightings, w, and region contributions, $w \times$ area, are shown for three conditions.

Weighting	Back of Ear		Ear Lobe		Scaphoid Fossa		Helix		Cymba Concha		Cavum Concha		Triangular Fossa		Concha Rim		Concha Ridge	
	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area	w	$w \times$ area
Non-Weighted	1	0.25	1	0.05	1	0.03	1	0.24	1	0.03	1	0.08	1	0.05	1	0.21	1	0.06
Concha Weighting	0.75	0.19	0.75	0.04	0.67	0.02	0.75	0.18	1.67	0.05	1.50	0.12	0.80	0.04	1.50	0.32	0.83	0.05
Fossa Weighting	0.60	0.15	0.60	0.03	5.65	0.17	0.60	0.14	0.60	0.02	0.61	0.05	5.80	0.29	0.57	0.12	0.50	0.03

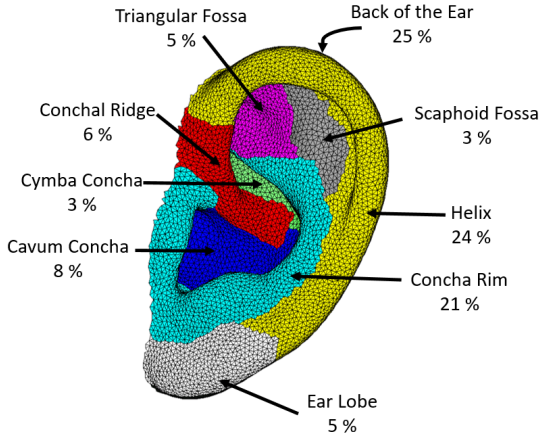


Figure 1: Various regions of the pinna are identified along with their respective fractional contribution to the total surface area.

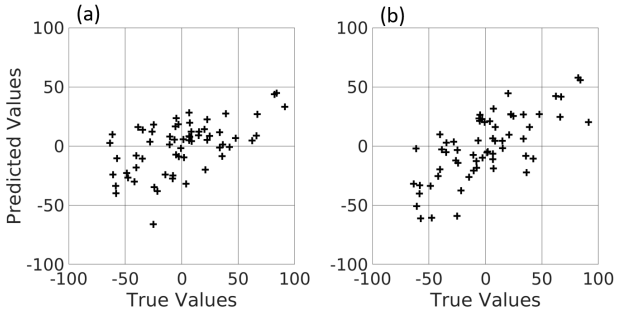


Figure 2: Scatter plots show the predicted and true values for the first principal component of the acoustic directivity patterns corresponding to a frequency of 3938 Hz. Plots are shown for data both without (a) and with (b) morphological weighting. The respective R^2 values are 0.32 and 0.52.

3. RESULTS

Using the SYMARE database, morphological weighting was applied to two regions of the ear separately - the concha and the fossa. Because it is our view that regions with larger surface area have greater influence on the LDDMM mapping algorithm, we determined the surface area for each of the selected regions of the ear. The surface area multiplied by the morphological weight provides a better indication of a region's contribution to the LDDMM mapping. At this stage, only a simple approach to morphological weighting has been taken: e.g., a region's contribution was multiplied by a small, arbitrary factor. Moderate weighting generally proved better, with smaller regions requiring larger morphological weights as is shown in Table 1. It should be noted that the weights are normalized so that the sum of the region contributions (region area multiplied by morphological weight) is unity.

The impact of the morphological weighting was measured simply using linear regression. At this stage, we have only considered the first principal component for the acoustic directivity pattern and the first principal component for the LDDMM ear model. While this assessment is limited, it is important to keep in mind that the ears are affine-matched and only a few principal components are required to adequately describe the acoustic directivity patterns. So a given morphological weighting was evaluated by applying linear regression to find the best linear relationship between the first principal components for the two respective sets of data - the LDDMM ear model and the acoustic directivity pattern. Example results are shown in Fig. 2. In this case, we examine a relatively low frequency (approximately 4 kHz) and find that morphological weighting applied to the concha makes a small improvement.

What is much more interesting is to examine the influence of the morphological weighting as a function of frequency. These data are shown in Fig. 3. For each frequency, we applied the linear regression model to predict the first principal component for a given ear's acoustic directivity pattern. Prediction errors were measured in units of one standard deviation for the population data. For the concha region (see Fig. 3a), we see that morphological weighting results in improvements at various frequencies around 3 kHz, 7 kHz, 10 kHz, and 13 kHz. We interpret the broad range of frequencies as indicating the concha may influence resonance modes at various frequencies. To further support these findings we examined the percentage of cases with improvements and found a similar pattern (see Fig. 3b). The morphological weighting for the fossa produced similar results albeit at slightly higher frequencies (refer to Figs. 3c and 3d. We do not intend for these data to indicate the concha and fossa play independent roles. Rather, the morphological weighting enables one to explore at which frequencies a particular region of the ear may have particular influence on the acoustic directivity of the ear.

The influences of the improvement in the prediction of the first principal component on the resulting acoustic directivity patterns are shown in Fig. 4. Because we only explore the first principal component, all other principal components are held fixed at their true values. We find that the improved prediction of the first principal component does result in small, but visible improvements in the acoustic directivity patterns.

4. DISCUSSION AND CONCLUSION

Morphological weighting provides an interesting tool to explore the morphoacoustic properties of the human outer ear. At this stage, our understanding is limited. Each frequency and each principal component may find improvements with different morphological

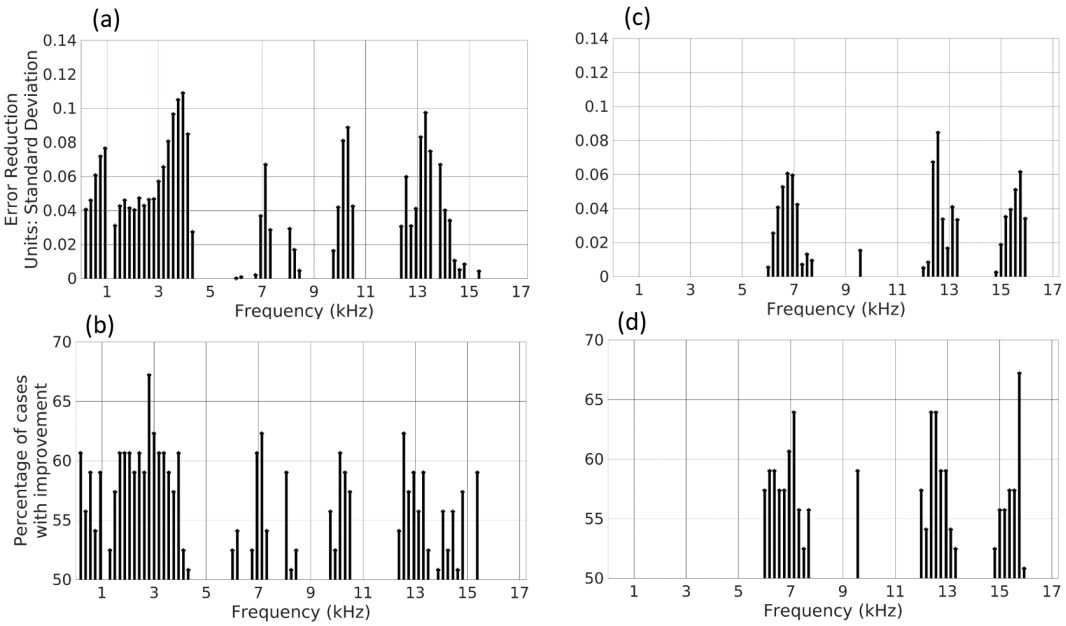


Figure 3: Stem plots show the impact of morphological weighting as a function of frequency. Data for the concha are shown in (a) and (b), while data for the fossa are shown in (c) and (d). The mean reduction in prediction error is shown in (a) and (c) using the population standard deviation as a unit measure. The percentage of ears for which the prediction improved is shown in (b) and (d).

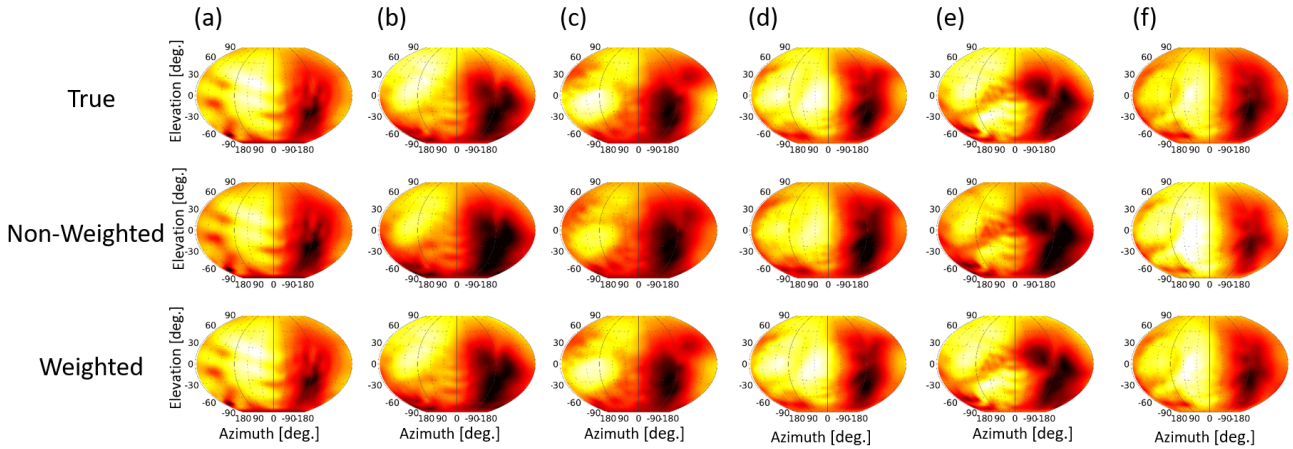


Figure 4: Changes in the acoustic directivity patterns that occur based on the prediction of the first principal component are shown. Azimuth and elevation angles are shown in degrees. The top row shows the true data, the second row shows the data *without* morphological weighting and the third row shows the data *with* morphological weighting. Data are shown for the concha at frequencies: (a) 3938 Hz; (b) 7125 Hz; (c) 10312 Hz; and (d) 13313 Hz. Data are shown for the fossa at frequencies: (e) 6938 Hz and (f) 12563 Hz. Best viewed in color online to see subtle differences.

weightings. We do not find this unreasonable because the acoustic properties of the outer ear result from the structure as a whole and the strength of any particular resonance mode may result from complicated interactions between various morphological elements. We have not yet explored a general optimization algorithm for morphological weighting, nor explored whether additive combinations of morphological weightings would make any sense. It is not even clear how many physical regions one should divide the ear into, nor what the possible interactions may be. Further, it is not yet clear whether a particular morphological weighting should be applied for all ears or just a particular class of ears. Nonetheless, we have made a start and believe there is much more to be learned and will so direct our future attention.

5. REFERENCES

[1] H. Moller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3, pp. 171 – 218, 1992.

[2] A. Tew, C. Hetherington, and J. Thorpe, "Morphoacoustic perturbation analysis," in *Proceedings of the Joint meeting of the 11th Congrès Français d'Acoustique and the 2012 Annual Meeting of the Institute of Acoustics from UK*, Nantes, France, 2012, pp. 867–872.

[3] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* Ieee, 2003, pp. 157–160.

[4] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," in *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing)*. Citeseer, 2000, pp. 235–238.

[5] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2408–2411.

[6] J. Chen, B. D. van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. of Am.*, vol. 97, no. 1, pp. 1493–1510, 1995.

[7] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. of Am.*, vol. 106, no. 3, pp. 1480–1492, 1999.

[8] M. A. Ramirez and S. G. Rodriguez, "HRTF individualization by solving the least squares problem," in *Audio Engineering Society Convention 118*, May 2005.

[9] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: A review," in *Virtual Reality, Second Intl. Conf., ICVR 2007*, 07 2007, pp. 397–407.

[10] S. Hwang, Y. Park, and Y.-s. Park, "Modeling and customization of head-related impulse responses based on general basis functions in time domain," *Acta Acustica united with Acustica*, vol. 94, pp. 965–980, 11 2008.

[11] M. Akagi and H. Hisatsune, "Admissible range for individualization of head-related transfer function in median plane," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Oct 2013, pp. 326–329.

[12] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Virtual autoencoder based recommendation system for individualizing head-related transfer functions," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2013, pp. 1–4.

[13] J. He, W. S. Gan, and E. L. Tan, "On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometric features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 639–643.

[14] W. Lei and Z. Xiangyang, "New method for synthesizing personalized head-related transfer function," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[15] M. Shahnawaz, L. Bianchi, A. Sarti, and S. Tubaro, "Analyzing notch patterns of head related transfer functions in cipic and symare databases," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 101–105.

[16] M. Buerger, S. Meier, C. Hofmann, W. Kellermann, E. Fischer, and H. Puder, "Retrieval of individualized head-related transfer functions for hearing aid applications," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 6–10.

[17] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *2017 International Conference on 3D Immersion (IC3D)*. IEEE, 2017, pp. 1–7.

[18] C. Jin, R. Zolfaghari, X. Long, A. Sebastian, S. Hossain, J. Glaun'es, A. Tew, M. Shahnawaz, and A. Sarti, "Considerations regarding individualization of head-related transfer functions," in *2018 Intl. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 6787–6791.

[19] R. Zolfaghari, N. Epain, C. T. Jin, J. Glaunes, and A. Tew, "Large deformation diffeomorphic metric mapping and fast-multipole boundary element method provide new insights for binaural acoustics," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2863–2867.

[20] R. Zolfaghari, N. Epain, C. Jin, A. Tew, and J. Glaunes, "A multiscale lddmm template algorithm for studying ear shape variations," in *Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on*, Dec 2014, pp. 1–6.

[21] R. Zolfaghari, N. Epain, C. T. Jin, J. Glaunes, and A. Tew, "Generating a morphable model of ears," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 1771–1775.

[22] R. Zolfaghari, N. Epain, C. T. Jin, J. Glauns, and A. Tew, "Kernel principal component analysis of the ear morphology," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 481–485.

[23] C. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. Tew, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *Multimedia, IEEE Transactions on*, vol. 16, no. 1, pp. 37–46, Jan 2014.

[24] U. Grenander and M. I. Miller, "Computational anatomy: An emerging discipline," *Quarterly of applied mathematics*, vol. 56, no. 4, pp. 617–694, 1998.

[25] M. Miller and L. Younes, "Group actions, homeomorphisms, and matching: A general framework," *International Journal of Computer Vision*, vol. 41, no. 1, pp. 61–84, 2001.

[26] M. Vaillant and J. A. Glaunés, "Surface matching via currents," in *Information Processing in Medical Imaging*, ser. Lecture Notes in Computer Science, G. E. Christensen and M. Sonka, Eds. Springer Berlin Heidelberg, 2005, vol. 3565, pp. 381–392.

[27] M. Vaillant, A. Qiu, J. A. Glaunés, and M. I. Miller, "Diffeomorphic metric surface mapping in subregion of the superior temporal gyrus," *NeuroImage*, vol. 34, no. 3, pp. 1149 – 1159, 2007.

[28] M. Vaillant, M. Miller, L. Younes, A. Trouvé, *et al.*, "Statistics on diffeomorphisms via tangent space representations," *NeuroImage*, vol. 23, no. 1, p. 161, 2004.

[29] E. Rasumow, M. Blau, M. Hansen, P. V. Steven, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2012–2025, 2014.