Estimation of Hearing Aid Head Related Transfer Functions using Anthropometric Features.

A thesis presented in partial fulfilment of the requirements for the degree of Bachelor of Science

Simon Nunayon

S.B. Degree Candidate in Electrical Engineering

Faculty Advisor: Demba Ba

Harvard University School of Engineering and Applied Science

Cambridge MA

Date April 1st 2024

Contents

Acknowledgments

I would like to express my appreciation to Professor Demba Ba for his invaluable guidance through this project. I must also thank Leo Gomez for expediting the distillation of my highly varied ideas. Beyond this, I am grateful to the ES 100 staff for developing the framework that has enabled me to combine my problem-solving and research interests with my passion for audio and human sound perception. I must also thank Professor Virginia Best from Boston University for giving me access to their KEMAR manikin at the 11th hour.

I am immensely grateful to my classmates and colleagues, who have consistently provided moral support, exemplary inspiration, and good humour.

Ultimately, this endeavour would not have been possible without the love and good cheer of my partner, roommates, friends and family. (Except Sofia, you know why). Lastly, I must thank my teammates and coaches for giving me a place where I am not an engineer with a looming thesis deadline.

Abstract

Virtual reality is emerging as an ecologically valid environment for auditory rehabilitation and diagnostics. Creating an auditory environment that accurately simulates a user's experience requires accurate knowledge of their Head-Related Transfer Functions (HRTFs). However, these vary due to differences in physical features and can lead to the incorrect simulation of sound if the HRTFs used are too dissimilar to a listener's. This issue is exacerbated by the use of a hearing device as sound is transmitted to the eardrum from a microphone located elsewhere on the ear, changing the path of sound and consequently the HRTFs. This project presents a method of identifying a suitable set of HRTFs from a dataset, performing better than chance and leads to a localisation error 1.2°greater than a subject using their HRTFs.

List of Figures

List of Tables

1 Define

1.1 Introduction

Virtual Reality aural rehabilitation is emerging as a potentially effective treatment for hearing loss as it more accurately simulates real-life listening environments. However, to accurately simulate 3D sound, Head Related Transfer Functions (HRTFs) must be applied [42] to sound sources. These functions are position-dependent and influenced by anthropometric features such as the head, torso and ears (Fig. 1). These features influence HRTFs as they determine how sound signals are incident upon the eardrum after being reflected or absorbed by the body. As a result, a difference in physical features leads to a difference in the sound incident upon the eardrum. The brain relies on an individual's set of functions to decipher the location of a sound.

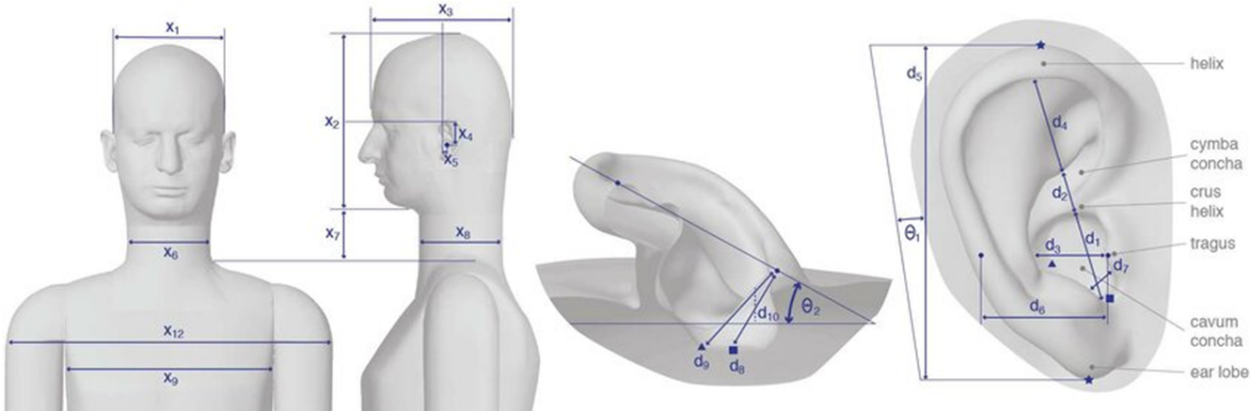


Figure 1: Anthropometric measurements of the head, torso, and ears [18]

However, non-individualised (generic) HRTFs are often used to spatialise sound for virtual reality or spatial audio, causing some people to incorrectly localise sound if the applied HRTFs are too far removed from their own. This issue is more prevalent for users of hearing aids and hearing aids that use microphones placed at the entrance to the ear canal (ITC) (Fig. 2). ITC hearing aids are specifically chosen for this project as their relatively inconspicuous nature lends them well to the approaches used.

Figure 2: In the Canal Hearing Aids[19]

Problem Statement

Virtual Reality presents controllable and realistic listening environments for diagnosis and rehabilitation. This requires suitable HRTFs. However, the use of an ITC sound processor leads to a different set of HRTFs that must be estimated without an anechoic chamber.

Need for a Solution

It has been observed there is a difference between HRTFs measured at the eardrum and HRTFs measured at the ear canal entrance. This figure 3 shows the HRTF (top - eardrum) for two subjects and the transfer functions that would map sound at the entrance to the ear canal to the eardrum (bottom - ECEBl). If the two HRTFs were the same, the bottom function would be a horizontal line.

Critical Project Goals

The critical goal of this project is an algorithm that presents a superior set of HRTFs. These will either be selected from a dataset or estimated based on the user's anthropometric features.

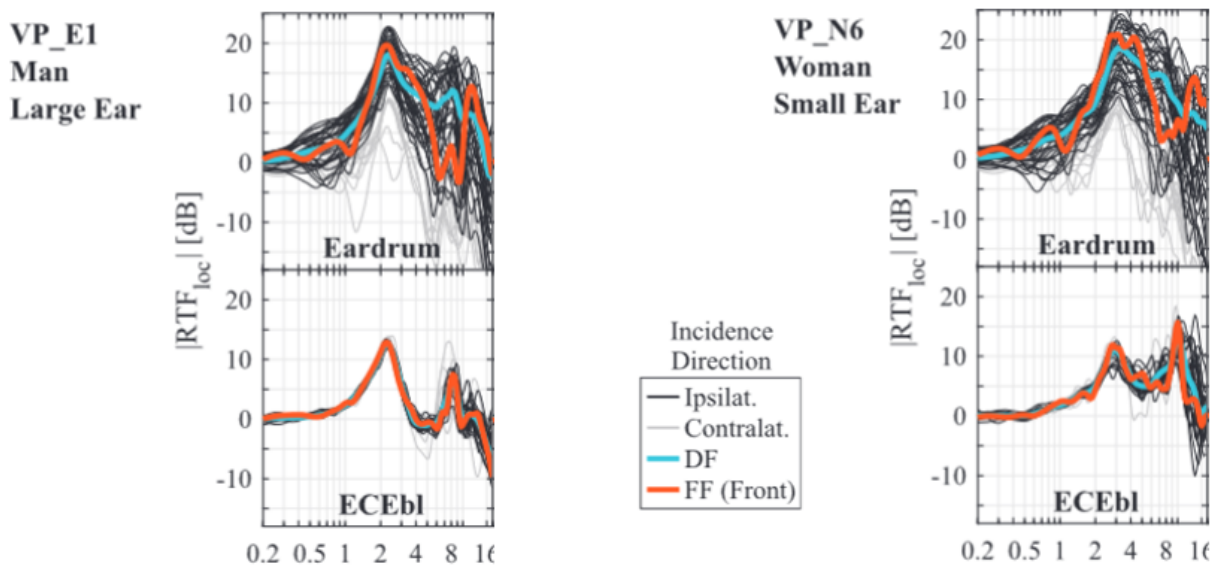


Figure 3: HRTFs for two subjects from Oldenburg study at the eardrum and the relative function to the blocked ear canal (DF - Diffuse field, FF - Free field) [12]

1.2 Background and Broader context

Hearing loss is a global problem that disproportionately affects people in low- and middle-income countries and approximately 37.5 million American adults[28] The World Health Organization predicts 2.5 billion people will suffer from some form of hearing loss by 2050 [11]. In America alone, 28.8 million adults could benefit from hearing aids. However, current hearing aids do not restore normal levels of hearing which contribute to feelings of loneliness and isolation.

Hearing aids work by amplifying a sound incident on the microphones and applying a multi-band compression algorithm tuned to the listener. After processing, this signal is played through a speaker located near the eardrum. Several methods are employed to generate the necessary parameters for the multi-band compression algorithm. A method most effective for patients with conductive hearing loss (where sounds cannot pass from the outer ear to the inner ear) is Real Ear Measurements (REM), where a microphone is placed near the eardrum and measures the strength of frequency bands incident upon the ear drum. This is used to verify the effectiveness of hearing aids and inform further adjustments.

However, sensorineural hearing loss (damage to hair cells and/or the auditory nerve) means even a perfect recreation of sound does not restore full hearing ability.

Furthermore, while all sounds can be represented as pure tones of different frequencies, the brain does not entirely interpret sound incidents on the eardrum in this manner. Thus, it is not surprising that on the Abbreviated Profile of Hearing Aid Benefit (APHAB - a measure for the effectiveness of a hearing aid) most people report a 5 % to 40 % benefit after six weeks of using hearing aids [10].

Many hearing aid fittings involve a speech-in-noise test [15], where patients are asked to repeat sentences they hear while noise is present. Based on their responses, further adjustments will be made to their hearing aids. However, some considerations must be made as it is currently impossible to exactly recreate a patient's real-world experience within an audiologist's office.

The first major issue is realistically simulating speakers in noisy or chaotic environments such as restaurants or train stations. Some labs possess facilities that can simulate entire environments, but these require multiple speaker arrays and a suitably large sound-isolated room. This makes them inaccessible to the majority of audiologists and their patients.

Another issue relates to the voices that patients regularly encounter. Real-world speech is accompanied by several context clues that are not provided during testing. Audiologists account for this as normal-hearing people do not score perfectly on tests. Additionally, if the voice used to test a patient's hearing is an unfamiliar accent, they may score poorly. After all, there is a measurable difference between the speech perception of a familiar voice and an unfamiliar voice. This further limits the audiologist's ability to understand the true extent of a patient's hearing loss.

Rehabilitation exercises are an opportunity to address the shortcomings of a hearing aid fitting. Traditional exercises involve another speaker saying two similar words while the patient has to differentiate between them. Rehabilitation software removes the need for another speaker and has become more accessible due to the range of mobile applications

Figure 4: Speaker array within an anechoic chamber [40]

available. These exercises must be done for 45 minutes to an hour daily for positive outcomes.

Despite this, there is limited evidence to suggest rehabilitation exercises improve the perception of speech outside the context of the exercises. This seems counter-intuitive as other forms of rehabilitation involve the consistent practice of a motion, but recent research suggests methods such as neurofeedback [24] or virtual-reality rehabilitation [8] have benefits in real-world conditions. This project would enable the delivery of ecologically valid environments to perform rehabilitation,

HRTFs

HRTFs are typically measured through the use of an anechoic chamber and a multiple-speaker array centred around a listener (Figure 4). From each position, white noise is played and the impulse response at the eardrum is measured with microphones. This is known as a Head-Related Impulse Response (HRIR). The Fourier transform of this response gives the HRTF at that position. When repeated over several positions, the full set of spatially dependent HRTFs can be generated.

HRTF enable sound to be spatialised in 3 dimensions, which is necessary for accurate VR aural rehabilitation. This is achieved by the convolution of a sound source with the HRTF for each ear at that position. Each location relative to the listener has a unique function, locations are characterised in polar coordinates of azimuth, elevation, and distance (Fig. 5).

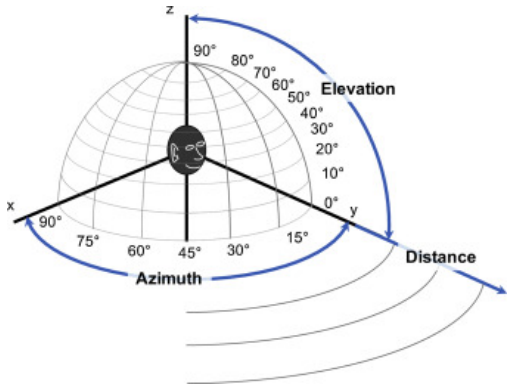


Figure 5: Polar coordinates used to locate sounds [29]

The brain typically relies on interaural time differences (ITDs) and interaural level differences (ILDs) to distinguish the location of a sound source in front of a listener and in the same plane as a listener. These methods fail when a sound source is behind or in a different plane to the listener but ITDs and ILDs are the same. This region is known as the cone of confusion.

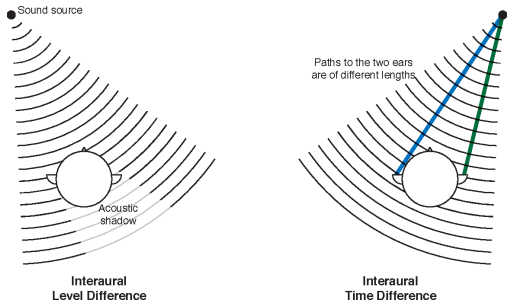


Figure 6: ITD and ILD [6]

In normal hearing listeners, the cone of confusion (COC) (Fig. 7) is easily decoded by spectral analysis. As shown in the figure, the attenuation of a signal due to reflection and absorption by the body differs depending on the location of the source. This spectral analysis also demonstrates peaks in HRTFs are not consistent between directions, which may lead to difficulties in directly estimating HRTFs.

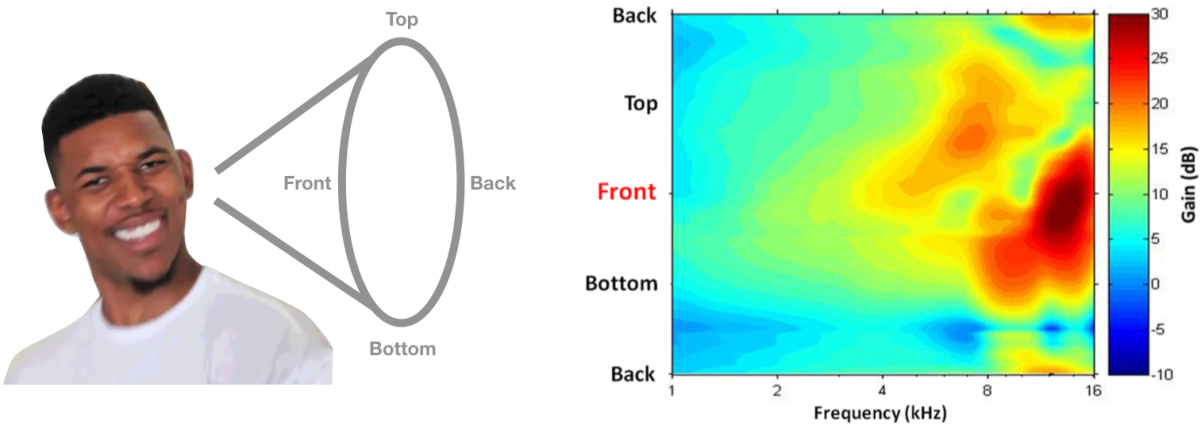


Figure 7: Decoding the cone of confusion [36]

Stakeholders

There are three main stakeholders for this project. First, the users of ITC hearing aids are the most significant stakeholder as success in this project may lead to successful rehabilitation through the use of multiple speaker situations. The second most significant stakeholders are researchers and clinicians designing virtual reality rehabilitation protocols. Lastly, audiologists are a significant stakeholder as rehabilitation protocols proven successful are more likely to be recommended to patients, improving outcomes.

This project is uniquely positioned to address their needs as it combines established methods for individualising HRTFs with an audiologist's proximity to a patient to maximise success. Established methods rely on information reported by patients. This often takes the form of anthropometric data (sometimes collected via pictures) An audiologist augments these processes as they can be trained in an accurate measurement protocol for anthropometric features.

Further Impact

This project's most direct impact is improving the validity of virtual reality rehabilitation research. While this may lead to improved patient outcomes, this project may directly serve to improve the performance of hearing aid users in localising sound around them. Some

hearing aid manufacturers account for the differences caused by microphone locations and apply some form of a correction function.

Referring back to Figure 3, it must also be noted the original direction of a sound has a limited impact on the average relative transfer function between the eardrum and the entrance to the blocked ear canal. This can be compared to other microphone locations in Figure 23 (Appendix). However, these are also generic functions and are not necessarily suitable for all users.

The process outlined in this project can be used to estimate a subject's eardrum HRTFs and their ITC HRTFs. This would present an individualised correction function to maximise hearing outcomes.

1.3 Existing Solutions and Previous Work

1.3.1 Existing Solutions

There exist multiple attempts to individualise HRTFs, but none are designed specifically for hearing aid users.

Non-individualised HRTFs

Generic HRTFs are often based on pre-recorded datasets of people or a manikin such as the KEMAR head and torso simulator (Figure 8). Spatial sound can also be constructed by using a binaural recording from a head simulator such as a Neumann KU 100. These simulators are anthropometrically relevant as they are designed with the average anthropometric features in mind, such as those defined in [35].

The use of incorrect HRTFs can lead to front/back confusion and a greater error in localising sounds [20]. However, people can adapt over time to a new set of HRTFs and retain this adaptation even when they return to their usual HRTFs. However, providing a generic HRTF with headphones for an ITC hearing aid user would be unsuitable for virtual

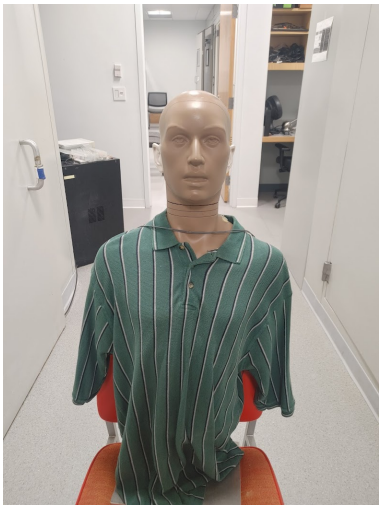


Figure 8: A KEMAR maninkin

reality rehabilitation as sound would not be presented as it is experienced in real life by the user.

HRTF datasets

HRTF datasets attempt to bridge this gap by matching users to a pre-recorded set of HRTFs. They achieve this through a variety of ways, from measuring anthropometric features to asking users to complete a sound localisation task to assess the effectiveness of a set of HRTFs.

These HRTFs are measured by placing microphones within a subject's ears and measuring the sound incident upon them when a sound is played from a speaker at a pre-determined distance, azimuth and elevation. This is done in regular steps of 5° to 15° to create an accurate picture of how the subject experiences sound.

There are numerous datasets available of standard HRTFs [23], which presents a solution for normal hearing users who wish to experience spatialised audio. These datasets can be leveraged by a range of approaches to deliver a suitable set of HRTFs to a listener. However, this breadth of data is not available for users of hearing aids making it an unsuitable approach.

HRTFs from behavioural tests

Another method used to estimate a user's HRTFs is a localisation assessment [9]. Users will be presented with a range of HRTFs and asked to localise sounds as well as judge the audio quality. For a set of HRTFs that matches their own, localisation error will be minimised and audio quality will usually be rated highly.

The dearth of ITC HRTF datasets makes this approach difficult to consider, and relying on human participants would limit the project's rate of iteration and improvement.

HRTFs from photos

Assessing HRTFs from photos is an approach commonly used in the consumer audio industry to spatialise sound. The images are often fed into a neural network and used to modify known HRTFs to estimate a user's HRTFs. The key engineering challenge is estimating the 3D measurements of the ear from a 2D photo, usually with no physical measure to standardise. In this scenario, distance from the ear can be approximated, but physical measurements of the head, torso and ear are the ground truth.

While this approach could be applied to estimating ITC HRTFs, this would incorporate a significant computer vision component into the project which would not fit within the temporal scope of the project.

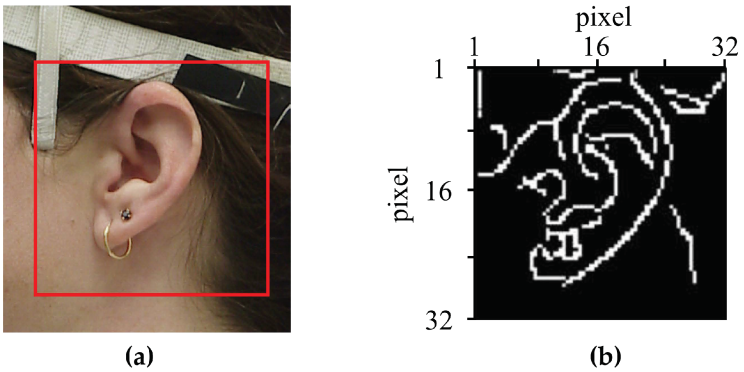


Figure 9: Taking ear measurements from a photo [25]

Direct estimation

An emerging state-of-the-art method is the direct estimation of HRTFs from anthropometric features [38] [39]. These features are either collected by measuring anthropometric features from listeners or measuring from a 3D scan or photos of a listener. However, both of these are unsuitable in their base form.

Firstly, the overarching issue of a lack of ITC HRTF training data. Secondly, it is not feasible to measure the number of physical features needed for these methods. Thirdly, requesting audiologists to invest in a 3D scanner, which can cost on the order of $1000 could lead to a significant financial barrier that would prevent hearing-impaired people from receiving effective rehabilitation.

HRTF modification

A potential method to modify HRTFs is by scaling entire measured functions based on morphological factors of the ear [26]. It is theorized that differences in the spectral characteristics of Head-Related Transfer Functions (HRTFs) among individuals are associated with the dimensions of their ear cavities, influencing the resonances within the ear. A difference in these sizes of the ear cavity, whether through enlargement or reduction, impacts these resonances, causing shifts in both the frequencies of notches and peaks within the HRTF. Specifically, enlarging the ear cavity leads to shifts towards higher frequencies, while reducing its size results in shifts towards lower frequencies. This modification of the ear's dimensions directly affects the positioning of notches and peaks within the HRTF. This approach could be considered without human subjects, but again, there are limited original ITC HRTFs to start with.

Current shortcomings and needs to meet

There are up to 55 morphological factors [21] that are not always measured by datasets that may influence sound incident at the entrance to the ear canal. A measurement in this

position will be influenced by different factors, so it is not sufficient to rely on parameters defined by another study as being most associated with a difference in HRTFs.

1.4 Design Independent Technical Specs

1.4.1 Specification

Requirement	Specification	Value
Fast	Run time	≤ 30 minutes
Easily Measurable	Number of head and torso measurements	≤ 8
	Number of ear measurements	≤ 6
Accurate	Azimuth gradation	$\geq 10^\circ$
	Elevation gradation	$\geq 15^\circ$
	RMS localisation error at 0° elevation	$\leq 3^\circ$
	RMS localisation error at 0° azimuth	$\leq 42^\circ$

Table 1: Design Independent Technical Specifications

1.4.2 Justification and Measurement

Run time

As measurements will be taken within an audiologist appointment, the HRTFs should be available to the patient by the end of their appointment. These appointments can vary in length between 30 minutes and 1 hour. This limits the audiologist's administrative backlog formed by consecutive appointments. This will limit the negative social and economic impact on the audiologist as it allows them to see patients at their usual rate. There is a slight environmental component as well - a short run time within the temporal bounds of an appointment means the audiologist does not require computing power outside their usual hours.

This will be measured by the average time taken to compute the HRTFs after a set of measurements are provided.

Number of head and torso measurements

The dimensions believed to be most important for ear canal HRTFs are the head's dimensions (depth, height, width etc), and shoulder width [30]. However, the audiologist has limited time to measure these within an appointment. Assuming 1 minute to measure a single feature, this would require 8 minutes to measure all head and torso measurements. This number is also greater than that seen in literature [30] [32] to maximise the accuracy of the project within the time constraints of an appointment.

This will be measured by the number of head and torso features identified that can be easily measured.

Number of ear measurements needed

The dimensions of the ear are harder to accurately measure than the head and torso, so the fewer needed, the better. The equipment needed to measure anthropometric features will also be considered - audiologists may lack the callipers or protractors to accurately measure ear features. Assuming 1 minute to measure a single feature, this would require 6 minutes to measure all head and torso measurements. This number is also greater than that seen in literature [30] [32] to maximise the accuracy of the project within the time constraints of an appointment.

Being able to easily measure these features and the head and torso features will maximise the global impact of this project as it reduces the need for expensive specialised tools.

This will be measured by the number of ear features identified that can be easily measured.

Azimuth gradation and Elevation gradation

To make the output of this method comparable to HRTFs in existing datasets, the HRTFs presented should minimise the need for interpolation. Interpolation is used by audio engines if a sound source is located in a position where the HRTF is not defined. As this point lies

between two positions where the HRTF is defined, the HRTF at that point can be estimated, but this may lead to spectral distortion [14]. As normal HRTFs are already unsuitable for hearing aid users, interpolation methods based on traditional HRTFs may lead to incorrect interpolation.

This will be assessed by taking the average azimuth gradation across all elevations, and the average elevation gradation across all azimuths.

RMS error at 0° elevation and RMS error at 0° azimuth

A metric used to quantify HRTF fit is the RMS localisation error on the horizontal plane (0° elevation) and medial plane (0° azimuth). This metric comes from localisation tests. In these tests, a sound is played from a direction unknown to the listener (either through a speaker array or HRTFs are used to simulate the position of a sound) and the listener indicates what angle they believe the sound is coming from [26].

The error at each position is calculated by subtracting the real location of the sound from the perceived location by the listener. This can be used to test the quality of a set of HRTFs across all positions: if the error is minimal (the listener identifies with high accuracy where the sound is coming from), then those HRTFs can be used to spatialise sound for that listener.

The chosen thresholds come from literature. 3° is the expected RMS localisation error for a listener with their HRTFs in the horizontal plane (0° elevation) [27]. However, 42° is the mean RMS localisation error for a listener using another person's HRTFs [26]. For a listener with their HRTFs, there is typically a greater RMS error in the vertical localisation (23° to 34°). It will be demonstrated later in this project that the distribution of vertical localisation errors with any other set of HRTFs is outside of this range. As a result, delivering a set of HRTFs that perform better than the mean from the Middlebrooks experiment [26] may prove this project a success.

This specification will be assessed by comparing the localisation error in both planes

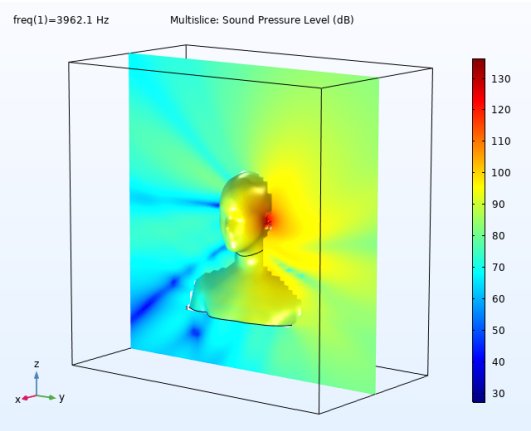
using statistical models based on psychoacoustic localisation experiments.

2 Design

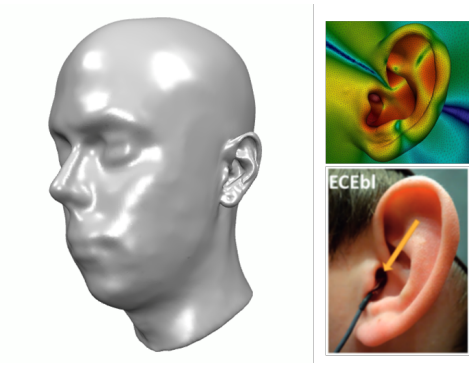
2.1 Design Justification

After reviewing the methods currently used to select or estimate HRTFs, this project will use two approaches to deliver two sets of HRTFs suitable for a listener. These two approaches will be linked by their use of the same reduced number of anthropometric features. However, an issue raised in the earlier evaluation of approaches was the lack of ITC HRTF data to train on.

This issue can be solved by using a technique made possible by modern computing power - simulating HRTFs through the boundary element method (BEM) with a 3D scan of a subject. In traditional HRTF measurement, a loudspeaker serves as an emitter and its impulse response is recorded at each eardrum. BEM reverse this via the reciprocity principle, where sound is emitted from the eardrum, and the response is measured at a location away from the listener.



(a) BEM at 3962Hz [7]



(b) Head mesh, ear mesh and ECEbl configuration

Figure 10: BEM modelling of HRTFs

A limitation of this method lies in the quality of the 3D scan: the laser beam used to construct the 3D mesh cannot capture the ear canal. As a result, the entrance to the ear canal is often smoothed over, and HRTFs are calculated from this point. Due to the lack of ITC HRTFs with any anthropometric data, this approach serves as the best source of

training data for this project as simulation occurs from the entrance of a blocked ear canal ie. where an ITC hearing aid microphone would be.

A preliminary set of experiments also suggested this could be suitable: some known anthropometric features of a KEMAR manikin were compared with all subjects in the HUTUBS dataset [34].

The subject with the least sum of differences in their features was deemed the physically closest, and this subject's simulated HRTFs were compared against the manikin's ear canal entrance blocked (ECEbl) HRTFs (Appendix: Figures 24, 25, 26). While this is a low-fidelity way of comparing HRTFs, (especially as they were measured at different distances), the agreement shown in their overall shapes affirms the use of simulated HRTFs. Ultimately, this provides a large quantity of data that would be difficult to match through self-conducted HRTF experiments.

The considerations above led to the following requirements for a dataset used for this project.

2.1.1 Design Dependent Technical Specs

Requirement	Specification	Value
Representative Datasets	Number of subjects (with anthropometric data)	≥ 50
	Number of ear dimensions provided	≥ 12
	Number of head dimensions provided	≥ 4
	Number of torso dimensions provided	≥ 8
Allows simulation	3D meshes	True/False

Table 2: Design Dependent Technical Specifications

Justification and Measurement

Number of subjects

HRTF datasets with anthropometric data can range between having 10 to 96 subjects

[5]. As a result, it would be preferable to maximise the number of subjects who can be used as training data.

Number of ear dimensions provided

The maximum number of ear dimensions available is 17 per ear, as defined in the ARI dataset [22]. While some of these dimensions will show limited impact on measured HRTFs, this maximises the ability to reduce the dimensionality of HRTFs. This will be measured by averaging the number of recorded dimensions across subjects in the datasets used.

Number of head dimensions provided

The maximum number of head dimensions available is 5. Due to their relatively low number, it would be wise to have most if not all of them. This will be measured by averaging the number of recorded dimensions across subjects in the datasets used.

Number of torso dimensions provided

The maximum number of neck and torso dimensions available is 11. Again, it would be wise to have most if not all of them. This will be measured by averaging the number of recorded dimensions across subjects in the datasets used.

Dataset selection

After assessing the literature, 4 HRTF datasets stood out as potential sources. The ARI dataset [22], HUTUBS [34], AXD [1], and the data from the Oldenburg study that collected HRTFs from a range of microphone locations [12].

Name	ARI	HUTUBS	AXD	Oldenburg
Number of subjects (w/ anthro data)	60	96	200 (0)	19 (0)
Number of ear dimensions	32	24	X	X
Number of head dimensions	5	5	X	X
Number of torso dimensions	10	7	X	X
3D meshes	X	Head only	Head and Torso	X

Table 3: Comparing HRTF datasets

The HUTUBS dataset meets all of the initial dataset specifications, but the other datasets

Requirement	Specification	Value	HUTUBS
Representative Datasets	Number of subjects (with anthropometric data)	≥ 50	96
	Number of ear dimensions provided	≥ 12	12
	Number of head dimensions provided	≥ 4	5
	Number of torso dimensions provided	≥ 8	8
Allows simulation	3D meshes	True/False	True

Table 4: Evaluating HUTUBS dataset against dataset requirements

show promise for a future iteration of this project. The ARI dataset presents a greater number of features, and the AXD dataset possesses more than double the number of subjects. A potential approach is scripting the measurement of anthropometric features from 3D meshes, but this may introduce unwanted errors. The Oldenburg dataset is promising as it has HRTFs measured at the entrance to the ear canal, making it the most directly suitable dataset for real-world validity. However, the reduced number of subjects and lack of anthropometric data is a significant barrier.

2.2 Design Approach

This leads to the following approach map (Figure 11).

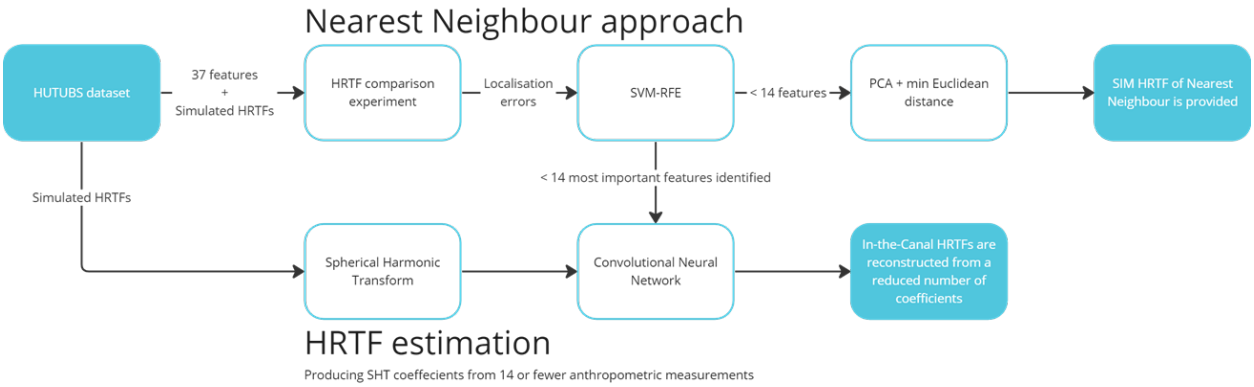


Figure 11: Design Approach

This project contains two approaches. The first, the nearest neighbour approach, seeks to find the physically closest subject in the HUTUBS dataset to the listener. In this approach,

an HRTF comparison experiment is simulated to identify which subjects in the dataset could swap HRTFs and correctly localise sound. This step is seen in several selection-based approaches [30] [32].

Following this, support vector machines (SVM) are constructed with the classification of matches (whether sounds are correctly localised), and recursive feature elimination (RFE) is used to reduce the number of anthropometric features needed to identify a positive match. This approach has also been seen in literature [30]. These features will also be used in the second approach to directly estimate HRTFs.

It is not sufficient to find the subject in the dataset with the smallest Euclidean distance from a patient as this assumes all of the identified features are equally important to identifying a match. Instead, principal component analysis (PCA) will be used to identify components that explain variance in the chosen anthropometric features. Here, the subject with the least minimum Euclidean distance (in this reduced vector space) can be identified and matched to the listener.

The second approach reconfigures a state-of-the-art method by using the anthropometric features chosen by SVM-RFE that predict a positive HRTF match. Using a convolutional neural network (CNN) [38] that directly estimates a reduced dimension form of the patient's HRTF. This dimension reduction technique is known as spherical harmonic transform and encodes frequency data as well as spatial data. By training the CNN on simulated HRTFs instead of measured HRTFs, ITC HRTFs can be directly estimated.

2.2.1 HRTF comparison experiment

The established method of verifying whether another subject's HRTF is suitable for a listener is time and resource-intensive. In this method, a subject wears headphones in a soundproofed room and sound is simulated to appear as if it comes from a specific. Using a laser pointer or some other device, the subject indicates to the nearest degree where they believe the sound is coming from. The angular distance in degrees between the simulated position and

the perceived position is known as the localisation error. This process is repeated from all angular positions, although different experiments use different gradations. However, the resources needed are not accessible, and the time needed to collect data exceeds the timespan afforded to this project.

A powerful alternative is leveraging statistical models that rely on data from such psychoacoustic experiments. Using only an individual's HRTFs, these models can predict localisation error in the horizontal plane [41], the vertical plane [3], or even in all directions [2].

The chosen model was the Wierstorf 2013 model for horizontal localisation. This was chosen for two reasons. Firstly, this model was seen as more robust as horizontal plane localisation highly leverages ITDs and ILDs (see Figure 6) to determine where a sound is coming from. Furthermore, the RMS localisation error of this model (roughly 1°between -60°and 60°) matches that seen in some measured data (1-2°) [37].

It can also be noted that horizontal localisation is more important for determining the location of a sound as real-world listening environments often have speakers in the same vertical plane as a listener. The use of such a model instead of human participants also enables greater iteration of this project.

2.2.2 SVM-RFE

The use of this process is inspired by the approach from Schönstein and Katz to select HRTFs from a dataset using morphological factors [30]. SVMs are a type of supervised machine learning algorithm used for classification and regression tasks. They are particularly effective in high-dimensional spaces and have been used to classify images or text. They perform extremely well with binary classification, non-linear relationships and small datasets, all of which fit this project.

The primary goal of a SVM is to find a hyperplane that best separates the data into different classes while maximising the margin between them. SVMs have a parameter 'C'

that controls the trade-off between having a smooth decision boundary and classifying the training points correctly. A small 'C' allows a softer margin with some misclassifications, while a large 'C' aims for a harder margin with fewer misclassifications. The chosen C value is 10, in line with the referenced work.

The constructed linear SVMs (which have shown equivalent performance to non-linear SVMs), can then have RFE applied to iteratively eliminate anthropometric features by fitting the model with all features, evaluating the importance of each feature and eliminating the least important feature. This process prevents overfitting and minimises the number of anthropometric features that need to be measured by an audiologist.

2.2.3 PCA and minimum Euclidean Distance

PCA presents a further reduced dimensional space to compare a test subject to the training data. While SVM-RFE already reduces the dimensionality of the anthropometric data, these chosen features may be of vastly different scales. For example, height may range between 1.4 metres and 1.8 metres, while ear fossa height can vary between 1.6 centimetres and 2.2 centimetres. This impacts the ability to find the physically closest subject in the dataset.

PCA extracts vectors that carry the most variance in the data, providing a compact and informative representation of the data while also reducing noise. First, the data in each feature is standardised to have a mean of 0 and a standard deviation of 1. From these, a covariance matrix is generated. This matrix describes the relationships between all pairs of variables in the dataset. It indicates the direction of the linear relationship between variables and the strength of that relationship.

PCA then performs eigenvalue decomposition on this matrix to identify eigenvectors and their respective eigenvalues. These eigenvectors are the directions within the data containing the most variance and their respective eigenvalues represent how much variance in the data they explain. After computing the covariance matrix, PCA performs eigenvalue decomposition on this matrix. Eigenvalue decomposition breaks down the covariance matrix into

eigenvectors and eigenvalues. Eigenvectors are the directions along which the data varies the most, while eigenvalues indicate the variance of the data along these directions.

When sorted by their explained variance, these eigenvectors can be used to capture some or all of the variance in the data. Typically, the number of components that explain 95% of the data together are used and the rest discarded. The training data and test subject are then projected onto this reduced dimensional space, where the nearest neighbour to the subject can be identified. The nearest subject will be chosen by multiplying each eigenvector by its eigenvalue to weight it according to its variance. The minimum Euclidean distance in this direction will be used to select the nearest subject in the dataset.

2.2.4 Spherical Harmonic Transforms

This approach seeks to combine a state-of-the-art method [38] of estimating eardrum HRTFs from anthropometric data with the constraints of a hearing aid user having a device fitted with an audiologist. The two key constraints are: not every physical feature can be measured, and eardrum HRTFs differ from ITC HRTFs.

In its base version, this approach predicts the spherical harmonic (SH) coefficients of a listener's HRTFs based on 41 anthropometric features. SHs are functions that represent solutions to Laplace's equation on the sphere. These functions are orthonormal over the surface of the unit sphere, meaning they form a complete basis set for functions defined on the sphere. As a result, any sufficiently smooth function on the sphere can be represented as a linear combination of SH functions. The coefficients allow for any function to be represented on this basis, providing a compact representation of HRTFs, known as their spherical harmonic transform (SHT).

They serve as a powerful form of dimensional reduction for HRTFs as they capture the spatial dependency of HRTFs as well as their frequency information. This makes it a superior dimensionality reduction method compared to PCA, which would require substantially more vectors and ignore the spatial dependency. Another method used is Isomap [17], but this

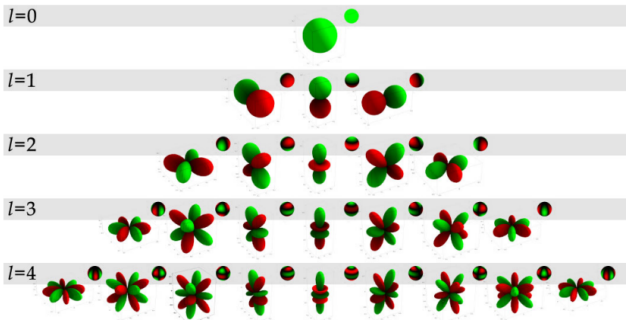


Figure 12: SH coefficients up to the 4th order[16]

technique preserves geodesic distances between data points rather than objective spatial information.

Similar to how PCA can maintain all explained variance or a reduced amount, SHT reduces dimensionality in the form of orders, where each order can be considered an increase in fidelity. This method uses SH coefficients up to the 7th order, leading to 64 SH coefficients. This is significantly smaller than the number of PCA components that would be needed to represent HRTFs.

2.2.5 Convolutional Neural Network

A deep-learning based model is used to map each subject's anthropometric data to the SH coefficients of their HRTF. The structure is shown in Figure 13. Measurements of the ear, head and torso are fed into fully connected layers to obtain their embedding. The frequency index of the HRTFs and whether the ear is left or right is treated as side information and also fed into fully connected layers. These outputs are concatenated and another fully connected layer fuses the information. This is fed into several layers of a 1D CNN that predicts the SH coefficients. Training loss is calculated with the mean square error of the ground-truth SH coefficients and the predicted SH coefficients.

Inverse SHT is applied to the predicted SH coefficients to reconstruct the HRTFs from the desired positions. The inverse Fourier transform will be applied to the HRTFs to form the necessary HRIRs. The accuracy of this approach in its base form is shown in Figure

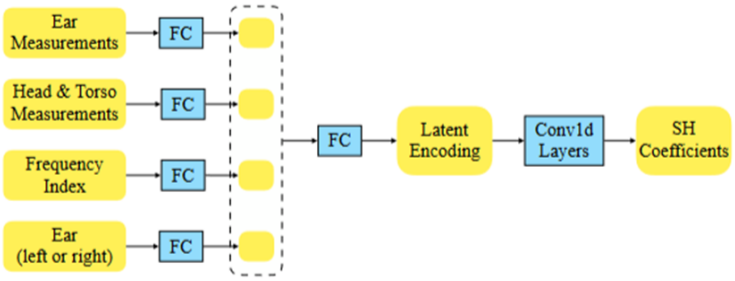


Figure 13: The data-flow diagram of predicting SH coefficients [38].

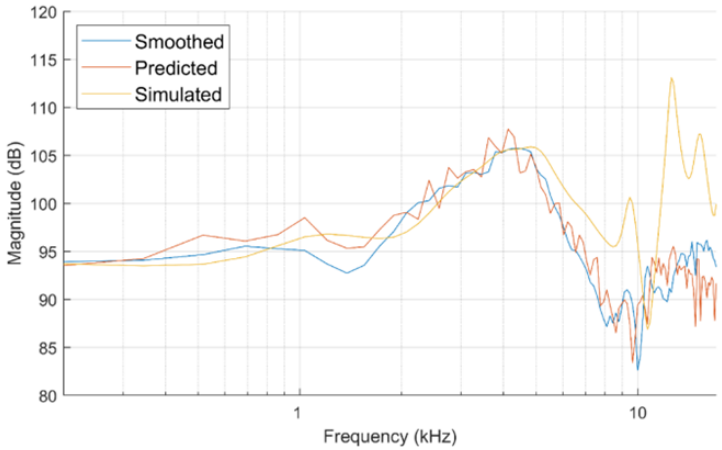


Figure 14: Comparison of smoothed, predicted and simulated HRTFs at 0°elevation and azimuth [38]

14. where it seeks to predict the SH coefficients of measured HRTFs. Here, smoothed refers to the HRTF constructed from ground SH coefficients, predicted refers to the HRTF reconstructed from the SH coefficients predicted by the CNN, and simulated refers to the simulated HRTF from a subject's 3D head mesh [38]. This also serves to demonstrate the difference between HRTFs measured at the eardrum and HRTFs simulated at the entrance to the ear canal.

2.3 Standards

The key engineering standard to be applied is the SOFA convention *SimpleFreeFieldHRIR* Version 1.0 [31]. This convention is a widely accepted method of providing HRTFs for conducting psychoacoustic experiments. The key information provided is the number of

measurements, the positions relative to the listener that the HRIRs have been recorded from, the sample length of the HRIRs and their sampling frequency. Where possible, this project will seek to match the metadata used in the HUTUBS dataset. (Appendix: Table 13)

3 Build

3.1 Nearest Neighbour

3.1.1 HRTF comparison experiment

The HRTF comparison experiment was conducted using the Wierstorf 2013 model from the Auditory Modelling Toolbox. Subjects 1 and 96 as well as 22 and 88 were duplicates of each other and so were removed from testing. Subject 1's HRTFs will be used to test both approaches and excluded from any training data.

This model is designed to test localisation predictions for wave-field synthesis set-up, but the 'estimate azimuth' function makes it suitable for this project. This model works by taking a look-up table of the ITD cues from a reference HRTF belonging to a certain user, across 13 positions between -60°and 60°azimuth. These ITD cues are calculated with the binaural model by Dietz et al. [13]. When a sound source is convoluted with a test HRTF, the Wierstorf model predicts where this virtual listener believes the sound is coming from.

In total, 105300 localisation predictions were calculated (90 HRTFs against 90 HRTFs over 13 positions), creating the need for a simpler metric to determine a match. Instead of using the error at every position for a subject with a set of HRTFs, the RMS error across all 13 positions for each HRTF comparison was computed. It is worth noting this model is developed using measured eardrum HRTFs. For measured HRTFs compared against themselves, an RMS error of around 0.5°to 1°was derived (for positions between -90°and 90°). When the simulated HRTFs were used (compared ear canal entrance HRTFs against each other), this increased to 1°to 1.5°. These errors were significantly diminished when the testing range was reduced to -60°and 60°

3.1.2 SVM-RFE

To build the support vector machines, the calculated RMS errors were classified according to three thresholds. These thresholds aimed to capture matches shown by three peaks in

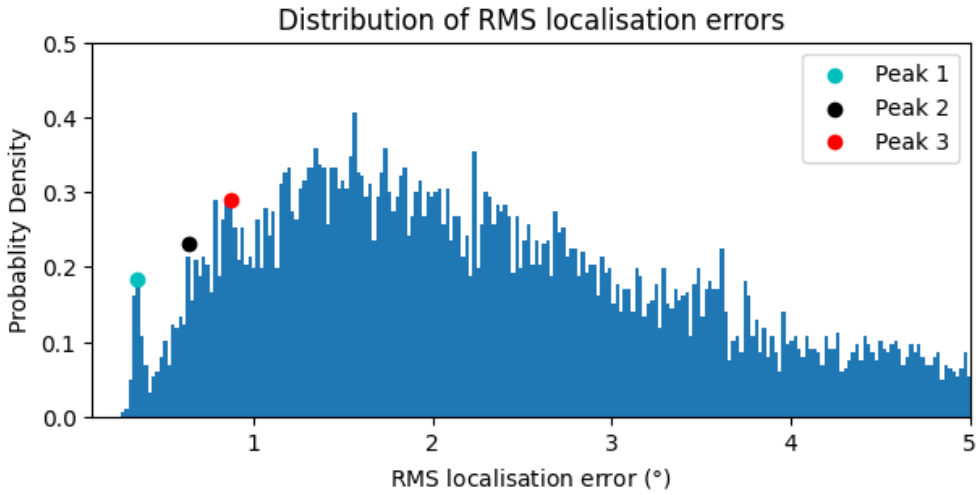


Figure 15: RMS errors of localisation predictions for all subjects against all others

the RMS errors of all the localisation predictions (Figure 15.). Multiple thresholds were used in an attempt to identify the point where an RMS error exceeding a subject with their own HRTF could still be considered a positive match. The performance of each threshold is assessed in section 3.1.3.

The first peak (0.35°) was generally the error of subjects performing localisation with their own HRTFs. This was confirmed by averaging the error of subjects with their own HRTFs and obtaining 0.33°. The second peak was visually determined as 0.64°, the third as 0.87°. The thresholds that captured these peaks were subsequently determined as 0.37°, 0.66°, and 0.89°. RMS errors below these were considered a positive match and classed as a '1', RMS errors below this were considered negative matches and classed as '0'.

Using this classification, a matrix of input data was generated based on a subject's anthropometric data. This data was processed in the same fashion as [30], where the value of each anthropometric feature was divided by the sum of the features. This represents each feature as a percentage of a subject's total features. Each HRTF comparison was then associated with a difference in the percentage of morphological features.

The anthropometric data of subjects 18, 79, and 92 were removed in addition to the subjects excluded from the HRTF comparison experiment as they lacked anthropometric

data. An earlier iteration of this process revealed the pinna flare angle on both ears to be features that predicted a positive HRTF match (Appendix: Table 12). As this feature cannot be easily measured accurately by an audiologist, it was also removed from the anthropometric data.

From here, a linear SVM was constructed using the three different sets of classifiers. RFE was then performed on the SVMs to reduce the number of anthropometric features from 41 to 14, and 3 sets of features were identified to be associated with a positive HRTF match. These features are presented in Tables 5 and 6:

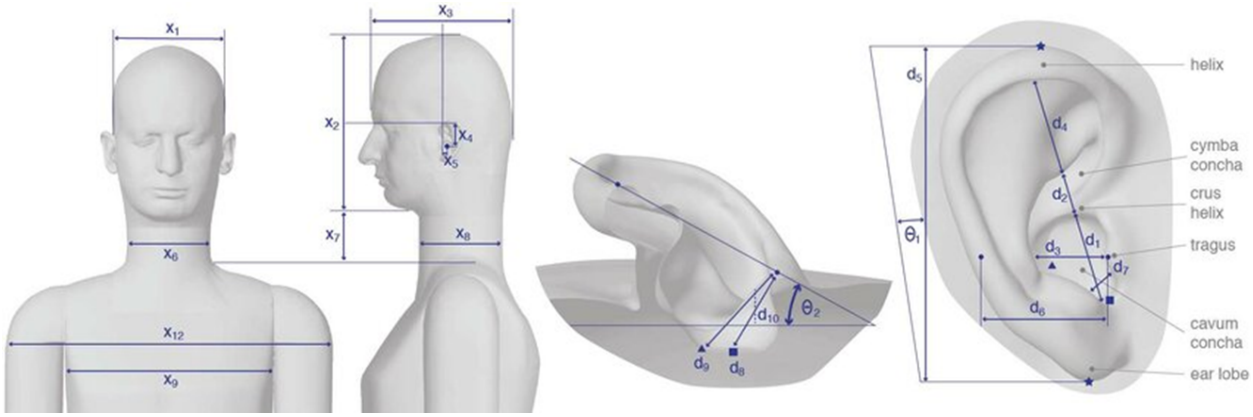


Figure 16: Anthropometric features

Variable Name	Feature Name	Class 1	Class 2	Class 3
x1	Head width			
x2	Head height			
x3	Head depth			
x4	Pinna offset down			
x5	Pinna offset back			
x6	Neck width			
x7	Neck height			
x8	Neck depth			
x9	Torso top width			
x12	Shoulder width			
x14	Height			
x17	Shoulder circumference			

Table 5: Torso features identified by the three classifiers in green

Variable Name	Feature Name	Class 1	Class 2	Class 3
Ld1	Left cavum concha height			
Ld3	Left cavum concha width			
Ld4	Left fossa height			
Ld8	Left cavum concha depth (down)			
Ld9	Left cavum concha depth (back)			
Ltheta1	Left pinna rotation angle			
Rd1	Right cavum concha height			
Rd5	Right pinna height			
Rd6	Right pinna width			
Rd9	Right cavum concha depth (back)			
Rtheta1	Right pinna rotation angle			

Table 6: Ear features identified by the three classifiers in green

3.1.3 PCA and subject selection

PCA was performed using the 14 anthropometric features identified with the 90 valid subjects. This data was normalised and centred across all subjects, allowing a covariance matrix to be generated. The vectors to project the anthropometric data onto can be found from the eigenvectors of this covariance matrix. It was also found that 94% of the variance for all three classification thresholds could be explained through 9 eigenvectors.

The reduced anthropometric data of subject 1 (determined by SVM-RFE) was then projected onto the respective vectors. To prevent all vectors from being given equal weighting, each vector is multiplied by its explained variance to appropriately scale it. When the minimum Euclidean distance was computed, subjects 77, 69 and 65 were identified as being most physically similar to the test subject. The localisation ability of subject 1 was tested using the identified subjects' HRTFs leading to the following results (Figure 17).

This demonstrated that the features identified by the strictest threshold performed best. These features will be used as inputs for the CNN.

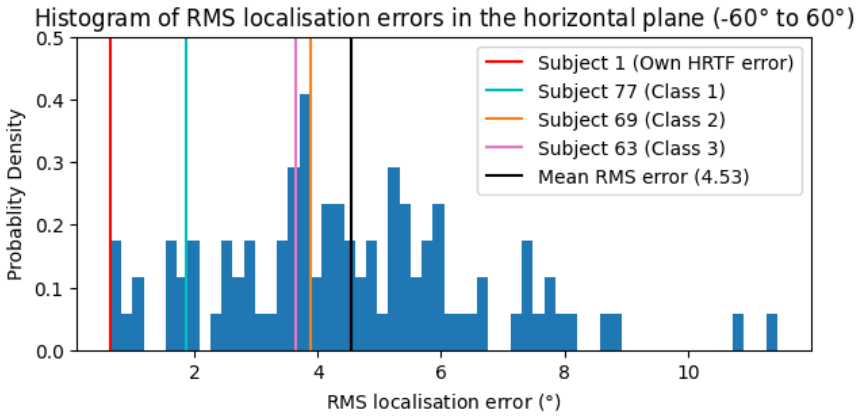


Figure 17: RMS localisation error of subject 1 using simulated HRTFs provided in the HUTUBS dataset

3.2 SHT-CNN

In the original configuration, the CNN predicted the value of 41 frequencies at 64 spherical coefficients from 41 anthropometric features. The HRTFs of a subject in the HUTUBS could be estimated through leave one out validation, where it is excluded from the training data. The learning rate is initially set to 0.0005 with 20% decay for every 100 epochs. The CNN is trained for 1000 epochs. To maintain as much of the performance of the CNN in its original configuration, minimal changes were made.

3.2.1 Reconfiguration

It was first reconfigured by discarding features outside of the chosen 14 but required symmetry in the number of ear features. To solve this issue, unique chosen ear features (where only the left or right feature was chosen) were duplicated onto the other ear, creating subjects with almost symmetrical features. If ears were assumed to be entirely symmetrical, this reduced the number of features to be measured by 1. When assessed by comparing the HRTFs of the same subject with all features against 13 features, there was an apparent reduction in accuracy.

The HRTFs at each position were reconstructed using the inverse SHT. However, the chosen frequencies were not equally spaced and required interpolation before constructing

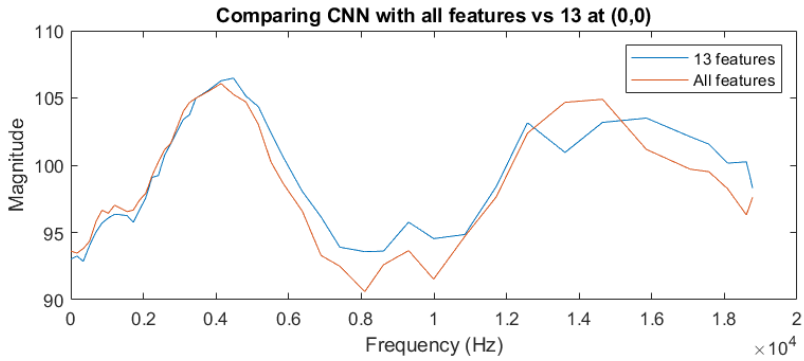


Figure 18: Comparing the reconstructed HRTFs at 0°elevation and azimuth using all anthropometric features(blue) and 13 features (orange)

the HRIR at that position using the inverse FFT. An issue raised when testing these HRTFs with the horizontal localisation model was the initially low sampling rate (34.1 kHz). This was chosen as the maximum frequency constructed was 17.05 kHz, causing the sampling frequency to be twice this.

The initial solution was to increase the number of frequencies predicted by the model to 45 instead as measured and simulated HRTFs have frequency content in this region. This increased the maximum predicted frequency to 18.78 kHz, increasing the sampling frequency to 37.55 kHz. Zero padding in the frequency domain up to 44.1 kHz was considered as this would bring the sampling rate closer to the original HUTUBS HRTFs, but this did not lead to an improvement in results. The effectiveness of this approach is assessed in section 4.1.2.

3.3 Iterations

This project iterated through two versions before arriving at the above implementation.

3.3.1 Version 0

This was a crude version with two objectives. First was whether finding the physically closest subject in a dataset to a test subject would lead to a similar set of HRTFs. Second was the suitability of simulated HRTFs to estimate ITC HRTFs. This version used a restricted number of anthropometric features as these were provided in the KEMAR documentation.

The three nearest subjects were identified by summing the relevant physical features and subtracting from the total of the KEMAR's physical features.

At this stage, perceptual models were not used to assess the fit of HRTFs. Instead, the RMS error between two transfer functions at the same position was used. (Appendix: Figures 27, 28). However, greater agreement was shown when the simulated HRTFs of the physically closest subject were compared to the ITC HRTF of the KEMAR manikin (Appendix: Figures 24 24 24).

3.3.2 Version 1

This was an initial implementation of the approaches outlined in section 2.2. Compared to the previous version, this aimed to use perceptually relevant features that best predicted a positive HRTF match.

Here, localisation predictions were performed between -90 °and 90 °at every integer degree position. This led to 1466100 localisation predictions. However, the data produced was inconsistent as some predictions, particularly at angles outside -60°and 60°, produced a nan value, which carried over in the RMS error calculation for that subject and test HRTF. This is likely caused by the cone of confusion in the horizontal plane (Figure 7), where simulated predictions fell outside the bounds where sounds were being simulated. Nonetheless, this incomplete data was used to select features to predict HRTF matches. The chosen features can be found in Tables 11 and 12.

PCA and the minimum Euclidean distance along these components were used to identify three subjects as physically closest to the test subject, the results of this can be found in Appendix: Figure 29. These results suggested the features obtained by classifying errors below the second strictest threshold selected as positive were better predictors than the strictest. However, both performed better than the mean of all subjects in the HUTUBS dataset.

As this ordering seemed counter-intuitive, the HRTF comparison experiment was per-

formed using bounds of -60°and 60°. This greatly reduced the range of localisation errors but led to significantly improved performance for the strictest threshold. This improvement was due to the identification of new subjects.

4 Evaluation and Verification

4.1 Measure

4.1.1 Tests against specification

The specification points to test both approaches against are defined in Table 7.

Specification	Value	Test
Run time	≤ 0.5 hours	Time from input to output
Number of head and torso measurements	≤ 8	Features identified in SVM-RFE
Number of ear measurements	≤ 6	
Azimuth gradation	$\geq 10°$	Average gradation over all elevations
Elevation gradation	$\geq 15°$	Average gradation over all elevations
RMS localisation error at 0° elevation	$\leq 3°$	Localisation model error between -60°and 60°
RMS localisation error at 0° azimuth	$\leq 42°$	Localisation model error between -30°and 30°

Table 7: Technical specifications and testing methods

Runtime

Runtime is measured by averaging the time it takes to deliver HRTFs after anthropometric data is received. This value is averaged over 3 runs.

Number of features

This is measured by examining the features identified in SVM-RFE, particularly if they are head and torso features, or ear features. These chosen features were seen to be constant when the same training data was used, meaning it was not necessary to take an average.

Gradation

Azimuth gradation is calculated by dividing 360°by the number of unique azimuth positions, Elevation gradation is calculated by dividing 180°by the number of unique elevation positions. This is also constant and an average is not taken.

Localisation error

Horizontal RMS localisation error is calculated using the horizontal localisation model seen in the HRTF comparison experiment (section 2.2.1). Based on subject 1's HRTFs, a localisation experiment is simulated using the HRTFs delivered by the approach. 5 trials are completed and an average is taken.

Vertical RMS localisation error is assessed in a similar fashion, although using the Baumgartner 2014 model [3] between -30°and 30°. These are the bounds used in the Middlebrooks experiment [26] that led to the specification threshold of 42°.

4.1.2 Test results

The results of the nearest neighbour approach are in Table 8, and the results of the direct estimation approach are in 9.

Specification	Value	Measured Value
Run time	≤ 0.5 hours	0.062 ± 0.013 seconds
Number of head and torso measurements	≤ 8	7 features
Number of ear measurements	≤ 6	7 features (6 for identical ears)
Azimuth gradation	$\geq 10°$	0.464°
Elevation gradation	$\geq 15°$	0.747°
RMS localisation error at 0° elevation	$\leq 3°$	$1.86° \pm 0.45°$
RMS localisation error at 0° azimuth	$\leq 42°$	$47.61° \pm 1.11°$

Table 8: Technical specifications and results of nearest neighbour approach

Specification	Value	Measured Value
Run time	≤ 0.5 hours	11.4 ± 1.2 minutes
Number of head and torso measurements	≤ 8	7 features
Number of ear measurements	≤ 6	6 features
Azimuth gradation	$\geq 10°$	5°
Elevation gradation	$\geq 15°$	9.47°
RMS localisation error at 0° elevation	$\leq 3°$	$37.05° \pm 0.8°$
RMS localisation error at 0° azimuth	$\leq 42°$	$51.84° \pm 1.37°$

Table 9: Technical specifications and results of direct estimation approach

4.2 Analysis and Verification

4.2.1 Nearest Neighbour

This project met most specification points except the number of ear measurements needed and the vertical localisation error.

The runtime was as expected as there are three processing stages for the input data. First is projecting it into the component space, next is finding the minimum Euclidean distance from all other points in the component space, and last is sorting the distances and identifying the point that distance corresponds to.

As the HRTF matches were based on horizontal localisation, it was expected that more head features would produce positive matches. This is caused by the inter-aural time difference (the delay between the same sound reaching different ears) and the inter-aural level difference (the difference in amplitude from the same sound at both ears). As the dimensions of the head describe the distance between the ears and the physical matter responsible for absorbing some sound, it was expected that there would be fewer ear features associated with a positive match. Most importantly, all features can be measured with intuitive tools such as outside callipers and photography.

The gradation in both respects significantly outperformed the specification, but this was expected as simulated HRTFs are being output (1730 positions). As these are generated using BEM, there is no lower limit for average gradation.

The horizontal localisation error is as expected, most likely due to the horizontal model being used to determine which features should lead to a positive HRTF match. This also demonstrates that a positive match for horizontal localisation does not necessitate a positive match for vertical localisation. Multiple runs of both localisation experiments showed the HRTFs delivered would consistently meet the specification for horizontal localisation, but not for vertical localisation.

4.2.2 Direct Estimation

This project met most specification points except the localisation errors in both the horizontal plane and vertical plane.

The runtime was faster than expected as the current implementation trains on all data excluding a chosen subject, before deriving HRTFs and subsequently HRIRs from the predicted SH coefficients. However, it should be noted that this value may depend on the processing power of a client's computer. While parallelization via a GPU was not used, the machine possesses an AMD EPYC 7763 64-core Processor with 64 GB of RAM, which is likely to be stronger than most audiologist's computers.

As this approach requires an equal number of ear features from both sides, some element of symmetry was required. Inspection of the HUTUBS anthropometric data revealed the average difference between the right and left cavum concha depth (Ld9) was 0.00054 cm. This corresponds to a difference of 0.54% between ears for this dimension. As a result, it is valid to make the assumption the ears are symmetrical for this feature.

Gradation meets the specification as expected as it uses the positions used to generate the HUTUBS measured HRTFs (440 positions). As HRTFs are derived from SH coefficients, it is again possible to derive HRTFs from any position around the user.

Most critically, localisation errors were significantly higher than expected of the SHT-CNN approach. The horizontal error performed worse than all simulated HRTFs in the HUTUBS dataset, suggesting a significant flaw in the HRIRs derived. The localisation predictions of subject 1 using the estimated HRTF are shown in Figure 19.

This shows the localisation model struggles to predict a range of horizontal positions from the reconstructed HRTFs. This error could be caused by an incompatibility between the reconstructed HRTFs and the model. One significant difference between the reconstructed HRTFs and standard HRTFs is the lower sampling frequency. Alternatively, as this model relies on ITDs computed through the Dietz model [13], this issue could be traced to this intermediary stage. Another issue is the SHT-CNN approach could be considered numerically

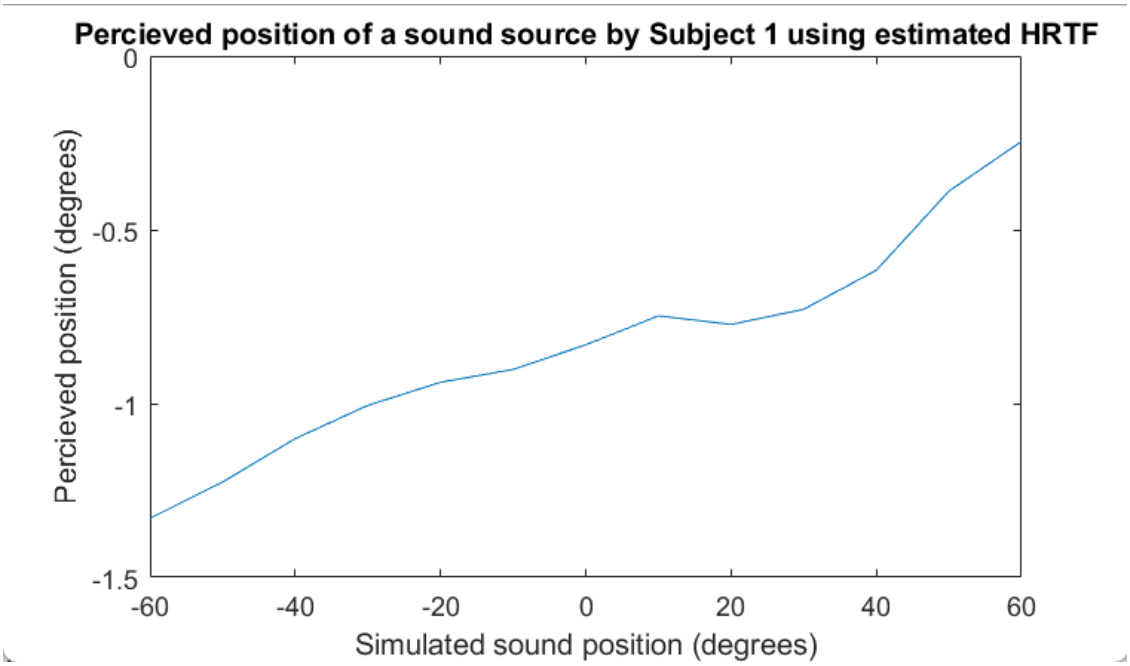


Figure 19: Perceived position of sound sources by subject 1 using the HRTFs reconstructed from predicted SH coefficients.

strong, but perceptually weak. This is to say that the transfer functions predicted show similarities with the ground truth functions 18, but it must be noted that not all frequencies equally contribute to localisation. As a result, if the error in the reconstructed HRTFs is concentrated near the most important frequencies, the perceptual error is magnified. It is also noted by the team behind this approach that prediction error accumulates with increased frequency. This is hypothesised to be caused by the insufficiency of anthropometric measurements for 'predicting fine structures in HRTFs' [38]. Finally, the CNN was tuned for measured HRTFs and fewer anthropometric features. Further reconfiguration is necessary for adequate testing and desirable results.

4.2.3 Feature Measurement

While not formally part of the specification, the accuracy with which anthropometric features could be measured was considered. Using the manikin shown in Figure 8 and a 3D scan obtained from [4], the SVM-RFE chosen features were measured and shown in Table 10.

Physical measurements were obtained using a Vernier calliper, a steel rule, and the app-based measuring tool *Angulus*. The accuracy of *Angulus* was assessed with a protractor before use. The values shown below are the average of two rounds of measuring. 3D measurements were made of the mesh file in *Autodesk Fusion 360*. Identical features were assumed for the KEMAR's ears as it is manufactured in that manner.

Feature name	Physical measurement (mm)	3D measurement (mm)	Error (%)
Head width	148.79	154.497	3.68
Head height	216	223.534	3.37
Head depth	198	193.517	2.26
Pinna offset down	6.59	26.233	74.9
Pinna offset back	9.13	4.126	121.23
Neck width	112.55	115.037	2.16
Neck height	71.62	73.839	3.01
Cavum concha width	19.67	20.70	4.98
Fossa height	16.84	10.214	64.87
Cavum concha depth back	20.58	20.82	1.15
Pinna Height	62.83	60.919	3.14
Pinna width	28.35	28.70	1.22
Pinna rotation angle	11.2 °	8.4 °	33

Table 10: Measurements of the KEMAR maninkin

The average error is 24.54%, but this can be explained by the highly inaccurate measurements of pinna offset down and back, and fossa height. Accounting for these, the average error is 5.80 %. These significant errors could be caused by a few issues. First, the ears attached to the measured KEMAR were not easily identifiable as the default ears. This could cause discrepancies with the 3D-scanned KEMAR. Secondly, for features such as pinna offset (which is the distance of the ear from the centre of the skull), it is difficult to determine the exact centre of a skull without specialist equipment. For suitable use with human subjects, robust measurement protocols must be identified, especially for these features.

5 Conclusion

5.1 Testing summary

The effectiveness of both approaches is summarised in Figures 20 and 21.

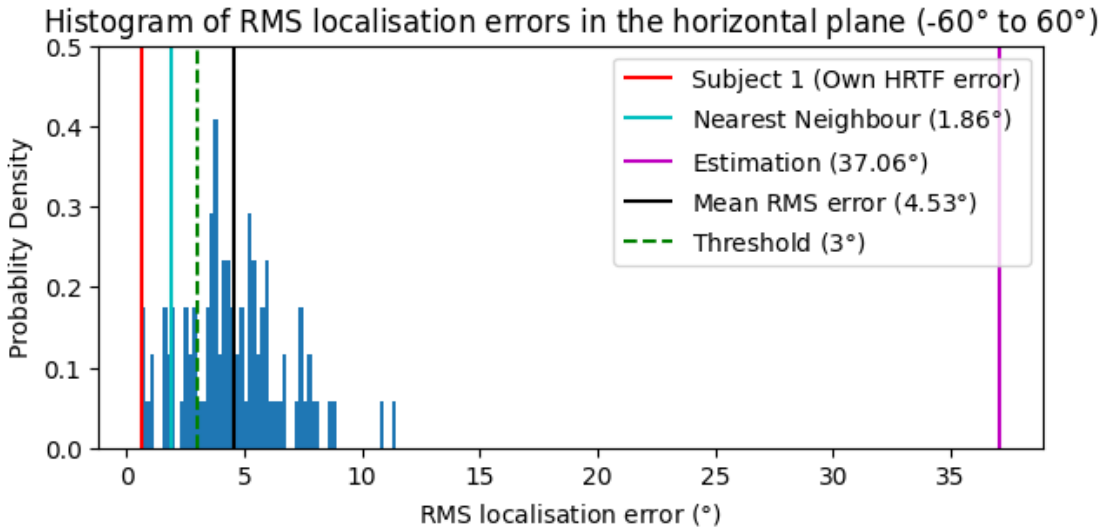


Figure 20: Horizontal RMS localisation errors of subject 1 using all simulated HRTFs in the HUTUBS dataset

It is clear the nearest neighbour approach is superior to the direct estimation approach in both horizontal and vertical localisation. The nearest neighbour approach error is below the threshold for the maximum horizontal RMS error and below the mean for all HRTFs in the HUTUBS dataset. Conversely, the direct estimation approach is above the mean in both cases and meets neither threshold.

5.2 Discussion

Some key reasons for the poor performance of the direct estimation approach have been identified. As mentioned previously, the direct estimation method could be considered numerically strong, yet perceptually weak. This could be improved upon by incorporating the localisation models used in the CNN's error computation for each training round.

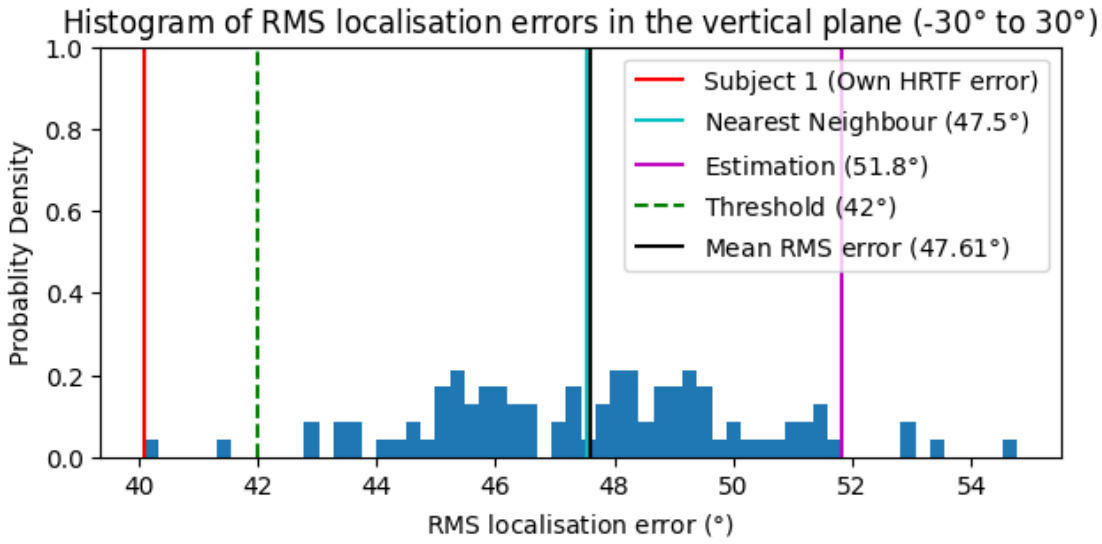


Figure 21: Vertical RMS localisation errors of subject 1 using all simulated HRTFs in the HUTUBS dataset

Secondly, relative to other deep-learning approaches, there is relatively little data available. This approach aims to find the relationship between anthropometric features and the SH coefficients of a subject's HRTFs. By significantly reducing the number of features used for prediction, the number of subjects may be insufficient to maintain prediction accuracy.

On the other hand, the nearest-neighbour method uses both perceptual measures and comparisons between subjects. This compares 90 data points (with anthropometric data and HRTFs) with 8100 datapoints (where each one possesses differences in anthropometric features and a classification of whether a set of HRTFs would be suitable for a subject). Nonetheless, if robust ITC HRTF and anthropometric data could be collected on a larger scale, this could improve the success of the direct estimation method.

It should be noted that the nearest neighbour method performed slightly better than chance for vertical localisation, and did not meet the specification for vertical localisation. This suggests a set of HRTFs that lead to sufficient horizontal localisation do not necessarily provide suitable vertical localisation. In a future iteration of the nearest neighbour method, it would be prudent to incorporate results from a vertical localisation model. However, preliminary experiments will be needed to determine how much emphasis should be placed

on vertical localisation.

Another specification point not met was the number of ear features identified by SVM-RFE. Right and left cavum concha depth (back) were both identified, but differed by 0.054% between subjects in the dataset. This minimal difference suggests they can be treated as a single measure, bringing the nearest neighbour approach in line with the specification.

5.3 Future Direction

3 directions for future work have been identified in addition to the solution mentioned above. First, in-situ testing with a subject for which ITC HRTFs exist. This process was started using ITC HRTFs from the Oldenburg dataset and real-life measurements taken of the KEMAR manikin, but the HRTFs provided appear to be incompatible with the horizontal localisation model.

This leads to the second direction - conducting ITC HRTF measurements as well as localisation experiments to construct a localisation model specific to ITC HRTFs. One issue with both models used is their foundation in eardrum HRTF experiments. Localisation errors were higher with simulated HRTFs rather than measured HRTFs. This measurement and experiment would also provide robust data for the direct estimation approach, likely improving its performance.

The third direction is combining the approaches. The direct estimation approach predicts SH coefficients of HRTFs from anthropometric data. In each training round, it compares the predicted SH coefficients to the ground truth SH coefficients. The nearest neighbour approach could be incorporated by giving greater weight to subjects in the training data most similar to the test subject. This would be another way of incorporating perceptually relevant information into the estimation process.

5.4 Larger Context

This algorithm is expected to operate with support from audiologists and virtual-reality researchers to maximise patient outcomes from auditory rehabilitation. The direct cost associated with this project is the measurement tools necessary to record anthropometric features. $30 to $50 is an exceedingly small cost for audiologists when compared to other equipment costing upwards of $3000 [33].

It is up to the audiologist whether this economic cost will fall on the patient. However, if this algorithm enables virtual-reality rehabilitation to maximise patient outcomes, then the social impact of hearing loss will be reduced. For individuals who struggle to understand speech in work and educational environments, this project indirectly provides economic benefits.

Nonetheless, it must also be considered that entirely immersive virtual reality headsets are expensive. While the multi-sensory aspect of speech is not fully understood, the ecological validity of virtual reality rehabilitation is rooted in recreating real-life listening environments. While this project enables the auditory aspect, low-cost methods of visual immersion must also be explored.

6 References

References

[1] *AXD Website*. URL: https://www.axdesign.co.uk/tools-and-devices/sonicom-hrtf-dataset (visited on 12/07/2023).

[2] Roberto Barumerli et al. "A Bayesian model for human directional localization of broadband static sound sources". In: *Acta Acustica* 7 (2023). Publisher: EDP Sciences, p. 12. ISSN: 2681-4617. DOI: 10.1051/aacus/2023006. URL: https://acta-acustica.edpsciences.org/articles/aacus/abs/2023/01/aacus210056/aacus210056.html (visited on 02/24/2024).

[3] Robert Baumgartner, Piotr Majdak, and Bernhard Laback. "Modeling the Effects of Sensorineural Hearing Loss on Sound Localization in the Median Plane". In: *Trends in Hearing* 20 (Jan. 1, 2016). Publisher: SAGE Publications Inc, p. 2331216516662003. ISSN: 2331-2165. DOI: 10.1177/2331216516662003. URL: https://doi.org/10.1177/2331216516662003 (visited on 02/24/2024).

[4] Hark Simon Braren and Janina Fels. *A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: 3D Data*. RWTH-2020-06760. Lehr- und Forschungsgebiet für Medizinische Akustik, 2020. DOI: 10.18154/RWTH-2020-06760. URL: https://publications.rwth-aachen.de/record/793260 (visited on 10/28/2023).

[5] Fabian Brinkmann et al. "A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses". In: *Journal of the Audio Engineering Society* 67.9 (Sept. 21, 2019). Publisher: Audio Engineering Society, pp. 705–718. URL: https://www.aes.org/e-lib/browse.cfm?elib=20546 (visited on 03/27/2024).

[6] Gestur Björn Christianson. "Information Processing in the Interaural Time Difference Pathway of the Barn Owl". In: Medium: PDF Version Number: Final. [object Object], Nov. 22, 2005. DOI: 10.7907/3WJJ-7294. URL: https://resolver.caltech.edu/CaltechETD:etd-11212005-110457 (visited on 04/01/2024).

[7] *Computing the HRTF of a Scanned Geometry of a Human Head*. COMSOL. URL: https://www.comsol.com/blogs/computing-the-hrtf-of-a-scanned-geometry-of-a-human-head/ (visited on 12/07/2023).

[8] N. A. Daikhes et al. "[The effectiveness of auditory training using virtual reality technologies in persons with chronic sensorineural hearing loss]". In: *Vestnik Otorinolaringologii* 86.6 (2021), pp. 17–21. ISSN: 0042-4668. DOI: 10.17116/otorino20218606117.

[9] R. Daugintis et al. "Initial evaluation of an auditory-model-aided selection procedure for non-individual HRTFs". In: *Forum Acusticum*. Accepted: 2023-09-22T13:55:20Z ISSN: 2221-3767. Sept. 22, 2023. URL: http://spiral.imperial.ac.uk/handle/10044/1/106574 (visited on 10/30/2023).

[10] Karina C. De Sousa et al. "Effectiveness of an Over-the-Counter Self-fitting Hearing Aid Compared With an Audiologist-Fitted Hearing Aid: A Randomized Clinical Trial". In: *JAMA otolaryngology– head & neck surgery* 149.6 (June 1, 2023), pp. 522–530. ISSN: 2168-619X. DOI: 10.1001/jamaoto.2023.0376.

[11] *Deafness and hearing loss*. URL: https://www.who.int/health-topics/hearing-loss (visited on 12/03/2023).

[12] Florian Denk et al. "Adapting Hearing Devices to the Individual Ear Acoustics: Database and Target Response Correction Functions for Various Device Styles". In: *Trends in Hearing* 22 (Jan. 1, 2018). Publisher: SAGE Publications Inc, p. 2331216518779313. ISSN: 2331-2165. DOI: 10.1177/2331216518779313. URL: https://doi.org/10.1177/2331216518779313 (visited on 11/04/2023).

[13] Mathias Dietz, Stephan D. Ewert, and Volker Hohmann. "Auditory model based direction estimation of concurrent speakers from binaural signals". In: *Speech Communication*. Perceptual and Statistical Audition 53.5 (May 1, 2011), pp. 592–605. ISSN: 0167-6393. DOI: 10.1016/j.specom.2010.05.006. URL: https://www.sciencedirect.com/science/article/pii/S016763931000097X (visited on 02/24/2024).

[14] Isaac Engel, Dan F. M. Goodman, and Lorenzo Picinali. "Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models". In: *Acta Acustica* 6 (2022). Publisher: EDP Sciences, p. 4. ISSN: 2681-4617. DOI: 10.1051/aacus/2021055. URL: https://acta-acustica.edpsciences.org/articles/aacus/abs/2022/01/aacus210029/aacus210029.html (visited on 03/27/2024).

[15] Dr Stella Fulman. *Why are Speech-In-Noise test important?* Audiology Island. Sept. 4, 2020. URL: https://audiologyisland.com/blog/why-are-speech-in-noise-test-important/ (visited on 07/26/2023).

[16] *Gen Afanasev — Graphics Programmer blog: Spherical Harmonics in Graphics: A Brief Overview*. URL: https://gen-graphics.blogspot.com/2017/11/spherical-harmonics-in-graphics-brief.html (visited on 04/02/2024).

[17] Felipe Grijalva et al. "Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, May 2014, pp. 4473–4477. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854448. URL: http://ieeexplore.ieee.org/document/6854448/ (visited on 11/04/2023).

[18] Peter Grosche. *The HUTUBS head-related transfer function (HRTF) database*. Peter Grosche - Audio DSP, Data Science, Music AI. Aug. 16, 2021. URL: https://petergrosche.github.io/publication/11303-9429/ (visited on 10/21/2023).

[19] *Hearing Aid Sizes and Models*. My Hearing Centre. URL: http://www.myhearingcentre.com.au/hearing-solutions/models-sizes/ (visited on 12/07/2023).

[20] Paul M. Hofman, Jos G. A. Van Riswick, and A. John Van Opstal. "Relearning sound localization with new ears". In: *Nature Neuroscience* 1.5 (Sept. 1998). Number: 5 Publisher: Nature Publishing Group, pp. 417–421. ISSN: 1546-1726. DOI: 10.1038/1633. URL: https://www.nature.com/articles/nn0998_417 (visited on 10/21/2023).

[21] *HRTF-Database*. URL: https://www.oeaw.ac.at/isf/das-institut/software/hrtf-database (visited on 04/01/2024).

[22] *Index of /data/database/ari*. URL: https://sofacoustics.org/data/database/ari/ (visited on 12/07/2023).

[23] *Index of /data/database/axd*. URL: https://sofacoustics.org/data/database/axd/ (visited on 12/07/2023).

[24] Subong Kim, Caroline Emory, and Inyong Choi. "Neurofeedback Training of Auditory Selective Attention Enhances Speech-In-Noise Perception". In: *Frontiers in Human Neuroscience* 15 (2021). ISSN: 1662-5161. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2021.676992 (visited on 08/16/2023).

[25] Geon Woo Lee and Hong Kook Kim. "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear". In: *Applied Sciences* 8.11 (Nov. 2018). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 2180. ISSN: 2076-3417. DOI: 10.3390/app8112180. URL: https://www.mdpi.com/2076-3417/8/11/2180 (visited on 10/30/2023).

[26] John Middlebrooks. "Individual differences in external-ear transfer functions reduced by scaling in frequency. Journal of the Acoustical Society of America 106:1480-1492". In: *The Journal of the Acoustical Society of America* 106 (Oct. 1, 1999), pp. 1480–92. DOI: 10.1121/1.427176.

[27] Douglas A. Miller and Mohammad A. Matin. "A model for predicting localization performance in cochlear implant users". In: *World Journal of Neuroscience* 3.3 (July 11, 2013). Number: 3 Publisher: Scientific Research Publishing, pp. 136–141. DOI: 10.4236/wjns.2013.33017. URL: https://www.scirp.org/journal/paperinformation.aspx?paperid=34955 (visited on 10/11/2023).

[28] *Quick Statistics About Hearing — NIDCD*. Mar. 25, 2021. URL: https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing (visited on 05/01/2023).

[29] M. Risoud et al. "Sound source localization". In: *European Annals of Otorhinolaryngology, Head and Neck Diseases* 135.4 (Aug. 1, 2018), pp. 259–264. ISSN: 1879-7296. DOI: 10.1016/j.anorl.2018.04.009. URL: https://www.sciencedirect.com/science/article/pii/S187972961830067X (visited on 10/30/2023).

[30] David Schönstein and Brian F G Katz. "HRTF selection for binaural synthesis from a database using morphological parameters". In: (2010).

[31] *SimpleFreeFieldHRIR - Sofaconventions*. URL: https://www.sofaconventions.org/mediawiki/index.php/SimpleFreeFieldHRIR (visited on 03/28/2024).

[32] Simone Spagnol. "HRTF Selection by Anthropometric Regression for Improving Horizontal Localization Accuracy". In: *IEEE Signal Processing Letters* 27 (2020). Conference Name: IEEE Signal Processing Letters, pp. 590–594. ISSN: 1558-2361. DOI: 10.1109/LSP.2020.2983633. URL: https://ieeexplore.ieee.org/document/9050904 (visited on 11/09/2023).

[33] Hearing Review Staff. *Cost-effective Pricing for Hearing Aids and Related Audiological Services*. The Hearing Review. Nov. 2, 2011. URL: https://hearingreview.com/hearing-products/hearing-aids/cost-effective-pricing-for-hearing-aids-and-related-audiological-services (visited on 04/01/2024).

[34] *The HUTUBS head-related transfer function (HRTF) database*. In collab. with Fabian Brinkmann et al. 2019. DOI: 10.14279/DEPOSITONCE-8487. URL: https://depositonce.tu-berlin.de/handle/11303/9429 (visited on 10/30/2023).

[35] Alvin R. Tilley and Henry Dreyfuss Associates. *The Measure of Man and Woman: Human Factors in Design*. Revised edition. New York: Wiley, Dec. 31, 2001. 112 pp. ISBN: 978-0-471-09955-0.

[36] *Using your ears and head to escape the Cone Of Confusion · Chris Said*. URL: https://chris-said.io/2018/08/06/cone-of-confusion/ (visited on 10/11/2023).

[37] Carl A. Verschuur et al. "Auditory Localization Abilities in Bilateral Cochlear Implant Recipients". In: *Otology & Neurotology* 26.5 (Sept. 2005), p. 965. ISSN: 1531-7129. DOI: 10.1097/01.mao.0000185073.81070.07. URL: https://journals.lww.com/otology-neurotology/abstract/2005/09000/auditory_localization_abilities_in_bilateral.23.aspx (visited on 10/29/2023).

[38] Yuxiang Wang et al. "Global HRTF Personalization Using Anthropometric Measures". In: (2021).

[39] Yuxiang Wang et al. *Predicting Global Head-Related Transfer Functions From Scanned Head Geometry Using Deep Learning and Compact Representations*. July 28, 2022. DOI: 10.48550/arXiv.2207.14352. arXiv: 2207.14352[cs,eess]. URL: http://arxiv.org/abs/2207.14352 (visited on 11/19/2023).

[40] *Weaving Sounds With One's Ears - Sound Localization*. URL: http://www.ais.riec.tohoku.ac.jp/Lab3/localization/index.html (visited on 03/27/2024).

[41] H. Wierstorf, A. Raake, and S. Spors. "Binaural Assessment of Multichannel Reproduction". In: *The Technology of Binaural Listening*. Ed. by Jens Blauert. Modern Acoustics and Signal Processing. Berlin, Heidelberg: Springer, 2013, pp. 255–278. ISBN: 978-3-642-37762-4. DOI: 10.1007/978-3-642-37762-4_10. URL: https://doi.org/10.1007/978-3-642-37762-4_10 (visited on 12/11/2023).

[42] D.Y.N. Zotkin et al. “HRTF personalization using anthropometric measurements”. In: *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, USA: IEEE, 2003, pp. 157–160. ISBN: 978-0-7803-7850-6. DOI: 10.1109/ASPAA.2003.1285855. URL: http://ieeexplore.ieee.org/document/1285855/ (visited on 10/21/2023).

7 Appendix

7.1 Budget

BOM (Bill of Materials)	Unit Cost $	Unit	# of Units	Total $	Exact or estimated?
Digital Protractor	$ 33.15	1	1	$ 33.15	Exact
Please do your best to fill in the following	**Total $**		**Exact or estimated?**		
Total Development cost: everything that was	33.15		Exact		
What is the minimum cost to make one	0				
Total cost of items purchased through the	33.15				
Total cost covered by the *Harvard Research*					
Total cost of items purchase personally, if any					
Total cost covered by a *non-Harvard lab*					
Please list below all material used in **ALL**					
Virtual Desktop					
Vernier Caliper					

Figure 22: Budget Report

7.2 Code

The code associated with this project can be found at `https://github.com/Sprog15/ITC-HRTFs`

7.3 Tables and Figures

Variable Name	Feature Name	Class 1	Class 2	Class 3
x1	Head width	■		■
x2	Head height	■		
x3	Head depth		■	■
x5	Pinna offset back	■		
x6	Neck width		■	■
x7	Neck height	■	■	■
x8	Neck depth	■	■	■
x12	Shoulder width	■		
x14	Height	■	■	
x17	Shoulder circumference	■	■	■

Table 11: Head and torso features identified by the three classifiers in a previous iteration

Variable Name	Feature Name	Class 1	Class 2	Class 3
Ld4	Left fossa height		■	■
Ld5	Left pinna height			■
Ld8	Left cavum concha depth (down)	■		■
Ltheta2	Left pinna flare angle		■	■
Rd3	Right cavum concha width		■	■
Rd4	Right fossa height	■		
Rd5	Right pinna height	■	■	■
Rd6	Right pinna width	■	■	
Rd8	Right cavum concha depth (down)	■		■
Rtheta1	Right pinna rotation angle		■	■
Rtheta2	Right pinna flare angle	■	■	

Table 12: Ear features identified by the three classifiers in a previous iteration

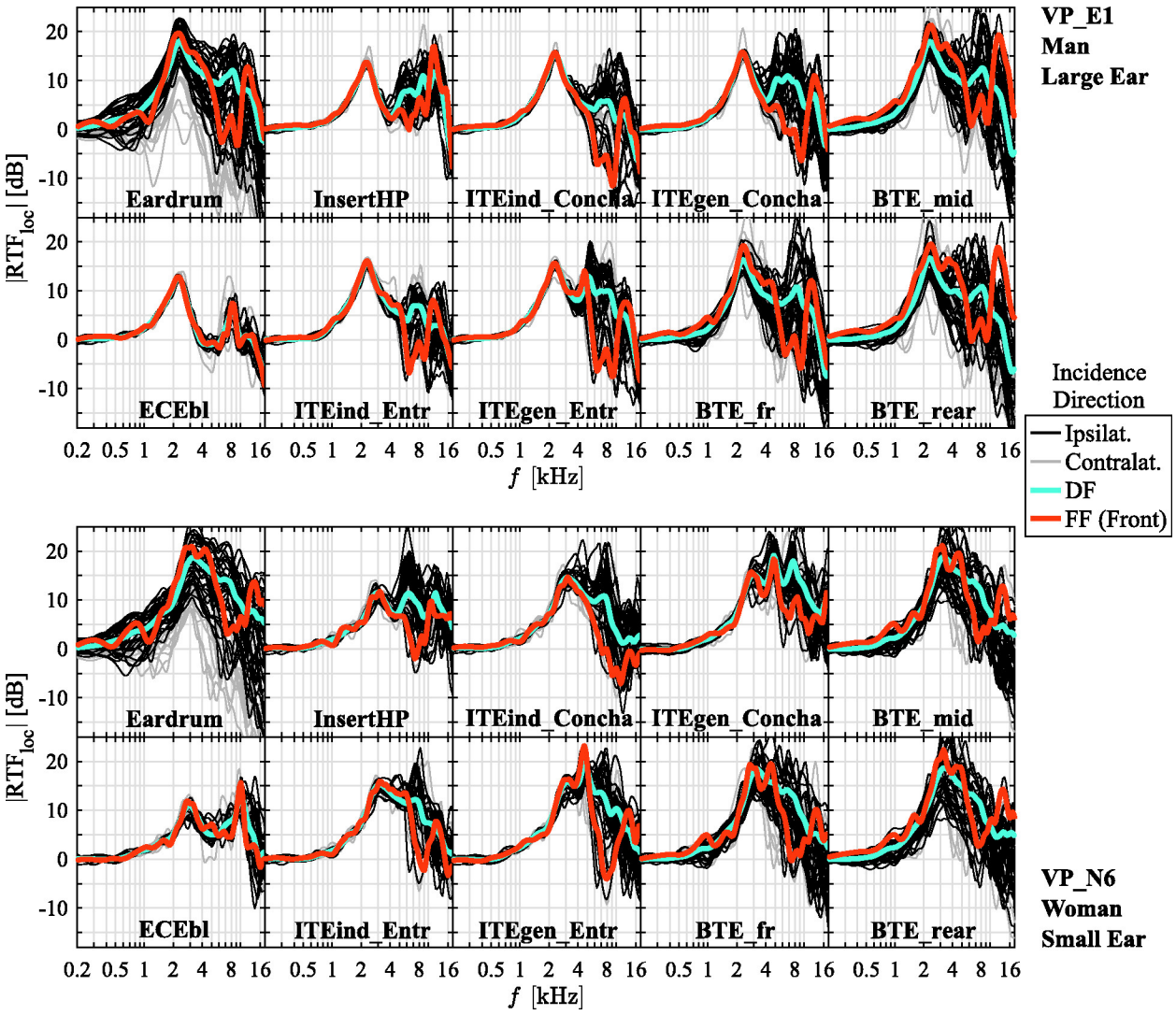


Figure 23: Relative transfer functions (RTF) between all hearing device microphone locations and the eardrum of the open ear for all incidence directions, as well as free-field (FF, i.e., frontal incidence) and diffuse-field (DF) incidence, for two individual subjects. For the eardrum, the corresponding HRTF is shown. VPE1 is a man with large ears and VPN6 a woman with small ears [12].

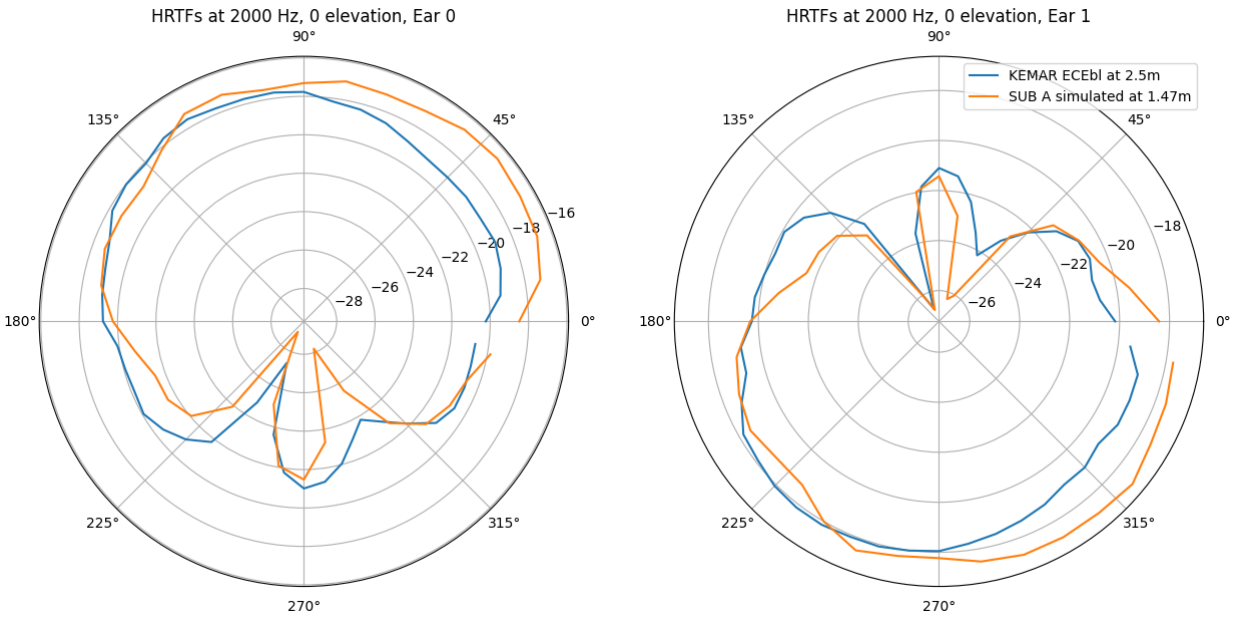


Figure 24: Comparing HRTFs at 2000Hz between the physically closest subject (simulated) and the KEMAR (ECEbl)

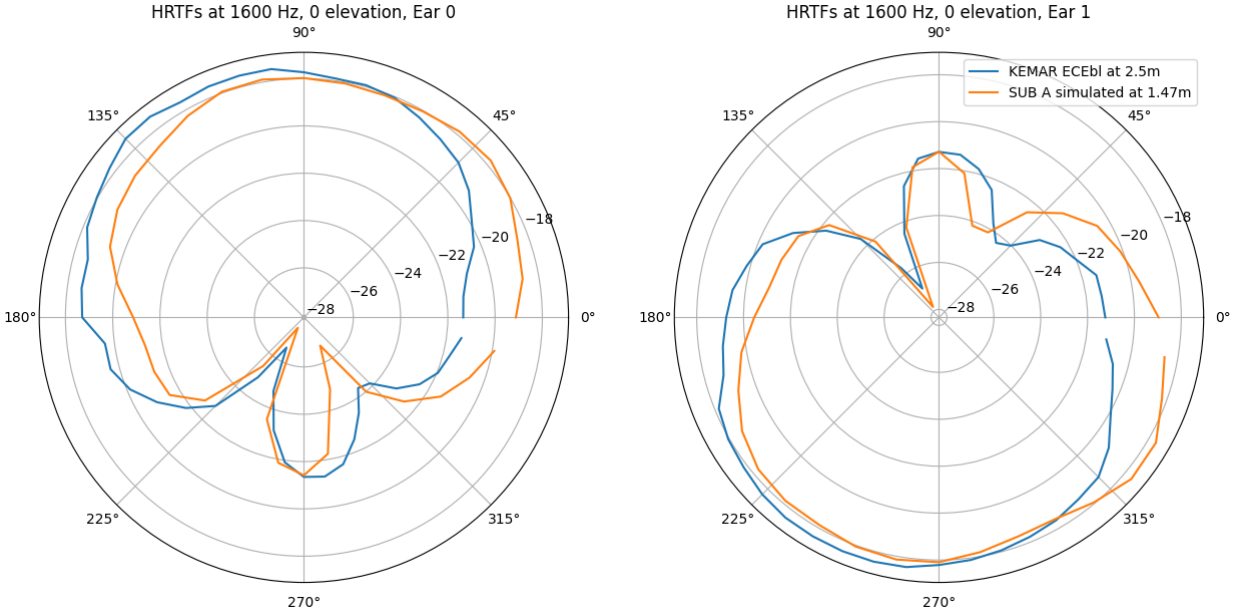


Figure 25: Comparing HRTFs at 1600Hz between the physically closest subject (simulated) and the KEMAR (ECEbl)

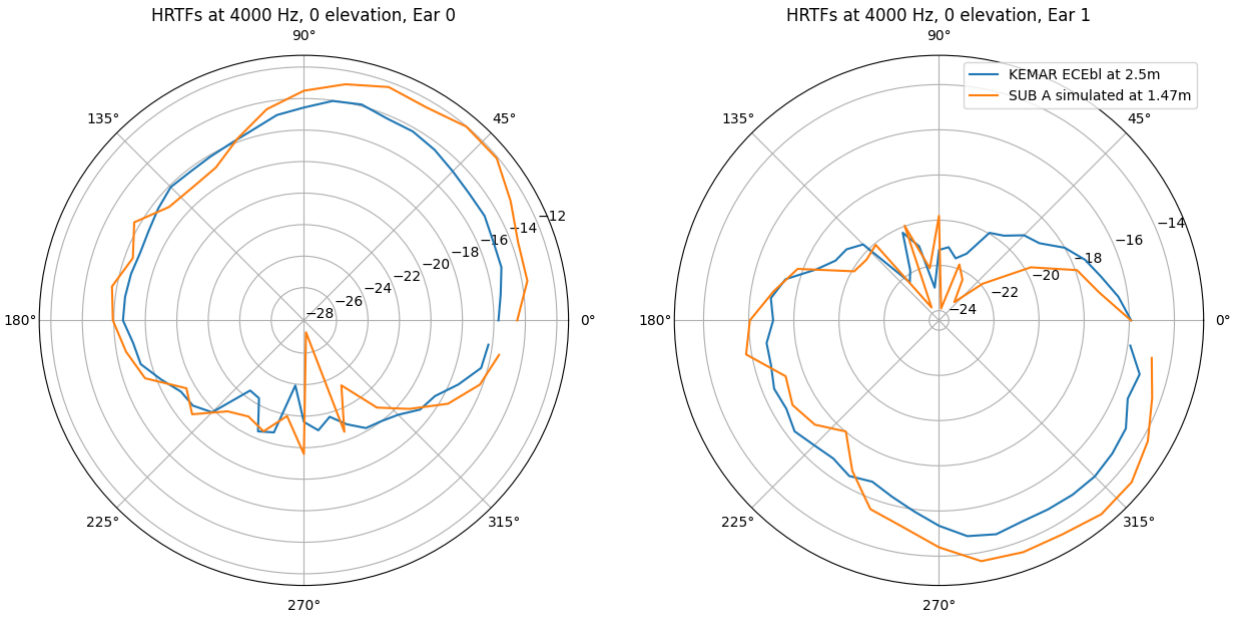


Figure 26: Comparing HRTFs at 4000Hz between the physically closest subject (simulated) and the KEMAR (ECEbl)

SOFA Fields	HUTUBS metadata
GLOBAL:RoomType	free field
ListenerPosition	[0,0,0]
ListenerPosition:Type	'cartesian'
ListenerPosition:Units	'metre'
ReceiverPosition	[0,0.74,0;0,-0.74,0]
ReceiverPosition:Type	'cartesian'
ReceiverPosition:Units	'metre'
SourcePosition	*Each position defined*
SourcePosition:Type	'spherical'
SourcePosition:Units	'degree, degree, metre'
EmitterPosition	[0,0,0]
EmitterPosition:Type	'cartesian'
EmitterPosition:Units	'metre'
Data.SamplingRate	44100
Data.SamplingRate:Units	'hertz'
Data.Delay	[0,0]

Table 13: SOFA SimpleFreeField Convention and HUTUBS metadata

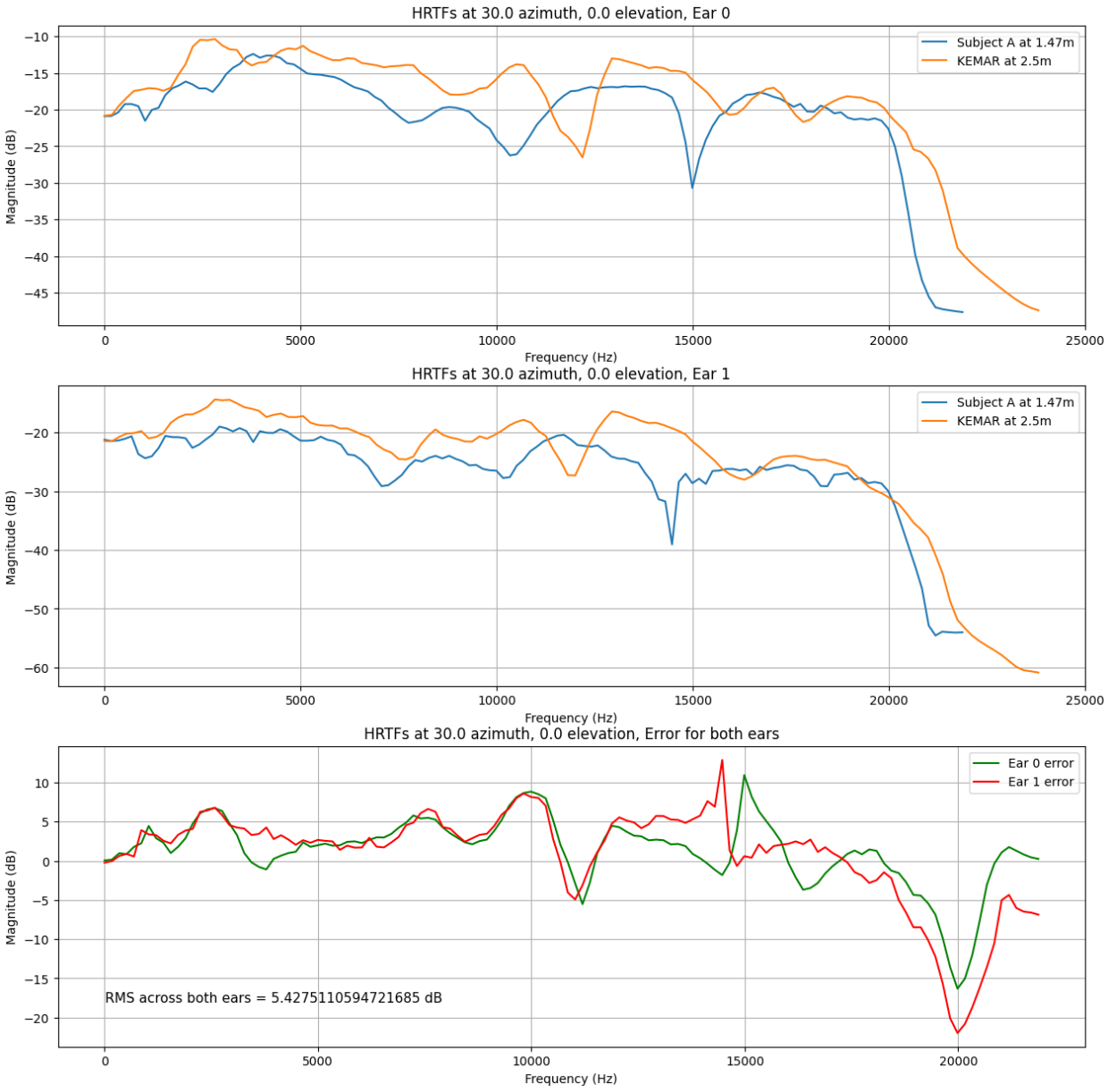


Figure 27: Comparing measured HRTFs between KEMAR and the physically closest subject

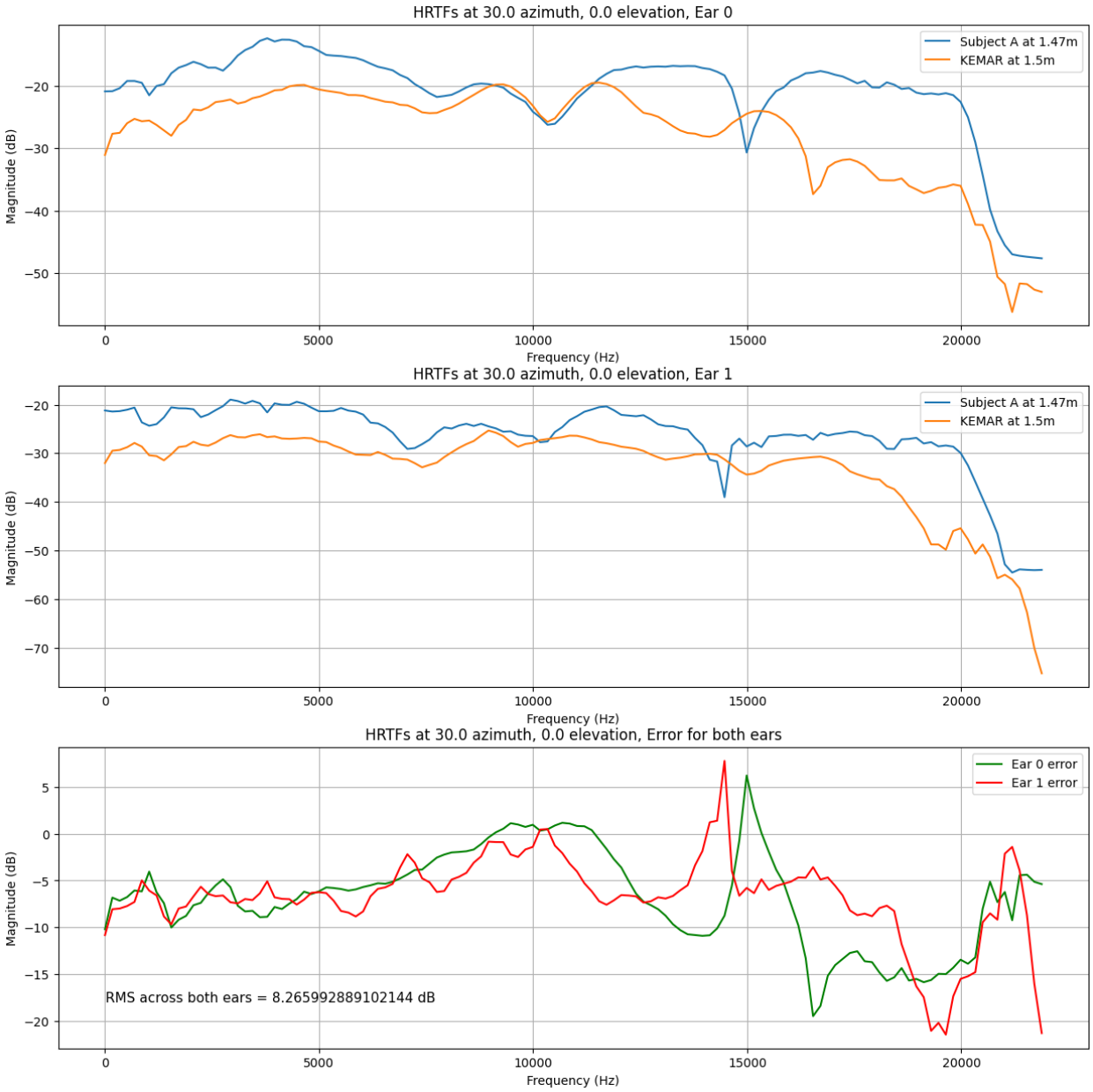


Figure 28: Comparing measured HRTFs between KEMAR at 1.5m and the physically closest subject

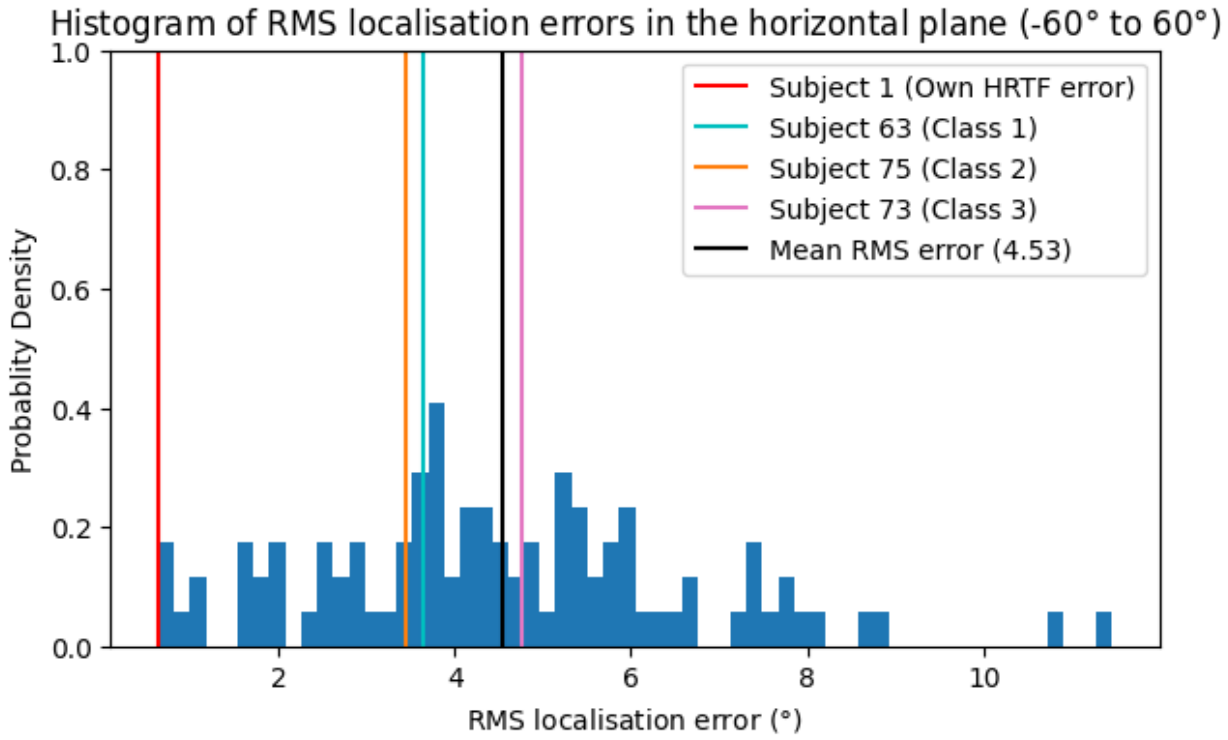


Figure 29: RMS localisation error of subject 1 using simulated HRTFs provided in the HUTUBS datasest