POLITECNICO DI MILANO
Master of Science in
Telecommunication engineering

Electronics, Information and Bioengineering Department

A method for HRTF personalization: Weighted sparse representation synthesis of HRTFs

Supervisor: Prof. Augusto Sarti
Co-supervisor: Dr. Muhammad Shahnawaz

Master graduation thesis by:
Zhu Mo, ID 833301

Academic year 2016-2017

POLITECNICO DI MILANO
Laurea Magistrale in
Ingegneria Delle Telecomunicazioni
INGEGNERIA DELLE TELECOMUNICAZIONI Dipartimento

di Elettronica, Informazione e Bioingegneria

Un metodo per la personalizzazione HRTF: Sintesi di rappresentazione sparse ponderata di HRTFs

Relatore: Prof. Augusto Sarti
Correlatore: Dr. Muhammad Shahnawaz

Tesi di Laurea di:
Zhu Mo, matricola 833301

Anno Accademico 2016-2017

Abstract

Audio is one of the most effective and convenient methods of communicating information. Nowadays, many personal devices like PDAs, mobile, tablets demand spatial audio reproduction to be achieved on personal devices.

One of the most popular ways to achieve spatial audio reproduction is using the headphone to reproduce the sound signal processed by the Head-related transfer function (HRTF). HRTF describes the spectral modifications that are characteristics of a source in a given location with respect to the listener. The time-domain equivalent of this transfer function is known as Head Related Impulse Response (HRIR).

As confirmed by many studies, HRTFs are highly idiosyncratic due to their strong dependence on the listener's anatomy and 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 a specialized equipment. It is necessary for us to find out an alternative technique of HRTF personalization

In this work, we introduce a simple and effective HRTF personalization method. Our method is based on weighted anthropometric sparse representation with preprocessing and postprocessing methods. We follow a strong assumption that the HRTF of a group can be represented using the same representation as is for the anthropometry.

Unlike, previous sparse representation methods, our method assigns different weights to different anthropometric features depending on their relevance.

All the experimentation presented in this study is done on CIPIC database e. We also compared the results of our approach with traditional sparse representation and three different closest-match based approaches. Our results demonstrate that by using only 17 anthropometric features, our method can outperform all previous approaches resulting an average spectral distortion value of 5.53 dBs.

Sommario

Audio e uno dei metodi piu efficaci e piu convenienti per comunicare le informazioni. Al giorno d'oggi, molti dispositivi personali come PDA, cellulari, tablet richiedono riproduzione audio spaziale da realizzare sui dispositivi personali.

Uno dei modi piu diffusi per ottenere la riproduzione audio 3D e utilizzare la cuffia per riprodurre il segnale sonoro elaborato dalla funzione di trasferimento della testa (HRTF). La funzione di trasferimento a testa correlata (HRTF) descrive le modifiche spettrali che sono caratteristiche di una sorgente in una data posizione rispetto all'ascoltatore. L'equivalente del dominio di tempo di questa funzione di trasferimento e conosciuto come Head Related Respulse Response (HRIR).

Come confermato da molti studi, i HRTF sono altamente idiosincratici a causa della loro forte dipendenza dall'anatomia degli ascoltatori e dalle funzioni personalizzate di trasferimento della testa (HRTFs) sono essenziali per la presentazione di audio spaziale autentico tramite rendering binaurale. Tuttavia, la misurazione di HRTF personalizzati per ogni utente e un compito noioso e richiede una attrezzatura specializzata. E necessario per noi scoprire una tecnica alternativa di personalizzazione HRTF

In questo lavoro introdurremo un metodo di personalizzazione HRTF semplice ed efficace. Il nostro metodo e basato su una rappresentazione ponderata antropometrica pesata con metodi di precaricamento e postprocessing. Seguiremo un forte presupposto che l'HRTF di un gruppo puo essere rappresentato utilizzando la stessa rappresentazione che e per l'antropometria.

A differenza dei metodi di rappresentazione sparse precedenti, il nostro

metodo assegna pesi diversi a diverse caratteristiche antropometriche a seconda della loro pertinenza, tutte le funzionalita antropometriche utilizzate possono essere misurate da tre immagini scalate di soggetto.

Tutta la sperimentazione e fatta sul database CIPIC. Abbiamo confrontato i risultati del nostro approccio con la rappresentazione sparsa in precedenza disponibile e individuando i tre diversi approcci basati su match-match. I nostri risultati dimostrano che utilizzando solo 17 funzioni antropometriche, il nostro metodo puo superare gli approcci precedenti con una media valore di distorsione spettrale di 5,53 dB.

Acknowledgments

I would like to express my special thanks of gratitude to Prof.Augusto Sarti who gave me the opportunity to do this thesis. It is a great opportunity for me to work in the lab and I came to know about so many new things. I am really thankful to him.

I would also like to thank my advisor, Muhammad Shahnawaz for his selfless help in guiding and encouraging me throughout my studies and work. It was impossible for me to finish this work without his guidances.

I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Audio is one of the most effective and convenient methods of communicating information. As many personal devices like PCs, mobile, tablets become ubiquitous, the demand to create an immersive aural experience through these personal devices is growing, which makes the spatial audio area become a popular area of research.

One of the most popular topics in the spatial audio area is the spatial audio reproduction. Binaural listening experiences can be created by using headphones, stereo speakers or 5.1 loudspeaker systems [1]. Many spatial audio reproduction systems have been introduced in previous studies. Those systems can be divided into two classes[2]: One is the stereo loudspeaker system, the speakers will reproduce the sound in the corresponding direction. The other one is headphones-based spatial audio reproduction[3], the signal of the sound processed by the Head-related transfer function (HRTF) will be reproduced by the headphone.

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 location of the sound source. The interaction between sound-field and listener's body can be encoded by a complex-valued and direction dependent transfer function, known as Head Related Transfer Function (HRTF). The time-domain equivalent of this transfer function is known as Head Related Impulse Response (HRIR).

Having the HRTF in hand enables us to reproduce the spatial audio over headphones. However, as confirmed by many studies, HRTFs are highly idiosyncratic due to their strong dependence on the listeners's anatomy. It means the best performance can only be guaranteed by using individualized HRTFs [4, 5]. Here we visually from the frequency domain to observe the differences bewteen four subjects'HRTF in the same direction, as shown in Fig. 1.1.

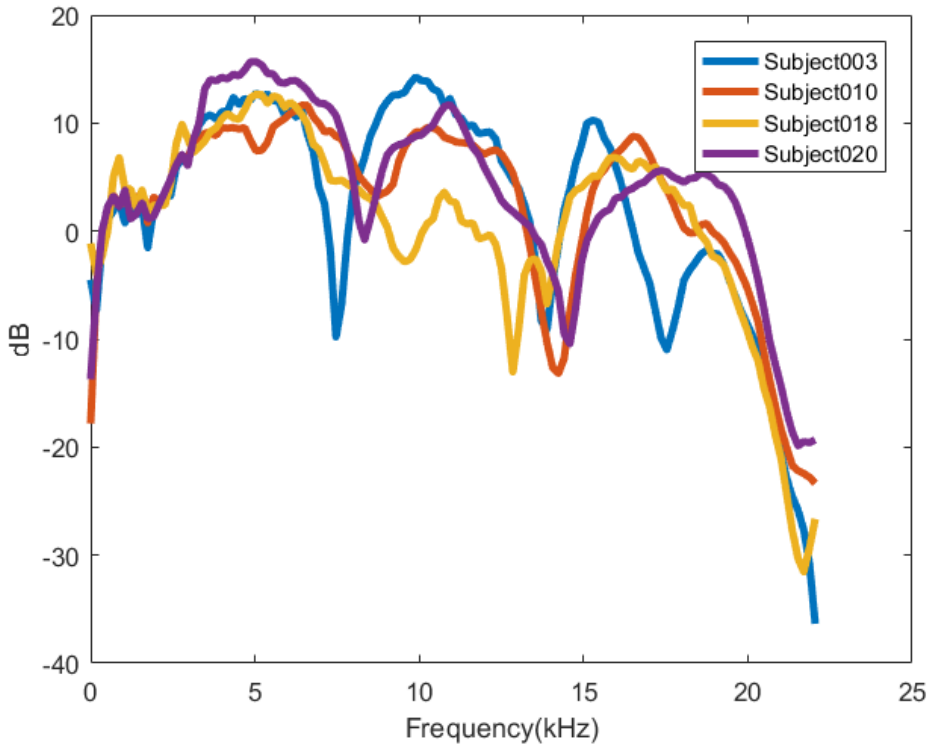


Figure 1.1: The HRTF spectrum of 4 different subjects from CIPIC database in the same direction

Unfortunately, measuring the HRTFs is very cost expensive and time-consuming and is limited to few labs[6, 7, 8]. This result in prevention of

its use in consumer applications. It necessitates to find an easy to use and effective method to produce a personalize HRTF.

Considering the dependence of the HRTFs on the anatomy of the listener, many efforts has been made on personalizing the HRTFs based on anthropometric features. Anthropometric features include the measurements of the anatomy, such as head width, height and depth, pinna and concha height and width etc.

Studies in [9, 10] proposed that the simplest possible approach to achieve the HRTF personalization is to use the anthropometric features to select the closest match from the database of non-individualized HRTFs. However, the closest match doesn't guarantee a good performance in all cases because it can return only one of the non-individualized HRTFs and does not let the user adjust the HRTF magnitudes.

Moreover, studies in [11, 12, 13, 14] find the relationship between the anthropometric features and the HRTFs. Many of these approaches try to find a linear relationship between anthropometric features and HRTFs. While studies [15, 16] investigate nonlinear simple relationships using neural networks. However, the performance of all these approaches depends on the choice of the selected features.

Recently, authors in [17], proposed a new HRTF personalization method based on sparse representation. The assumption is that the magnitude of HRTF can be described using the same sparse representation as the anthropometric features in the training data. Base on this strong assumption, HRTF of a new listener can be synthesized by sparse representation of its anthropometric features and the HRTFs in the database.

Unlike previous sparse representation based approaches, which considered all anthropometric features equally important, we assign weight factors to reflect the relative influence of anthropometric parameters in the calculation of sparse representation. The results show that this method can result in an improved performance in personalization when compared to other methods resulting in an average value of SD 5.53 dbs between actual and synthesized HRTFs.

1.2 Motivation

Broadcasting three-dimensional (3D) video content has been around for a while and geneally a stereo audio is transmitted with the video. video. However, a rich immersive experience demands for synthesizing of the 3D sound [1]. YouTube has deployed this technology in 2016 and its online videos can provide such experience. This is one of the key areas which is expected to grow in the future. Another market where a dire need is perceived is the gaming world. Games with 3D sound synthesis are not a matter for past, rather a pushing demand to provide better virtual sound experience is growing day-by-day.

Nowadays, many personal devices like PDAs, mobile, tablets demand spatial audio to be delivered on personal devices. One of the most popular ways to achieve spatial audio reproduction is using the headphone to reproduce the sound signal processed by the Head-related transfer function (HRTF)[18]. HRTF describes the spectral modifications that are characteristics of a source in a given location with respect to the listener. So it plays an important role in spatial audio reproduction techniques.

However, as the strong relationship between HRTF and personal anthropometry, even a small difference of anthropometric shape and size can create a significant influence on HRTFs for sound location. Perceptual distortions may occur in spatial hearing using generic HRTFs without the individual difference. Therefore, it is necessary to personalize HRTFs. As we already know the difficulty of HRTF measurements and the strict requirement of measurement experiment. The measurement of HRTF can not be widely used, it is necessary to look for an alternative way to do the HRTF personalization.

Due to the inherent relation between HRTFs and anatomy of a person, anthropometric data are widely used for HRTF personalization[9]. The process requires usually to create a model and train it on the databases of non-individualized HRTFs in hand. Then this model can be used to select the closest matching HRTF from the database or synthesize a new HRTF.

There are two main problems of HRTF personalization based on anthropometric measurements:

First is the complexity of the anthropometric measurements. Although it's much simpler for us to measure the anthropometric features than to measure the HRTF, Some anthropometric data, such as the pinna rotation angle and pinna are angle etc, still need to be measured in precise measurement tools and it is still a time-consuming work. These limits the widely used of HRTF personalization. It demands to select the key anthropometric features and to measure features in a simple way.

Second is the accuracy of the HRTF personalization. As many former studies in[11, 19, 20], the closest match based approach can lead to a significant difference between the original HRTF and the matched HRTF. This may lead to a blurred sound image and result in many psychoperceptual errors.

1.3 Objective

The objectives of this thesis are summarized as follows:

• To apply a weight calculation approach to reflect the different relevance of anthropometric features.

• To use the most relevant and easily measurable anthropometric features.

• To supply a more accurate HRTF personalization method based on anthropometric features.

1.4 Organization of the thesis

Chapter 2 provides an overview of the spatial audio, HRTF, the measurement of HRTF. and the state of articles of HRTF personalization approach. Furthermore, different types of HRTF personalization methods will be discussed.

Chapter 3 explains our HRTF personalization approach. First, the selection and acquisition of anthropometric feature will be introduced. Then

the preprocessing methods for anthropometry features and HRTFs will be discussed. Last is the calculation of sparse representation with weighted anthropometric features, including the calculation of anthropometric weights.

Chapter 4 talks about the implement and experiment steps, including database selection and evaluation criteria selection and the performance of our approach. Also, we compare the results of our approach with previously available sparse representation and finding the closest-match based approaches.

Finally, chapter 5 summarizes the major contribution of this thesis and suggests future work to be developed.

Chapter 2

Background

2.1 Spatial Audio

Recently a lot of interest has been seen in the spatial audio area. The definition of spatial audio can be found in [2]: Spatial audio is the perception of sound in 3D space and anything else related to such a perception, including sound acquisition, production, mastering, processing, reproduction, and evaluation of the sound, it can also be called as three dimensional (3D) audio.

One of the most popular topics in the spatial audio area is the sound localization. Humans are capable to perceive and localize sound in 3D-space. many previous studies describe the three dimensions of sound localization as distance, azimuth, and elevation[2], as shown in Fig. 2.1. The distance refers to the length of the direct path from the sound source to the centre of the head. The horizontal plane is the plane which is horizontal to the ground at ear-level height. Median plane is a vertical plane which is perpendicular to the horizontal plane and with the same origin at the center of the head. The azimuth is the angle between median plane and the path from the sound source to the centre of the head. The elevation is the angle between horizontal plane and the path from the sound source to the centre of the head[2]. These three dimensions can also be divided into two aspects of perceived localization such that the direction of the sound source(include azimuth and elevation) and the distance between the centre of head and sound source.

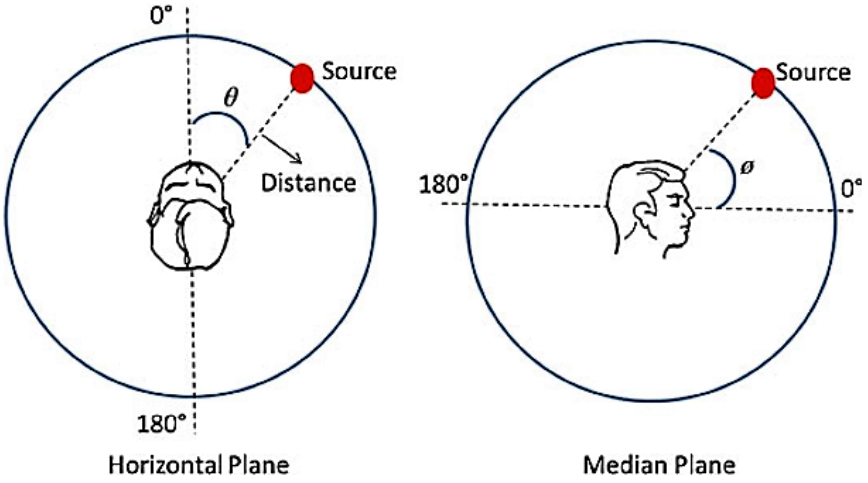


Figure 2.1: The coordinate system for sound localization [2]

For sound localization, human brains combine various cues from perceived sound and other sensory information, such as visual images[2]. Interaural time difference (ITD) and Interaural level difference (ILD) are the most important cues on sound direction localization[21]. ITD refers to the difference of time that the sound arrives the left ear and right ear from the sound source and ILD represents the difference in loudness and frequency distribution between the two ears, as illustrated in 2.2.

As [22] introduced, ITD represents the ability of a person to measure the interaural difference at low frequencies. ILD is mainly caused by the attenuation of the sound levels in the ear further to the source due to the head shadowing effect, compared to the ear nearer to the source. Therefore, ITD is more important in low frequency and ILD is more important in high frequency. [2].

However, ILD and ITD are not guarantee to work for all cases. As we can obtain the same value of ITD and ILD from the sound source in a conical surface[23], it will lead to many perceptual errors, such as front-back confusions[24]. The spectral cue can be used to help perceive the accurate elevation directions. Due to the relevance between spectral cues and the anthropometry of the listener, the idiosyncrasy anthropometry of the subject makes spectral cues individual.[22]

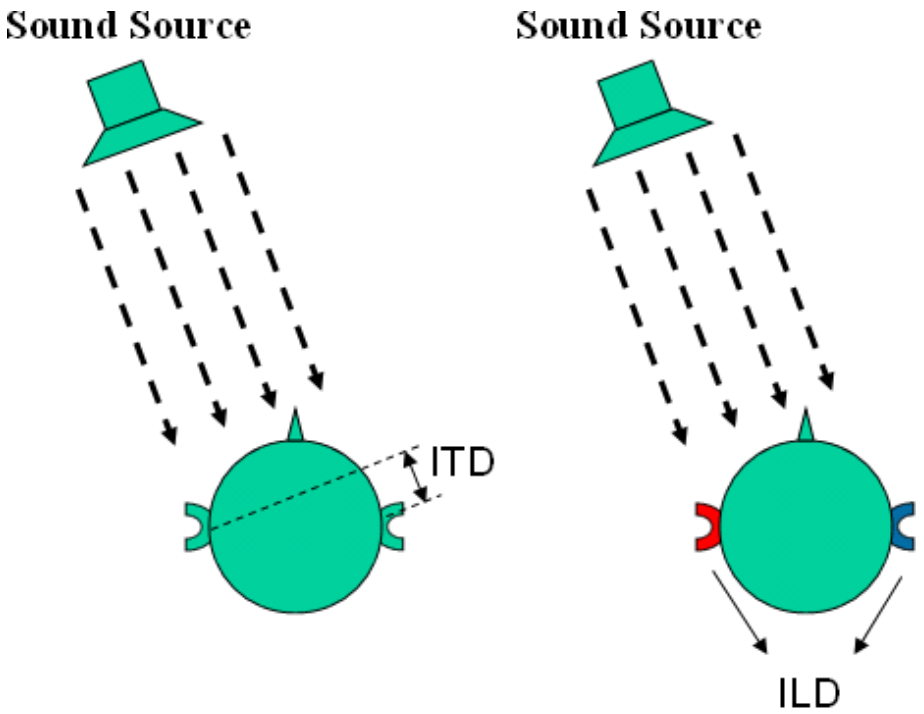


Figure 2.2: The example of ITD and ILD [2]

2.2 Head related transfer function

The sound waves emitted by the sound source are scattered by the head, pinna, and torso. The physical process can be regarded as a acoustic filtering system, whose characteristics can be obtained by the system's frequency domain transfer function description. HRTF is the frequency domain transfer function of this acoustic filtering system. In [25], HRTF has been defined as:

$$H_L = H_L(r, \theta, \phi, \omega, a) = P_L(r, \theta, \phi, \omega, a)/P_0(r, \omega) \tag{2.1}$$

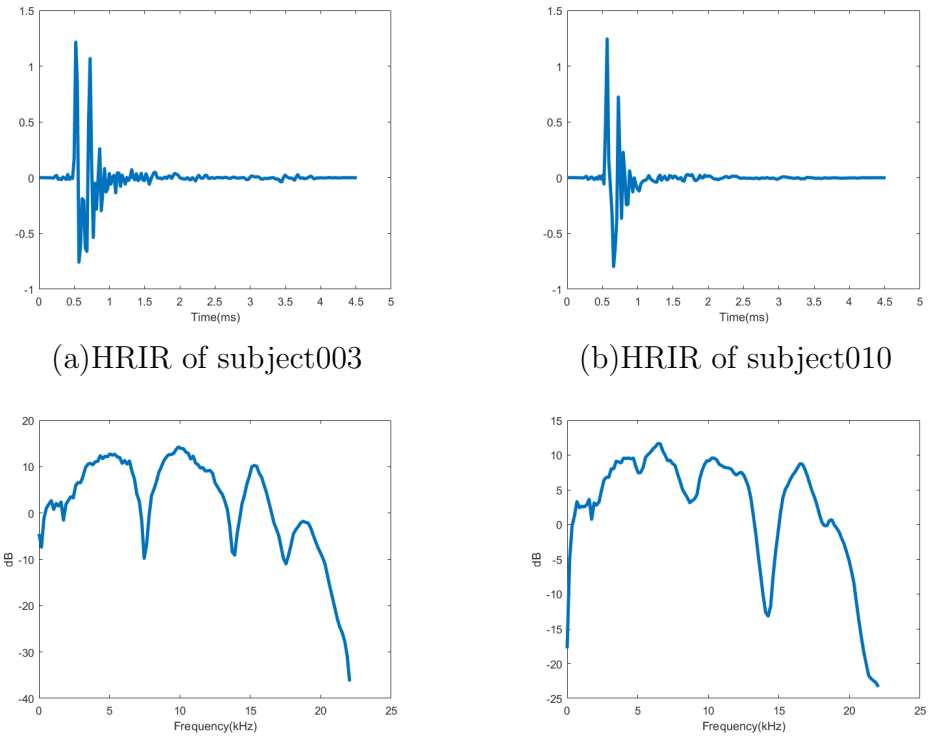
$$H_R = H_R(r, \theta, \phi, \omega, a) = P_R(r, \theta, \phi, \omega, a)/P_0(r, \omega) \tag{2.2}$$

where P_L and P_R are sound pressures in the listener left and right ear. P_0 is a sound pressure at the head center point without the head.

In general, H_L and H_R is the function of the elevation angle of the sound source θ, the azimuth angle of the sound source ϕ, the distance from the sound source to the center of the head r, and the angular frequency of the acoustic wave ω. HRTF is independent on the distance of the sound source r only if the source is in the far-field.

Otherwise, the anatomy of different people are not the same, such as head width, height and depth, pinna and concha height and width etc, which lead to the idiosyncrasy of people's HRTF. The parameter a in the function represents the individual features.

The time-domain equivalent of this transfer function is known as Head Related Impulse Response (HRIR), It is the inverse Fourier transform of the HRTF.



(a)HRIR of subject003 (b)HRIR of subject010

(c)HRTF[dB]of subject003 (d) HRTF[dB]of subject010

Figure 2.3: HRIR and HRTF for two subjects in the same direction [19]

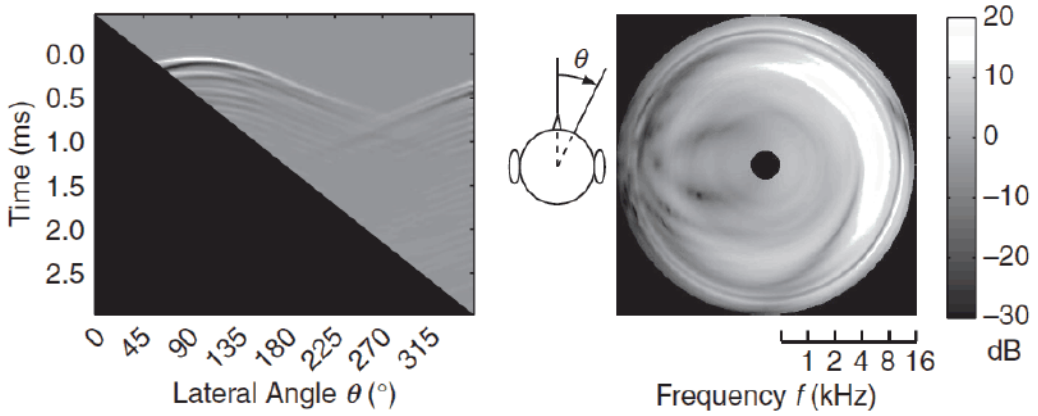


Figure 2.4: HRIR and HRTF(azimuth dependence) for one subject [21]

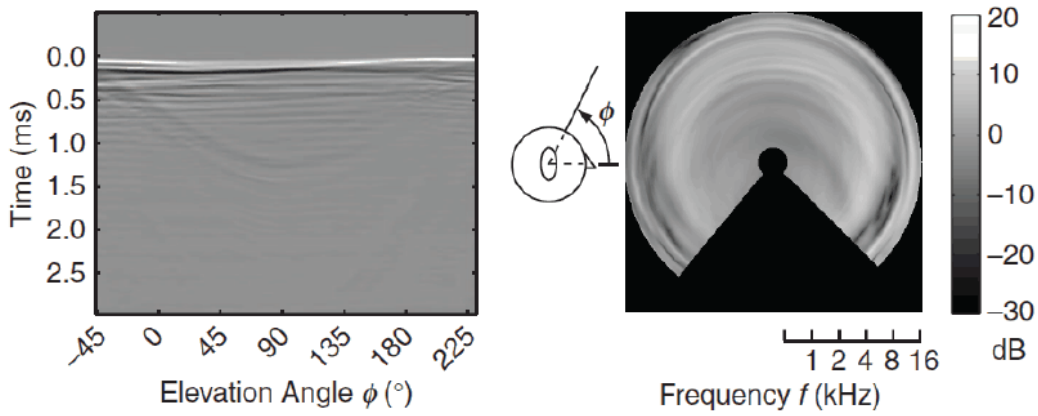


Figure 2.5: HRIR and HRTF(elevation dependence) for one subject [21]

2.3 The measurement of HRTF

Having the HRTF in hand enables us to reproduce the spatial audio over headphones. However, as confirmed by many studies, HRTFs are highly idiosyncratic due to their strong dependence on the listeners's anatomy[9]. Even a small difference of anatomy can create a significant influence on HRTFs for sound location. Perceptual distortions may occur in spatial hearing using generic HRTFs. Therefore, the best performance can only be guaranteed by using individualized HRTFs [4, 5].

As early as the 1940s, researchers have already tried to measure HRTF, but most of these measurements were just focus on the horizontal plane or median plane, or can not be used as general data because of the low accuracy of measured data[26]. Nowadays, The measurement of HRTF is currently a subject of much research, a variety of measurement methods for acoustic system transfer functions can be used for HRTF measurements. The actual measurement can obtain high precision personalized HRTF. Many research institutions and universities have already set up some HRTF database, such as CIPIC database[6], SYMARE database[8].

HRTF is usually measured in a acoustically conditioned environment. An anechoic room with speakers placed in a spherical pattern is used. The listener is seated at the center point of this virtual sphere, such that the imaginary line passing through his ears is through the diameter[6]. In principle, all speakers are at equal distance from the listener, as illustrated in Fig.2.6.

High-quality earphones are placed inside each ear canal of the listener, which record the stimulus generated from the speakers[27], as shown in Fig.2.7.

Once the setup is ready, speakers are energized one-by-one with a wide band signal based on Golay codes. The goal of such a signal is to cover all the frequencies and also discriminate the fine delays in propagation using its auto correlation property[28]. Measurements are taken and are post processed to generate the Head Related Impulse Response (HRIR)[6]. In this process, all the artifacts of the room are mitigated and a pure impulse response is

Figure 2.6: HRTF measurement setup at Sound and Music Computing (SMC) Lab, COMO, Polimi

Figure 2.7: The position of Microphone in HRTF measurement [28]

obtained. Precision and accuracy of such a response matters a lot as the 3D sound generation solely relies on it. This HRIR is transformed into frequency domain, resulting in an HRTF.

However, there are some limitations of HRTF measurements, such as cost-consuming, time-consuming and lack of flexibility, which limited the widespread use of HRTF. For this reason, finding an alternative method to personalize HRTF is important.

2.4 HRTF personalization

2.4.1 HRTF personalization using closest-match

Closest-match base on the differences of anthropometric features differences

In order to achieve prediction of personalized HRTF rapidly, a closest-match method base on the CIPIC database has been introduced in [6]. The brief idea of this method is to find out the subject whose physiologic structure is most

similar to the test one and return this subject's HRTF as the personalized HRTF of the test subject.

The method is based on finding the closet-match to the outer ear shape of the test subject using a set of seven pinna features, as illustrated in Fig.2.8.

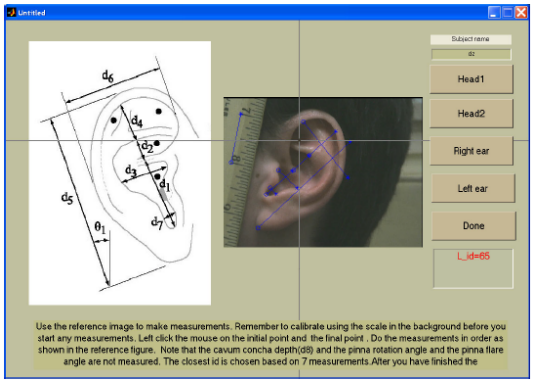


Figure 2.8: The pinna features used in HRTF personalization [9]

This method achieved the purpose of quick personalization. However, due to the diversity of anatomy, using pinna shapes involved in matching, which will lead to some errors.

Also, as we have pictures of the ear in the CIPIC database. So some image processing algorithms can be use to match the pictures and find the closest match

Closest-match base on the PCA

In [20], the closest-match method has been improved by using principal component analysis(PCA)[29] and correlation analysis to select out the key anthropometric features. Then, these key anthropometric features will be used to select out the closest HRTF in CIPIC database.

In this methods, PCA has been used to calculate the relevance of the anthropometric features. As the result of principal component analysis, the PC weights can be used to run the regressions on the anthropometric features and select the more relevant features. The selected features can be used to make the closest match in the database.

The parameter E has been defined as the measurement of similarity in [20]:

$$E = \sum_{m=1}^{M} \frac{(\hat{d}_m - d_m)^2}{\sigma_m^2} \tag{2.3}$$

where $\hat{d}_m$ corresponds to the m-th anthropometric feature of the test subject, d_m corresponds to the m-th anthropometric feature of the subject in database and σ_m is the variance of the m-th anthropometric feature in the database.

However, due to the diversity of anthropometric feature, using a small number of anthropometric measurements involved in matching, which will lead to some errors. More importantly, the closest match doesn't guarantee a good performance in all cases because it can return only one of the non-individualized HRTFs and does not let the user adjust the HRTF magnitudes.

Closest-match base on the notch frequencies

Authors in [11] exploits the use of a revised pinna reflection model on a 2-D image as a selection mechanism for HRTFs.

According to McAulay-Quatieri partial tracking algorithm, the three main frequency notches of a specific median-plane HRTF can be extracted by calculating the distance between a point lying approximately at the ear canal entrance and each point lying on the three pinna contours thought to be responsible for pinna reflections.

Then, with the HRTF data in CIPIC database, each subject can obtains three frequency notches. The test subject need to take the picture of ear, three contours'notch can be confirmed from the picture. Compare these three notches with the notches in the database, the HRTF set in the database whose mismatch is the lowest will be selected as the closest-match HRTF.

2.4.2 HRTF personalization using sparse representation

Recently, authors in [17] proposed a new HRTF personalization method based on sparse representation.

The sparse representation method based on a strong assumption that the magnitude of HRTF can be described by the same sparse representation as the anthropometric features in the training data.

Base on this assumption, a new subject's HRTF can be synthesized by sparse representation of its anthropometric features and the HRTFs in the database.

The basic steps of this method are shown in Fig. 2.9: First, looking for a sparse vector β that represents the test subject's anthropometric features as a linear combination of the anthropometric features from the database. Then apply this sparse vector on the HRTF from the database to synthesize the HRTF of the test subject.

The experiment results in [17]show that this method can improve the personalization performance.

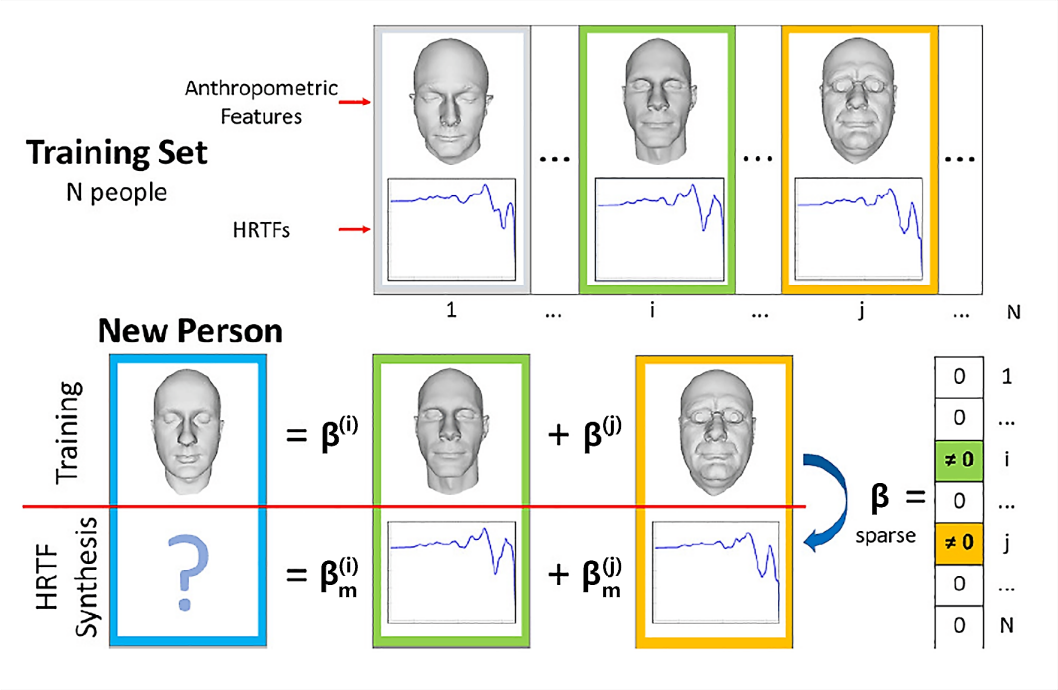


Figure 2.9: Block diagram of the proposed sparse representation[17]

Chapter 3

Overview of the System

In this chapter, we will describe the architecture of the HRTF personalization approach based on the weighted sparse representation of anthropometric measurements step by step, as illustrated in Fig.3.1

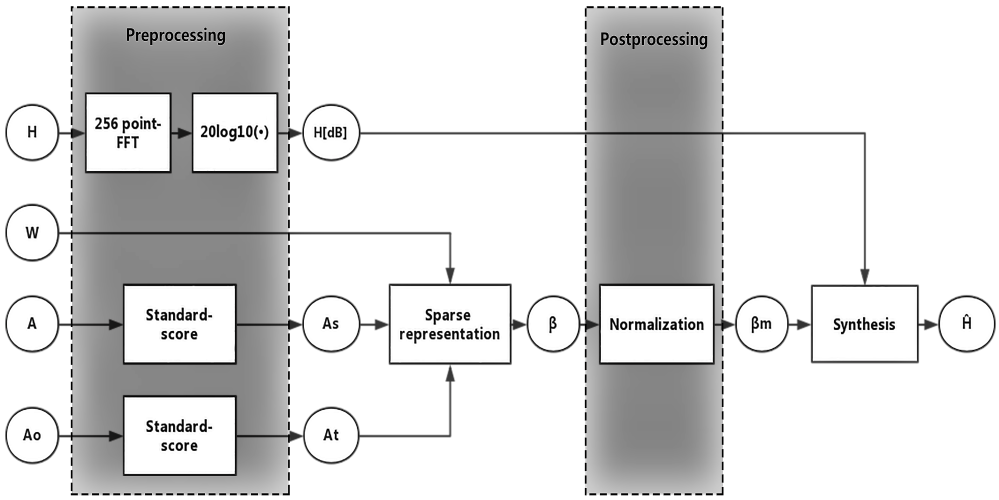


Figure 3.1: Block diagram of HRTFs personalization using weighted sparse representation of anthropometric features

In all previous sparse representation techniques [17, 30], the anthropometric parameters are considered equally relevant. Also, authors in [17, 30] consider both anthropometric measurements of left ear and right ear as a

single vector when determining a new subject's sparse representation. As a consequence, each subject only has one sparse representation of anthropometric features.

However, it is not the case. Some of the features are more relevant than the others[11]. For example, pinna features are the mostly more relevant than the shoulder. Also, the anthropometric measurements of left ear and right ear can be different.

Our contributions are twofold. First, assign weights of the anthropometric features using partially on-off strategy using the approach described in [19] and use these weights to devise a weighted sparse representation approach. Second, give each subject two separately sparse representation for both left ear and right ear in our work.

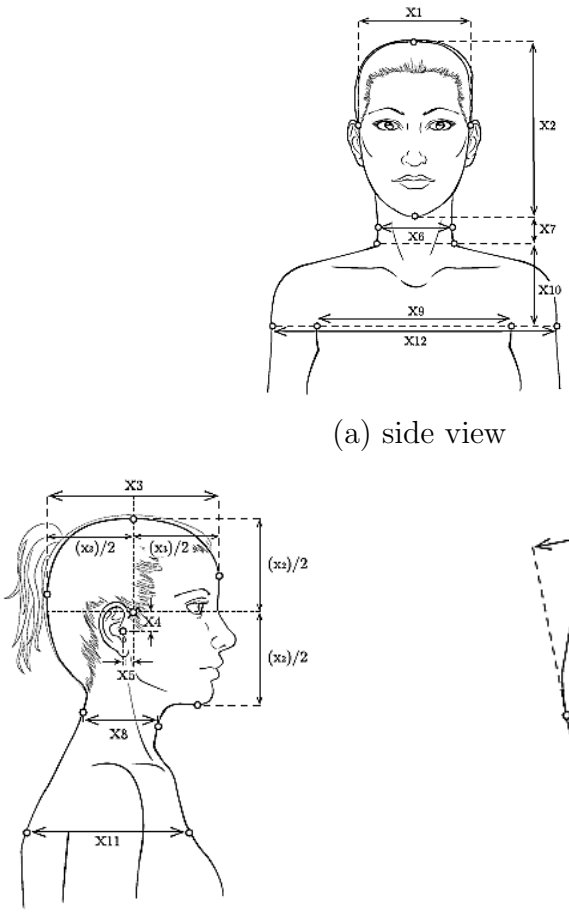
As the anthropometric measurements are on different scales and are from different ranges, some preprocessing and postprocessing methods for anthropometric data and HRTF data has been introduced in [30]. Then, authors in [30] presented the best performance combination of those processing methods. We follow these suggested preprocessing and postprocessing methods in our work.

Now we will discuss the detail of each step in this system.

3.1 Anthropometry

3.1.1 Anthropometric feature selection

The study in [19] reported that 19 anthropometric parameters (one pinna) can be directly measured using only three scaled pictures as illustrated in Fig. 3.2. All these 19 anthropometric parameters are listed in Table 3.1. However, the value of x_5 and x_7 are usually too small to be measured in the picture. For this reason, we only use the remaining 17 anthropometric parameters.



(a) side view

(b) front view

(c) ear area

Figure 3.2: Anthropometric parameters can be measured from side view, front view and ear area[19]

3.1.2 Anthropometric feature acquisition

As authors in [6] has explained that anthropometric measurements can be obtained from still photos by placing a measurement tape behind the ear while capturing the photos, as illustrated in Fig. 3.3. It allows those distances and sizes not included in the anthropometry data to be determined. Therefore, we can calculate the ratio of the picture and then use this ratio and the length of anthropometric features on the picture obtain the measurement of

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 3.1: 19 Anthropometric parameters can be measured from scaled picture

anthropometric features as follow:

$$Ratio = \frac{l_s}{l_p}, \quad A = Ratio \cdot A_p \tag{3.1}$$

where l_s represents the standard length of the item. l_s is the measured length on the picture. A_p is the measured length of anthropometric features on the picture. A is the measurement of anthropometric features.

(a)5cm long measurement tape (b)ruler

Figure 3.3: The example of scaled picture [6]

Recently, authors in [31, 19]presented another method of anthropometric feature acquisition from the three scaled pictures. A photographic studio has been set up and all the parameters of this photographic studio are fixed, such as the type of the camera, the distance from testing subject to the focal plane, the power of fluorescent lamps, the distance between halogen lamps etc. Details are shown in Fig. 3.4

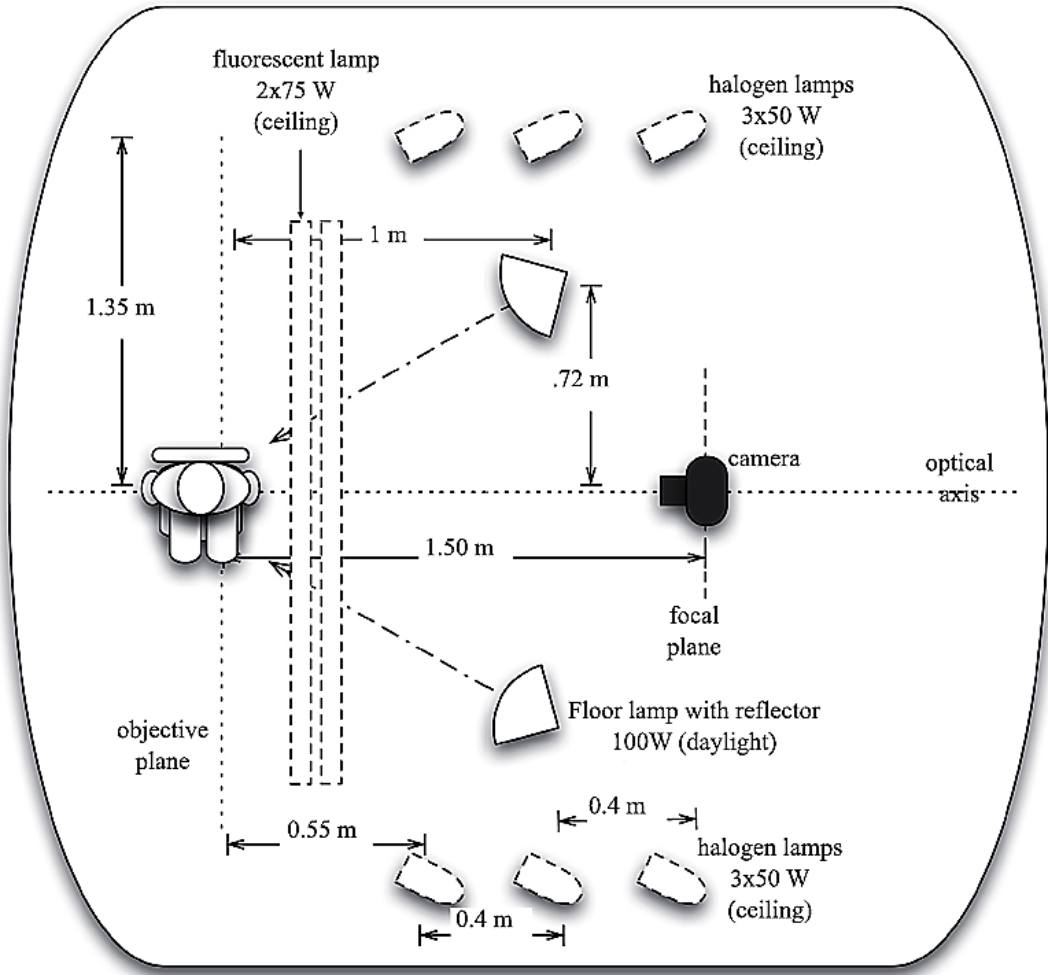


Figure 3.4: Top view of the photographic studio [19]

After setting up this photographic studio, all test subjects only need to take pictures of three different views. Due to the ratio of each photograph taken in this photographic studio has already been known, all 19 measurements of anthropometric features can be calculated.

However, even though the method introduced by [19] can achieve higher accuracy in anthropometric measurements, it is still cost and time-consuming to set up the photographic studio.

3.2 Preprocessing for Anthropometry Feature

According to many HRTF databases with anthropometric measurements, the scale of different anthropometric features are different. It is necessary for us to adjust those anthropometric data measured on different scales to a notionally common scale, which means normalization[32]. Normalization can also reduce the complexity and error in the calculation of weighted sparse representation.

Authors in [30]introduced three different types of preprocessing methods for anthropometric features and use the normalized anthropometric parameters A_t instead of using their scalar magnitudes A_o directly, including:

• **Min-Max**: each anthropometric feature subtracts the minimum value in the set of anthropometric features and divided by the difference between the maximum value and the minimum value in this set of anthropometric feature.

$$A_t = \frac{A_o - min[A_d]}{max[A_d] - min[A_d]} \tag{3.2}$$

$$A_s = \frac{A - min[A_d]}{max[A_d] - min[A_d]} \tag{3.3}$$

where $A_d = [A \quad A_o]$. A corresponds to the original anthropometric parameters of all subjects in the database. A_s is max-min normalized anthropometric parameters A in the database, the value of each element in A_s should in the range of 0 to 1 .

• **Standard score**: each anthropometric feature subtracts the mean value in the set of anthropometric feature and divided by the standard deviation of this set of anthropometric feature, as insulated in Fig. 3.5 .

$$A_t = \frac{A_o - mean[A_d]}{std[A_d]} \tag{3.4}$$

$$A_s = \frac{A - mean[A_d]}{std[A_d]} \tag{3.5}$$

where $A_d = [A \quad A_o]$. A represents the original anthropometric parameters of all subjects in the database. A_s is the result of standard score of the anthropometric parameters A in the database.

• **Standard deviation**: each anthropometric feature divided by the standard deviation of this set of the anthropometric feature directly.

$$A_t = \frac{A_o}{std[A_d]} \tag{3.6}$$

$$A_s = \frac{A]}{std[A_d]} \tag{3.7}$$

where $A_d = [A \quad A_o]$. A corresponds to the original anthropometric parameters of all subjects in the database. A_s is the standard deviation of the anthropometric parameters A in the database.

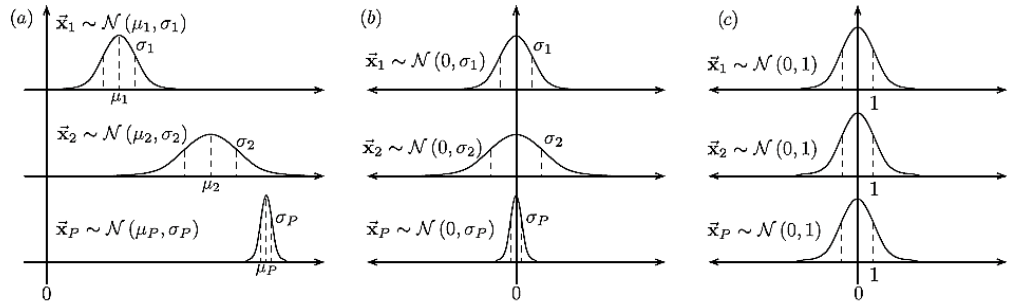


Figure 3.5: Standard score normalization of anthropometric data distribution (a) original distributions, (b) after subtracting the mean and (c) after dividing by the standard deviation [19]

As [30] already compared these three type of preprocessing methods for anthropometric data and introduced that use the standard score of anthro-

pometric parameters has the best performance. That's why we select the calculation of Standard score as the preprocessing method in this work.

3.3 Preprocessing for HRTF

As many studies have suggested that the performance of personalization method heavily depends on the choice of initial representation of HRTF[30, 11].Except using HRTF magnitude directly, log magnitude and power of HRTF can also be used as the representation of HRTF. It is necessary for us to choose a type of preprocessing method for HRTF data.

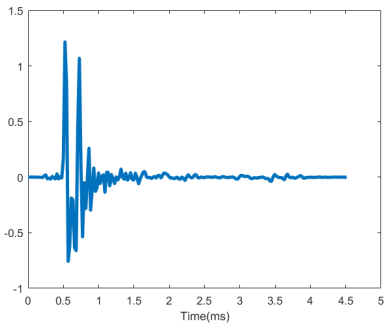
- **Log magnitude of HRTF**:

$$H_{[dB]} = 20 \log_{10} |H| \tag{3.8}$$

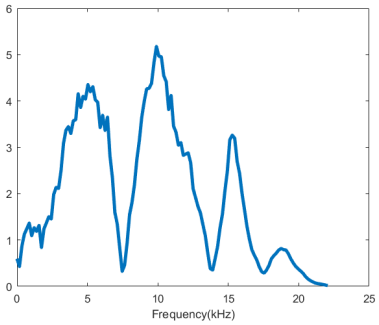
- **Power of HRTF**:

$$H_{power} = [H]^2 \tag{3.9}$$

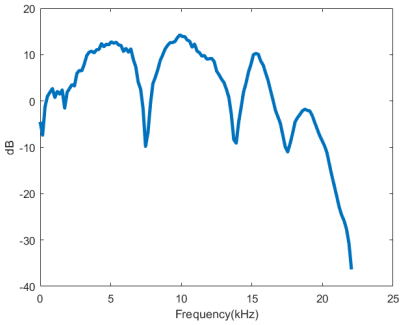
The steps of preprocessing method of HRTF data as shown in Fig. 3.6. First, we calculate HRTFs from HRIRs by computing the 256 point FFT, then we follow the suggestion in [30] by using log-scale magnitude can result in an improved performance. In our work, we used dB scale HRTFs instead of complex amplitude HRTFs.



(a)HRIR



(b)HRTF(256 point FFT)



(c) HRTF[dB]

Figure 3.6: Preprocessing steps of HRTF data: from (a) to (c)

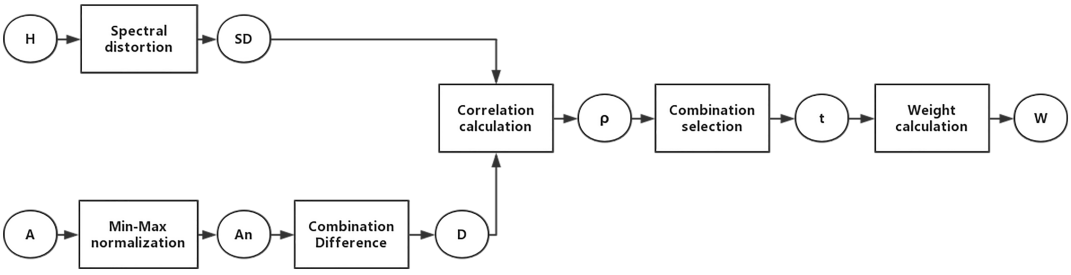


Figure 3.7: Block diagram of weight calculation

3.4 Sparse representation method for HRTF personalization

3.4.1 Introduction of weights for anthropometry features

Unlike the previous sparse representation based approaches [17, 30], which considered all anthropometric features equally important, we assign weight factors to the anthropometric parameters depending on their relevance. The weights are calculated using the approach presented in [19], as illustrated in Fig.3.7. For each subject in CIPIC database, we use 25 out of 27 anthropometric parameters in weight calculation(x_{14} height and x_{15} seated height are excluded). We did this to achieve a more general representation of the relevance of the anthropometric features. In order to adjust anthropometric parameters on different scales to a common scale, we normalize these 25 sets of anthropometric measurements with min-max method:

$$A_n^{(i)} = \frac{A^{(i)} - min[A^{(i)}]}{max[A^{(i)}] - min[A^{(i)}]} \quad \forall i = 1, 2, ..., 25 \tag{3.10}$$

where $A_n^{(i)}$ corresponds to the i-th set of normalized anthropometric parameters, $A^{(i)}$ corresponds to the i-th set of anthropometric parameters.

In order to obtain all possible combinations of 25 anthropometric parameters, partially on-off strategy has been used. So one anthropometric parameter only have two types of situations: included or excluded. Therefore, one subject has $2^{25} - 1 = 33,554,431$ different possible combinations of anthropometric parameters (excluding the situation where all parameters are outside the combination). Then, we compare subjects in pairs by calculating the difference between their combinations as follow:

$$DI^{(i,j,k)} = \|\sum A_n^{(i,k)} - \sum A_n^{(j,k)}\|, \quad \forall k = 1, 2, ..., 2^{25} - 1, \tag{3.11}$$

where $DI^{(i,j,k)}$ corresponds to the difference between sum of i-th subject and sum of j-th subject in k-th combination.

Next, we calculate the average spectral distortions (SD) of HRTFs between all subject pairs:

$$SD(H^{(i)}, H^{(j)}) = \sqrt{\frac{1}{D}\frac{1}{N}\sum_{d=1}^{D}\sum_{n=1}^{N}(20\log_{10}\frac{\|H^{(i,d)}(n)\|}{\|H^{(j,d)}(n)\|})^2} \tag{3.12}$$

where $H^{(i,d)}$ corresponds to i-th subject's HRTF in d-th direction. N is the number of frequency bins and is equal to 128. D is the number of directions and is equal to 1250. Then we obtain a matrix of average spectral distortions $SD^{S\times S}$, where $S = 35$ is the number of total subjects in considered for the experiments.

After then, we calculate the correlation between possible combinations of anthropometric parameters and the spectral distortion:

$$\rho^{(i,k)} = corr(DI^{(i\times 35,k)}, SD^{i\times 35})) \tag{3.13}$$

where $\rho^{(i,k)}$ corresponds to the Pearson's correlation coefficient of i-th subject in k-th combination, $DI^{(i\times 35,k)}$ corresponds to the difference matrix between sum of i-th subject and sum of other 35 subject in k-th combination, $SD^{S\times S}$, represent the average spectral distortions matrix between i-th subject and other 35 subject.

If a combination gives the biggest value of ρ, we can define this combination as the best anthropometric combination for i-th subject. Finally, we can obtain a total of 35 best anthropometric combinations.

Then, anthropometric feature's weight can be measured by the frequency of occurrence of this anthropometric parameter in all best combinations:

$$W^{(i)} = \frac{t^{(i)}}{S} \tag{3.14}$$

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 and S is the number of best combinations and is equal to 35.

In the weight vector $W = [W^{(1)}, W^{(2)}, ..., W^{(F)}]$, each element corresponds to the weight of a anthropometric paramter, F is the number of anthropometric parameters we used and is equal to 25.

3.4.2 Sparse representation of anthropometry features

The basic assumption in our approach is that, the HRTF data is in the same relation as these anthropometric features. Then, the sparse vector of anthropometric features can be used directly to synthesize this subject's synthesis.

We used sparse representation, to estimate the standard score of the new subject's anthropometric parameters A_s, as a linear superposition of the standard score of the anthropometric parameters of the users in the database [17]:

$$A_t \approx \beta A_s \tag{3.15}$$

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 linear superposition.

Thus, the problem of looking for an optimal sparse vector can be considered as a minimization problem:

$$\beta = \arg\min_{\beta}(\|W(A_t - \beta A_m)\|_2^2 + \lambda\|\beta\|_1), \beta^{(i)} \geq 0, \tag{3.16}$$

where W represents the weights of different anthropometric parameters. As suggested by [30], we added a non-negative constraint on the β as recommended, e.g. $\beta^{(i)} \geq 0$.

The regularization parameter λ of this minimization problem is a non-negative parameter.

3.4.3 Postprocessing for sparse vectors

As the sum of the beta vector which we obtain from the minimization problem as Eq. 3.16 may not be 1. In order to make sure the magnitude of the HRTF stays the same[30], we normalize the values of the Beta vector such that the sum of the beta vector is equal to 1.

$$\beta_m^{(i)} = \frac{\beta^{(i)}}{\sum_{s=1}^{S} \beta^{(s)}} \tag{3.17}$$

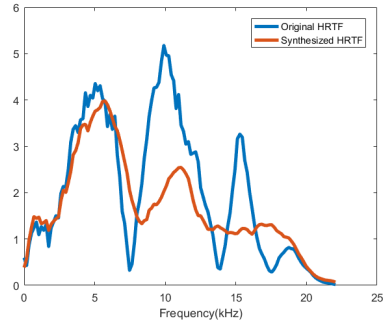
3.4.4 HRTF synthesis

As the assumption in [17] that the HRTFs can be represented using the same sparse representations as the anthropometric 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.

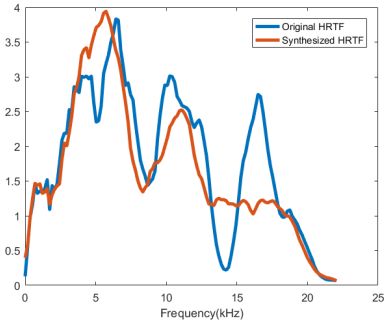
$$\hat{H}_{[dB]} = \sum_{s=1}^{S} \beta_m^{(s)} H_{[dB]}^{(s)} \tag{3.18}$$

However, the new synthesized HRTF $\hat{H}_{[dB]}$ is in dB scale, so we need to transfer the synthesized result into the scalar magnitude. The comparison between new synthesized HRTF and original HRTF are presented in Fig. 3.8

$$\hat{H} = 10^{\frac{\hat{H}_{[dB]}}{20}} \tag{3.19}$$



(a) subject003



(b) subject022

Figure 3.8: The comparison between new synthesized HRTF and original HRTF

3.4.5 Regularization parameter

Authors in [30] suggested, adding the only one parameter λ into the minimization problem can prevent over-fitting[33]. The model of over-fitting are shown in Fig 3.9. So a number of λ need to be tested by using the anthropometric measurements and measured HRTFs in the database and one will be selected as the optimal value of λ [34].

To find the value of λ, we can solve the minimum problem in Eq.3.16 using Least Absolute Shrinkage and Selection Operator (LASSO)[35]. Select the λ which results in the smallest cross-validation error as the optimal one. We used root mean square error as cross validation measure here as in eq 4.2.

In order to match the scale of λ to preprocessed anthropometric parameters and tune the value of λ easily, we normalize λ as [30] suggested:

$$\lambda = \frac{\lambda_0}{1-\lambda_0}\|A_t\|_2^2 \tag{3.20}$$

where A_t corresponds to preprocessed anthropometric parameters of the new subject. In this case by tuning the value of λ_0 from 0 to 1, we can obtain any nonnegative value of λ.

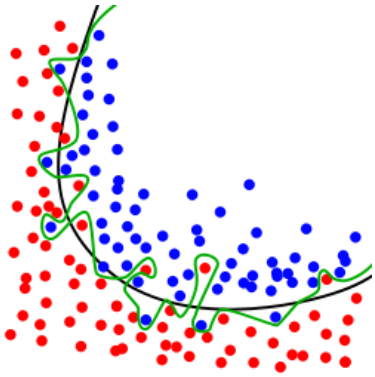


Figure 3.9: An example of the over-fitting model[36]

Chapter 4

Implementation and Results

In this chapter we will analyze the performance of our proposed approach. We used spectral distortion as the evaluation criteria and apply leave one out cross-validation approach [37]to build up the evaluation protocol. We also compared the results of performance with the previous sparse representation methods [17] and other three closest-match based HRTF personalization methods[12, 19, 20].

4.1 Database Selection

In order to compare the performance of our proposed approach with previous sparse representation techniques and other three closest-match based HRTF personalization, we need to select a suitable HRTF database with anthropometric measurements.

After considering many HRTF database containing anthropometry data[6, 8, 17, 25, 38, 39], as presented in table 4.1, we choose CIPIC database as the database in our work.

CIPIC database is a publicly available database of HRIRs that also contains measured HRIRs for 45 different subjects for 1250 different directions. CIPIC database can be obtained online and the number of anthropometric

HRTF databse	Year	subjects	Direction	anthro features
CIPIC[6]	2001	45	1250	27
Nishino et al [38]	2005	86	72	9
Xie et al[25]	2007	57	72	17
TUM LDV[39]	2013	35	2160	8
Microsoft Research[17]	2014	250+	512	52
SYMARE[8]	2014	61	393	3D model

Table 4.1: List of HRTF databases with anthropometric measurements

features is large enough for sparse representation calculation. More importantly, CIPIC also contain the pictures of the pinna, which can be used to find the closes match[12]. However, only 35 subjects have all 27 anthropometric data, so we only use the data of these 35 subjects in this work.

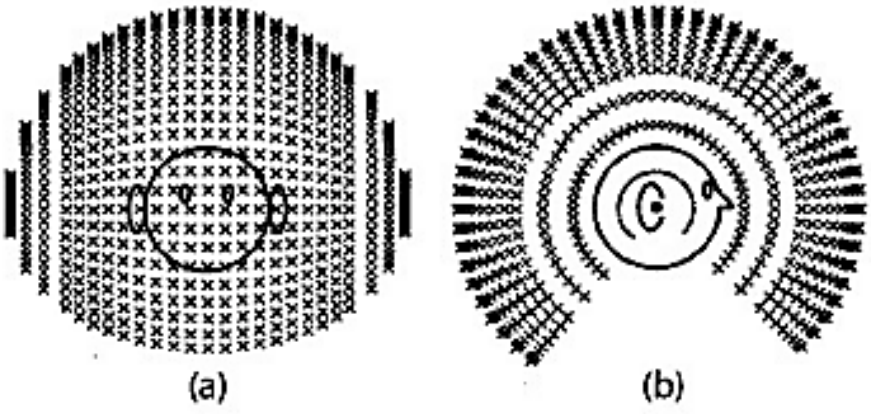


Figure 4.1: Loud speaker positions for the HRIR measurements in cipic database.(a) front view (b) side view [6]

4.2 Evaluation Criteria

4.2.1 Spectral Distortion

To compute 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.[11, 17]. Eq. 4.1 shows the expression to compute the spectral distortion SD.

$$SD^{(d)}(H,\hat{H}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(20\log_{10}\frac{\|H^{(d)}(n)\|}{\|\hat{H}^{(d)}(n)\|})^2} \quad [dB] \tag{4.1}$$

where $H^{(d)}$ represents the original and $\hat{H^{(}}d)$ is the synthesized HRTF in d-th direction. N is the number of frequency bins in considered frequency range, in our research N is equal to 128.

Then we can use the root mean square error (RMSE) to compare the two sets of HRTFs for all 1250 directions:

$$SD(H,\hat{H}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(SD^{(d)}(H,\hat{H}))^2} \quad [dB] \tag{4.2}$$

where D is 1250 in CIPIC database.

4.2.2 "The Best" and "The Worst" baselines in CIPIC

As we want to compare the performance of our approach with three different closest-match methods introduced by [19, 40, 20], we calculate "The Best" and "The Worst" average spectral distortion baselines for these 35 subjects in CIPIC database in all 1250 directions[17].

When a subject successfully selects one other subject which has the minimum average spectral distortion difference to it, this selection is defined as

"The Best". On the contrary, "The Worst" result here represents the choice which results in the largest spectral distortion.

The results are presented in Table 4.3. The result depict that in case of using the closest-matching approach the best and worst results will be bounded by the values of "The Best" and "The Worst" baselines.

4.3 Evaluation Protocol

We set up our evaluation protocol based on leave one out cross-validation approach(LOOCV)[37], as presented in Fig. 4.2. This approach works on a simple idea. Suppose we have a set of n subjects. For every trial we take one of these subjects as our test subjects while the remaining of the n-1 subjects will be regarded as the train subjects.

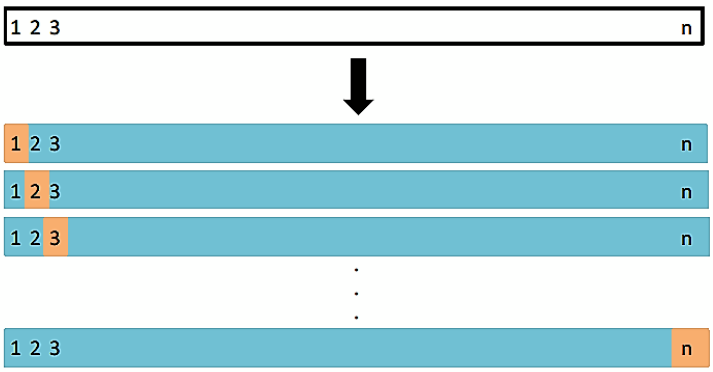


Figure 4.2: The basic schematic display of LOOCV

Among all 45 subjects present in CIPIC database, only 35 of these subjects have all 27 anthropometric measurements. So for our studies we consider only this subset of users. Each of 35 subjects will be taken out one by one as the test subject and the remaining 34 subjects will be regarded as the train subjects at the same time.

Having both the anthropometric and HRIR data in hand, the first thing we did was to calculate the relevance weights of the anthropometric features. For this purpose we used the partial on-off method. For further details on

how this method works, readers are invited to read section 3.4.1. Unlike the previous studies, we did the weight calculation for both ears seperately.

Then these subjects will be selected out one by one as the test subject. We can find a sparse representation of the test subject's anthropometric features as a linear superposition of the anthropometric features of the remaining subjects in the database. After that, the test subject's HRTF can be synthesized by using the β vector and HRTF data in the database.

We will calculate the average spectral distortion between synthesized HRTFs and original HRTFs. The average value of these 35 average spectral distortions can be regarded as the evaluation value of HRTF personalization performance in these 35 subjects from CIPIC database.

The three closest matching based methods are also evaluated using the same steps.

4.4 Results and discussion

The results of our experiments are presented in Table 4.2, Table4.3 and Table4.4. We also present these results in in Fig 4.3, 4.4 and 4.5 for more intuitive comparison.

	Left Ear	Right Ear	Average
Weighted Parameters (17)	5.5235	5.5351	5.5293
Unweighed Parameters (17)	5.6298	5.6359	5.6328
Unweighed Parameters (27)	5.5770	5.5707	5.5738

Table 4.2: Spectral distortion values for sparse representation approach for different setups

Results presented in Table 4.2 and Fig 4.3 show that the average spectral distortion of weighted sparse representation using 17 anthropometric parameters is 5.53dB, which is better than the unweighed sparse representation even when 27 anthropometric parameters are used(5.57dB) and that of unweighed sparse representation using 17 anthropometric parameters (5.63dB).

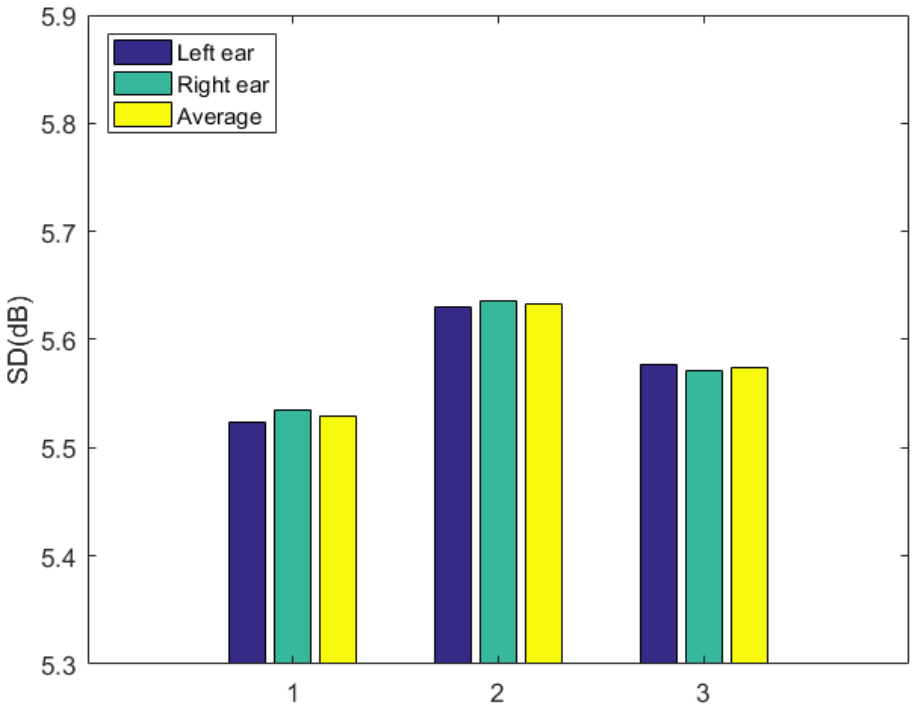


Figure 4.3: Result of spectral distortion of different sparse representation methods. 1 weighted sparse representation of 17 parameters, 2 unweighted sparse representation of 17 parameters, 3 unweighted sparse representation of 27 parameters;

	Left Ear	Right Ear	Average
Weighted Parameters (17)	5.5235	5.5351	5.5293
Pinna reflection[40]	7.3403	7.3403	7.3403
Closest-match using PCA[20]	7.6287	7.1844	7.4065
Weighted closest-match[19]	7.5451	7.2239	7.3845

Table 4.3: Average spectral distortion in all 1250 directions of weighted sparse representation of 17 parameters, closest-match method using pinna reflection, closest-match method using PCA selection, closest-match method using weighted anthropometric parameters in [dB]

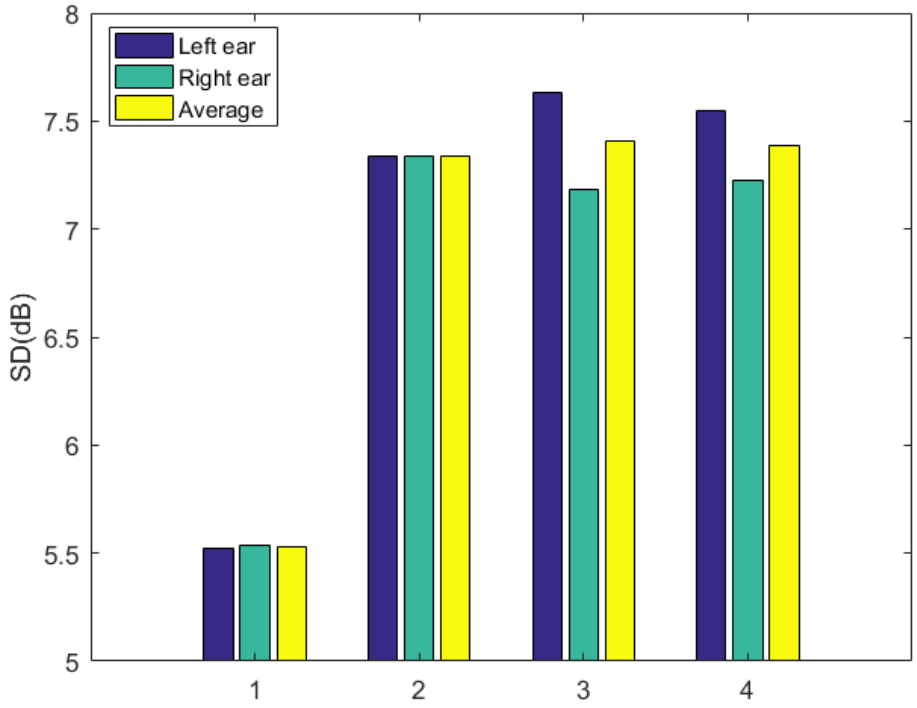


Figure 4.4: Result of spectral distortion of different HRTF personalization methods. 1 weighted sparse representation of 17 parameters, 2 closest-match method using pinna reflection, 3 closest-match method using PCA selection, 4 closest-match method using weighted anthropometric parameters

	Left Ear	Right Ear	Average
Weighted Parameters (17)	5.5235	5.5351	5.5293
"Best" baseline	6.2306	6.0317	6.1311
"Worst" baseline	9.5628	9.0821	9.3324

Table 4.4: Average spectral distortion of weighted sparse representation of 17 parameters and the "Best" and "Worst" baselines in CIPIC database in [dB]

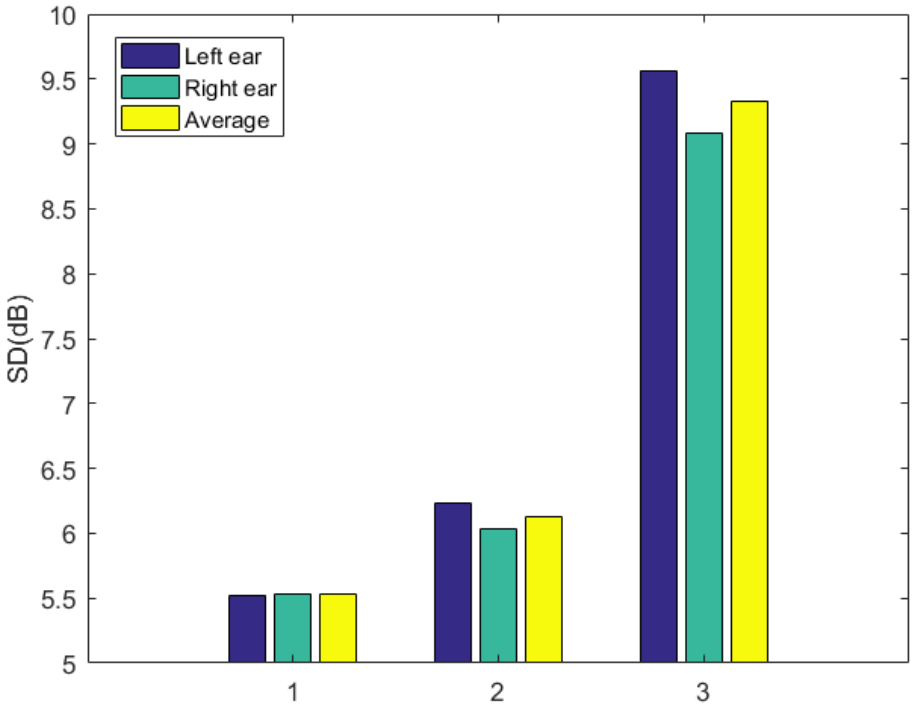


Figure 4.5: Result of spectral distortion of different HRTF personalization methods. 1 weighted sparse representation of 17 parameters, 2 the "Best" baselines in CIPIC database, 3 the "Worst" baselines in CIPIC database

Obviously, even though we use less anthropometric parameters, our approach still provides better results. More importantly, we can not directly obtain all these 27 anthropometric parameters through three scaled pictures, but all these 17 anthropometric parameters we used can be acquired from three scaled pictures of the subject. These results also proved that by considering the relevance of different anthropometric parameters the performance of HRTF personalization can be improved.

Results presented in Table 4.3 and Fig 4.4 show that the average spectral distortion of closest-match method using pinna reflection is 7.34dB, the average spectral distortion of closest-match method using PCA selection is 7.40dB, and the average spectral distortion of closest-match method using weighted anthropometric parameters is 7.38dB. These results indicated that our proposed approach outperforms the three closest-match methods.

Considering "The Best" baseline (6.13dB) and "The Worst" baseline (9.33dB), we can first find out that the result of all these three closest-match based methods are in the range of "The Best" baseline and "The Worst" baseline. It can be proved that the result of "The Best" baseline can represent the ideal result that the closest match can ever be reached.

However, we find that the average spectral distortion of weighted sparse representation using 17 anthropometric parameters is lower than "The Best" baseline. Because using closest match based methods can return only one of the non-individualized HRTFs in the database and does not let the user adjust the HRTF magnitudes, which may nor perform the best. Using weighted sparse representation method can adjust the HRTF magnitudes according to the anthropometric features of the test subject, which can acheive a better result.

Chapter 5

Conclusion and Future work

5.1 Concluding remarks

In this work, we introduced a simple and effective HRTF personalization method based on weighted sparse representation with preprocessing and post-processing methods.

Our work is defined in the field of HRTF personalization base on anthropometric features. We use 17 anthropometric features to personalize the HRTF. All of these features can be measured from subject's three scaled pictures. Using the partial on-off approach we calculated the weights for every anthropometric feature. The weights reflect the relevance of every feature in the process of personalization. We investigated that using some simple pre and post processing techniques can result in a better performance of the personalization method. We selected spectral distortion as the experimental evaluation criteria and applied leave one person out cross-validation approach to do the experiment.

Finally, we compared our proposed approach with previous sparse representation using 27 anthropometric parameters. Even though we use less anthropometric parameters, our approach still provides better results.

We also compared proposed approach with other closest-match based per-

sonalization methods. The result of experiment indicated that our approach outperforms the three closest-match methods.

5.2 Future work

Although the subjective tests shows that our approach outperforms almost all previous personalization methods but it still needs to be verified using perceptual localization testing.

The future work includes to reproduce the meaning 3D audio content using the HRTFs from our personalization method and run some psycho perception tests. We also aim to set up a simple mobile phone app for listeners which can be used to measure the anthropometric feature and provide a personalized HRTF.

Bibliography

[1] L. Savioja, A. Ando, R. Duraiswami, E. A. P. Habets, and S. Spors. Introduction to the issue on spatial audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):767–769, 2015.

[2] J. J. He. Literature review on spatial audio. In *Spatial Audio Reproduction with Primary Ambient Extraction*, pages 7–37. Springer, 2017.

[3] D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, 2001.

[4] 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*, 94(1):111–123, 1993.

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

[6] 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*, pages 99–102. IEEE, 2001.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[8] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghariand A. Van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe. Creating the sydney york mor-

phological and acoustic recordings of ears database. *IEEE Transactions on Multimedia*, 16(1):37–46, 2014.

[9] 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*, pages 157–160. IEEE, 2003.

[10] 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*, volume 3, pages III–57. IEEE, 2003.

[11] 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.

[12] 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)*, pages 231–236, 2015.

[13] 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)*, pages 101–105. IEEE, 2016.

[14] 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.

[15] 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.

[16] L. Li and Q. Huang. HRTF personalization modeling based on RBF neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3707–3710. IEEE, 2013.

[17] 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)*, pages 4468–4472. IEEE, 2014.

[18] E. A. Macpherson and J. C. 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.

[19] E. A. T. Gallegos, F. O. Bustamante, and F. A. Cosío. Personalization of head-related transfer functions HRTF based on automatic photoanthropometry and inference from a database. *Applied Acoustics*, 97:84–95, 2015.

[20] 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*, 329(19):4093–4106, 2010.

[21] V. R. Algazi and R. O. Duda. Headphone-based spatial sound. *IEEE Signal Processing Magazine*, 28(1):33–42, 2011.

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

[23] D. R. Begault and L. J. Trejo. 3-D sound for virtual reality and multimedia. 2000.

[24] J. Tobias. *Foundations of modern auditory theory*. Elsevier, 2012.

[25] B. S. Xie, X. L. Zhong, D. Rao, and Z. Q. Liang. Head-related transfer function database and its analyses. *Science in China Series G: Physics Mechanics and Astronomy*, 50(3):267–280, 2007.

[26] E. A. G. Shaw. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6):1848–1861, 1974.

[27] H. Gamper. Head-related transfer function interpolation in azimuth, elevation, and distance. *The Journal of the Acoustical Society of America*, 134(6):EL547–EL553, 2013.

[28] O. Warusfel. Listen HRTF database. *online, IRCAM and AK, Available: http://recherche. ircam. fr/equipes/salles/listen/index. html*, 2003.

[29] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[30] 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)*, pages 639–643. IEEE, 2015.

[31] J. Felsand S. Fingerhuth. Anthropometric data acquisition using photogrammetric techniques to obtain acoustic head-related transfer functions of children. In *Proceedings of the CTU Conference*, 2004.

[32] Y. Dodge. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand, 2006.

[33] I. V. Tetko, D. J. Livingstone, and A. I. Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

[34] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[36] D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[37] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[38] S. Hosoe, T. Nishino, K. Itou, and K. Takeda. Measurement of head-related transfer functions in the proximal region. In *Forum Acusticum*, pages 2539–2542, 2005.

[39] M. Rothbucher, P. Paukner, M. Stimpfl, and K. Diepold. The TUM-LDV HRTF database. Technical report, Technical Report, Technische Universität München, 2013.

[40] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 4463–4467. IEEE, 2014.

[41] D. Schönstein and B. Katz. Sélection de HRTF dans une base de données en utilisant des paramètres morphologiques pour la synthèse binaurale. In *10ème Congrès Français d'Acoustique*, 2010.

Appendix A

The result of 25 anthropometric features' weight factor in CIPIC database

The results of 25 anthropometric features' weight factors in CIPIC database are presented in Table. A.1. Each anthropometric feature has two different weight factors. All these weight factors are in the range of 0 to 1. The higher the weight factor is, the more relevant this anthropometric feature is.

Var	Measurement	Left Ear	Right Ear
x_1	head width	0.5714	0.5429
x_2	head height	0.5143	0.4857
x_3	head depth	0.5714	0.5429
x_4	pinna offset down	0.4286	0.3429
x_5	pinna offset back	0.1429	0.1143
x_6	neck width	0.2000	0.2857
x_7	neck height	0.5429	0.3714
x_8	neck depth	0.4286	0.6286
x_9	torso top width	0.2286	0.1714
x_{10}	torso top height	0.4000	0.4857
x_{11}	torso top depth	0.3143	0.0857
x_{12}	shoulder width	0.5429	0.4286
x_{13}	head offset forword	0.6857	0.6571
x_{16}	head circumference	0.2857	0.1143
x_{17}	shoulder circumference	0.4286	0.4857
d_1	cavum concha height	0.3143	0.2571
d_2	cymba concha height	0.1429	0.1714
d_3	cavum concha width	0.2000	0.2857
d_4	fossa height	0.5714	0.6286
d_5	pinna height	0.1429	0.0857
d_6	pinna width	0.6286	0.4000
d_7	intertragal incisure width	0.4286	0.3143
d_8	cavum concha depth	0.1429	0.4571
θ_1	pinna rotation angle	0.5143	0.5429
θ_2	pinna flare angle	0.4000	0.8000

Table A.1: The result of 25 anthropometric features' weight factor in CIPIC database

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

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

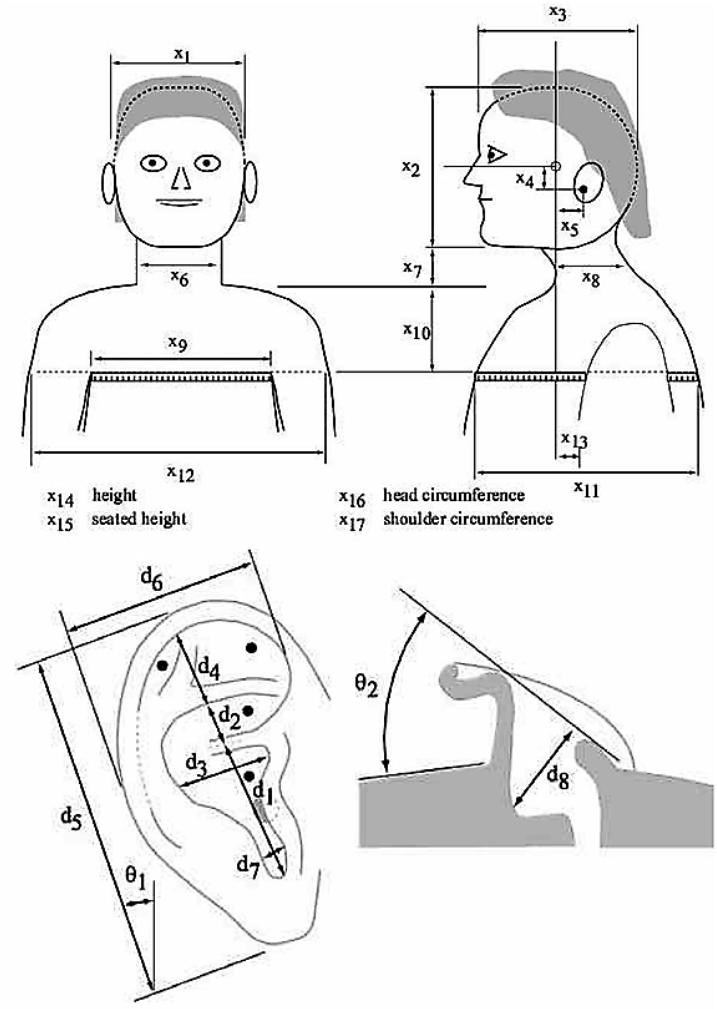


Figure A.1: Head, torso and pinna measurements in CIPIC database [41]

Appendix B

The result of 5 subjects' sparse representation in CIPIC database

The result of 5 subjects'(subject003, subject010, subject018, subject020 and subject027) sparse representation in CIPIC database are present in Table. B.1 and Table. B.2. Each column represent a sparse vector of a subject. These sparse vectors are the solutions of minimization problems in Eq. 3.16. The subject's sparse vector can be used to synthesize the HRTF of this subject with the HRTF data in the CIPIC database.

Subject003	**Subject010**	**Subject018**	**Subject020**	**Subject027**
0.0203	0.0201	0.0327	0.0345	0.0247
0.0319	0.0360	0.0344	0.0255	0.0293
0.0317	0.0289	0.0237	0.0188	0.0306
0.0232	0.0301	0.0309	0.0331	0.0297
0.0301	0.0260	0.0318	0.0441	0.0344
0.0316	0.0312	0.0302	0.0266	0.0257
0.0219	0.0274	0.0409	0.0319	0.0278
0.0389	0.0408	0.0299	0.0138	0.0275
0.0299	0.0234	0.0298	0.0217	0.0297
0.0275	0.0381	0.0205	0.0393	0.0284
0.0344	0.0234	0.0275	0.0251	0.0323
0.0308	0.0323	0.0295	0.0265	0.0279
0.0266	0.0316	0.0257	0.0268	0.0267
0.0334	0.0322	0.0360	0.0364	0.0254
0.0234	0.0343	0.0301	0.0464	0.0203
0.0307	0.0311	0.0271	0.0284	0.0374
0.0320	0.0363	0.0291	0.0309	0.0280
0.0363	0.0249	0.0301	0.0203	0.0331
0.0216	0.0333	0.0289	0.0345	0.0366
0.0297	0.0267	0.0283	0.0242	0.0256
0.0214	0.0292	0.0233	0.0136	0.0277
0.0319	0.0340	0.0290	0.0449	0.0339
0.0341	0.0224	0.0239	0.0258	0.0230
0.0223	0.0289	0.0279	0.0323	0.0365
0.0239	0.0259	0.0287	0.0225	0.0268
0.0316	0.0261	0.0281	0.0167	0.0264
0.0386	0.0267	0.0293	0.0405	0.0268
0.0429	0.0250	0.0341	0.0359	0.0307
0.0263	0.0369	0.0352	0.0184	0.0366
0.0269	0.0264	0.0340	0.0331	0.0323
0.0295	0.0228	0.0276	0.0149	0.0288
0.0293	0.0360	0.0315	0.0338	0.0229
0.0254	0.0263	0.0220	0.0454	0.0300
0.0285	0.0238	0.0267	0.0317	0.0348

Table B.1: The result of 5 subjects' sparse representation of left ear

Subject003	**Subject010**	**Subject018**	**Subject020**	**Subject027**
0.0166	0.0156	0.0343	0.0292	0.0243
0.0358	0.0354	0.0343	0.0250	0.0265
0.0291	0.0286	0.0249	0.0206	0.0311
0.0224	0.0264	0.0311	0.0197	0.0243
0.0286	0.0295	0.0302	0.0540	0.0387
0.0292	0.0367	0.0272	0.0188	0.0252
0.0240	0.0238	0.0354	0.0222	0.0297
0.0386	0.0357	0.0247	0.0210	0.0272
0.0303	0.0184	0.0349	0.0224	0.0311
0.0291	0.0321	0.0193	0.0346	0.0322
0.0292	0.0262	0.0294	0.0321	0.0273
0.0352	0.0280	0.0261	0.0186	0.0249
0.0288	0.0266	0.0288	0.0371	0.0302
0.0286	0.0365	0.0370	0.0181	0.0287
0.0217	0.0338	0.0266	0.0549	0.0215
0.0371	0.0276	0.0287	0.0361	0.0347
0.0332	0.0399	0.0294	0.0240	0.0268
0.0283	0.0275	0.0302	0.0384	0.0358
0.0231	0.0259	0.0304	0.0315	0.0341
0.0316	0.0357	0.0286	0.0277	0.0291
0.0195	0.0338	0.0210	0.0172	0.0325
0.0342	0.0322	0.0336	0.0333	0.0242
0.0377	0.0187	0.0216	0.0193	0.0291
0.0202	0.0303	0.0281	0.0385	0.0288
0.0291	0.0239	0.0307	0.0268	0.0307
0.0250	0.0334	0.0258	0.0305	0.0330
0.0413	0.0326	0.0310	0.0516	0.0279
0.0359	0.0267	0.0334	0.0253	0.0306
0.0260	0.0404	0.0378	0.0296	0.0369
0.0322	0.0227	0.0341	0.0318	0.0285
0.0262	0.0290	0.0268	0.0176	0.0332
0.0299	0.0377	0.0323	0.0301	0.0236
0.0294	0.0228	0.0228	0.0351	0.0279
0.0312	0.0242	0.0277	0.0255	0.0279

Table B.2: The result of 5 subjects' sparse representation of right ear