CUSTOMIZATION OF HRTF USING ANTHROPOMETRIC FEATURES

Yuxiang Wang
University of Rochester
ywang310@ur.edu

You Zhang
University of Rochester
you.zhang@rochester.edu

ABSTRACT

An HRTF(Head-related Transfer Function) describes how human ears receive a sound from a certain spatial direction. This transfer function is individual and direction dependent and is vital for virtual acoustic display. However, due to its uniqueness, using generic HRTF for virtual acoustic display may result in compromised results, leading to vague or misplaced acoustic images. Thus, it's meaningful to look into the customization of HRTF from a subject's physical appearance.In this project, we find certain relations from features in HRTF to the features of a subject's anthropometric details. And furthermore, by taking a few certain measurements such as taking a photo, we are be able to predict a new subject's HRTF.

1. INTRODUCTION

Head-related Transfer Function(HRTF) describes how human ears receive a sound from a certain spatial direction. [3]. Acoustic waves from the external sound source is filtered in a certain way by the physical shape of a listener's head and torso, which is described in HRTF. For hearing subject , the HRTF contains all the information he needs for creating a perceptual transparent acoustic scene. However, due to its uniqueness, using generic HRTF for virtual acoustic display may result in compromised results, as one's HRTF may differ significantly from the other.

So it makes an meaningful project to look into the customization of HRTF from a subject's physical appearance. In this project, we propose a practical approach for connecting the features of subjects' physical appearance, to their HRTFs' certain features. An existing HRTF database is used for our study, and some attempts were made to find such connections. In finalized version, we will provide a demo of generating one's HRTF using a picture of his/her ears.

2. PROPOSED METHODS

In this project, we made attempts to in this highly engineering oriented problem step by step:

First, we choose a proper HRTF database to look into. We shall choose a database that provides both subjects' HRTF and their corresponding anthropometric features. Second, we decide the features in the HRTF that is suitable to formulate relations to. Then, we make attempt to find connections from the subjects' physical features to the features in their HRTF. We try the connections for HRTF at one direction in the beginning, then to move on to the HRTFs at mid plane, then we try to generalize it to the HRTFs at all the directions.

2.1 HRTF database selection

There are various types of HRTF database, depending on their measuring grids and formatting. Three typical HRTF grids are shown here:

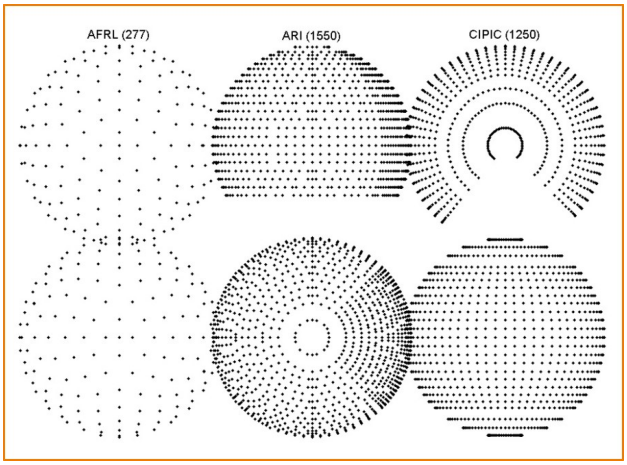

Figure 1. Typical HRTF databases

For the purpose of this project, we choose CIPIC database(on the right) [1], as it uses interpolar coordinate system, which is more than suitable for studying the connections between physical features to acoustic features, as it panned the source locations vertically. This database contains 45 subjects, and for each of the them, we have:

1) HRTF in time domain.

2) Corresponding anthropometric features. The standards for measuring subjects' anthropometric features were proposed by CIPIC group themselves, and it shown below in Figure 2.

We use some of these features after some preprocessing.

2.2 Pinna notch extraction

From various existing studies we learn that, not all the information contained in the HRTF are useful cues for sound localization [3]. In particularwe choose to use the so called pinna notches as the features of interest.

Human external ear functions as a filter to external sound, and results in crest and trough in the frequency domain.

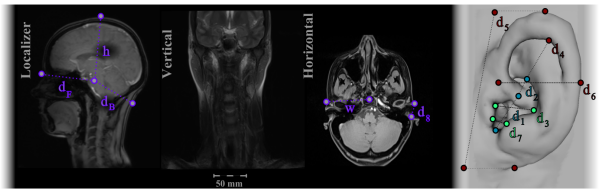

Figure 2. Standards for subjects' anthropometric features

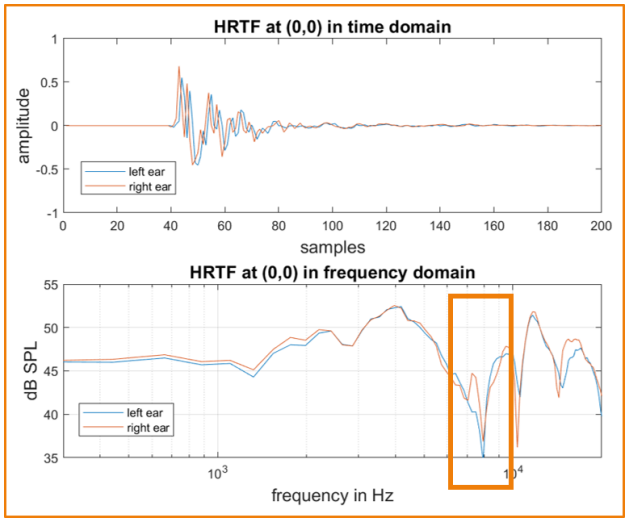

Figure 3. Example of pinna notches in HRTF at frontal direction

For human being, we are sensitive to these "notches" in the frequency [4] [3]. The location and depth of these notches will serve as important cues for localization, so it makes perfect sense to choose them as the feature we try to build connections to.

2.3 HRTF low frequency representation

From acoustic point of view, the HRTF in lower frequency domain can be regarded as the scatter ring pattern of two concatenated spheres, which simulates human head and torso. Some analytical derivation can be made for this step, as is shown in the so called, Snowman model:

For at certain point r, we solve the wave or Helmholtz equation to identify pressure, which is generated by a sinusoidal point source located at r with intensity Q_0:

$$\nabla'^2 P(\boldsymbol{r}', \boldsymbol{r}, f) + k^2 P(\boldsymbol{r}', \boldsymbol{r}, f) = -jk\rho_0 c Q_0 \delta(\boldsymbol{r}' - \boldsymbol{r}) \tag{1}$$

where k is the wave number; ρ is the density of air and c is the speed of sound; $\delta(r' - r)$ denotes the Dirac delta function. Only the pressures at the two field points that specify two ears on the head surface are of interest in HRTF calculation.

In HRTF calculation, therefore, the pressure at the ear can be equivalently obtained by calculating the pressure $P(r, r', f)$ at an arbitrary spatial position r:

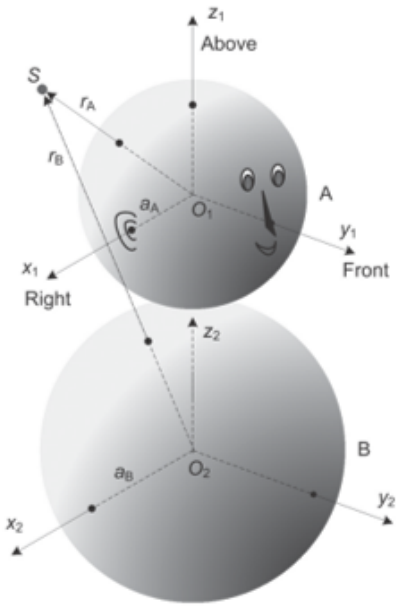

Figure 4. Snowman model for analytical calculation

$$\nabla^2 P(\boldsymbol{r}, \boldsymbol{r}', f) + k^2 P(\boldsymbol{r}, \boldsymbol{r}', f) = -jk\rho_0 c Q_0 \delta(\boldsymbol{r} - \boldsymbol{r}')$$

$$\lim_{r \to r} \left[\frac{\partial P(\boldsymbol{r}, \boldsymbol{r}', f)}{\partial r} + jkP(\boldsymbol{r}, \boldsymbol{r}', f) \right] = 0 \tag{2}$$

For the practical calculation in this project, we can write $P(r, r', f)$ in such way as [5]:

$$P_0(\boldsymbol{r}, \boldsymbol{r}', f) = j \frac{k\rho_0 c Q_0}{4\pi |\boldsymbol{r} - \boldsymbol{r}'|} \exp(-jk|\boldsymbol{r} - \boldsymbol{r}'|) \tag{3}$$

The scattered pressure $P_A(r, r', f)$ caused by spherical head A can be expanded as a series of complex-value spherical harmonic (SH) functions $Y_l m$, which can be written as:

$$P_A(\boldsymbol{r}, \boldsymbol{r}', f) = P_A(\boldsymbol{r}_A, \boldsymbol{r}'_A, f) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}^A h_l(kr_A) Y_{lm}(\Omega_A) \tag{4}$$

similarly, the scattering from the torso can be written as:

$$P_B(\boldsymbol{r}, \boldsymbol{r}', f) = P_B(\boldsymbol{r}_B, \boldsymbol{r}'_B, f) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l'} A_{lm}^B h_l(kr_B) Y_{lm'}(\Omega_B) \tag{5}$$

For the purpose of this project, we will not expand this part further. In the results part, we could combine this result as the lower frequency templates for a whole HRTF database.

2.4 Learning the mapping from anthropometric features to HRTF features

The anthropometric features include many measurements of parameters in human body and ears [2]. Some features may not be relevant to HRTF features. In our first step, Principle Components Analysis (PCA) is used to analyse the contribution of each feature and the correlation between features. Then, a model based on autoencoder is built to learn the mapping function from anthropometric features to HRTF features. The mapping can be used to generate higher frequency components of the HRTF impulse response. The learning procedure is illustrated in Figure 5

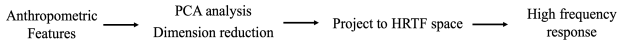

Figure 5. Proposed learning procedure

2.4.1 Dimension reduction for anthropometric features

In the CIPIC database that we use, 17 out of 37 anthropometric parameters are the features directly related to ears. They may not be independent since everyone's ears pattern are similar. In this section, we use PCA to analyse the importance of different anthropometric features. A threshold for the contribution is then set to decide which features are most useful for the mapping from anthropometric to HRTF. The dimension of data is then reduced according to the contribution, in order to feed into models for mapping features. The visualization of the three main components of PCA is shown in Figure 6.

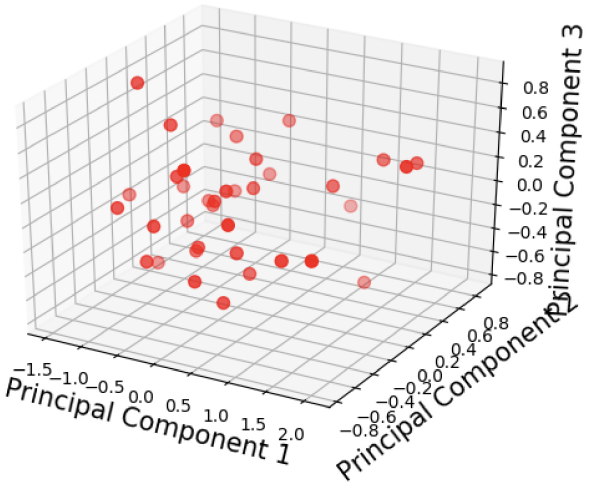

Figure 6. Three principal components from PCA

2.4.2 Model architecture

Inspired by autoencoder that can learn the latent feature representation to reconstruct the input data, we proposed a model based on similar architecture to learn the mapping from anthropometric features to HRTF impulse responses. The encoder learns the location and depth of different pinna notches and decoder is used to generate the higher frequency components for impulse responses. Since the number of pinna notches are at most four, we set eight parameters for the output of the encoder to model both the depth and location for each notch. The model architecture is shown in Figure 7.

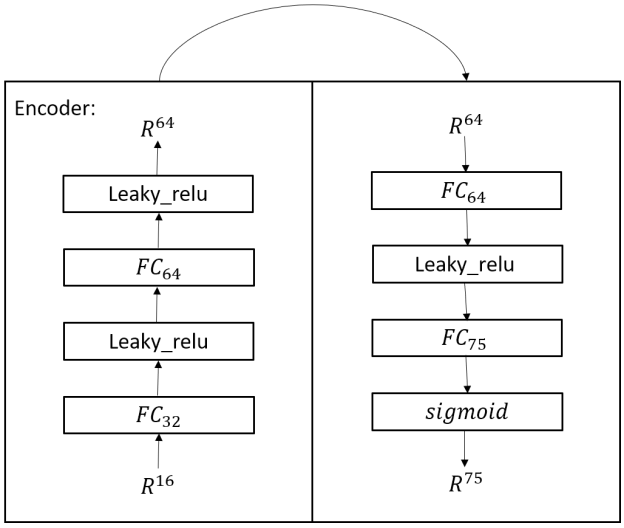

Figure 7. Model architecture for mapping anthropometric features

3. EXPERIMENT SETUP

3.1 Dataset augmentation

We use CIPIC database containing HRIRs of 45 subjects and their corresponding anthropometric measurements. 200 samples are provided for each time domain impulse response. Our proposed method is designed to learn the mapping function from anthropometric measurements to the frequency domain impulse response. We generate 600 samples for frequency domain impulse responses with each time domain samples. The last 200 samples are considered as higher frequency components of impulse response. The location and depth of pinna notches in these 200 samples vary from person to person. As a result, the higher frequency components are import for HRTF personalization.

3.2 Evaluation metric

At this step, we use the test set of the original database to perform the evaluation of our learned mapping, to avoid certain mistakes in our experiments.

We use the subjects' anthropometric datas as input, and try to generate our predicted pinna notch locations and depths for them. Then we compare our predicted results to the ground truth, and use the Euclidean distance as the quantified parameter for the evaluation. If the distance falls beyond certain threshold, we shall consider taking more adjustments to our setup.

4. PRELIMINARY RESULTS

4.1 Pinna notch extraction

Using CIPIC database, we've generated the aligned frequency layout in the midsagittal plane, where the pinna notches are represented in deep valleys.

Furthermore, with methods using the same principles of peak detection, we could generate our detected pinna notches, similar to the ones described in the textbooks:

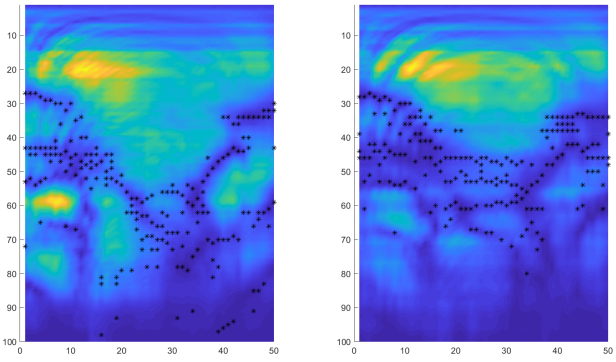

Figure 8. Pinna notch extraction of the HRTF layout in the midsagittal plane

With the pinna notch results for each subject in the database, we continued to perform experiments of the machine learning step.

4.2 Feature learning for the frontal direction HRTF

And here is the validation results on the validation set:

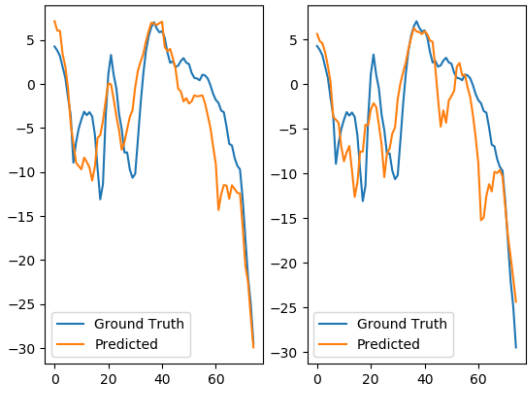

Figure 9. Comparison of the higher frequency components of HRTF frequency impulse responses (ground truth vs. predicted) (a) for left ear (b) for right ear

Comparing the predicted HRTF to the ground truth, we could see the overall trend of the higher frequency is corresponding quite well. The pinna notches are clearly presented in the predicted ones. Still, the detailed frequency shapes near the pinna notches is not ideal, and it's worth further tuning of the model.

Furthermore, we present here our preliminary results for the HRTF prediction. After training our algorithm on the CIPIC database, we measured Yuxiang's pinna parameters according to the CIPIC standards, and feed these parameters into the trained model.

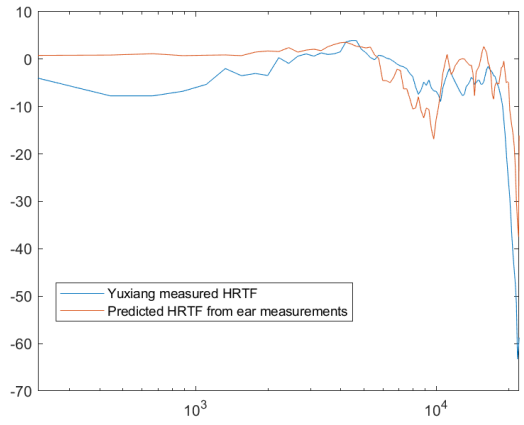

Figure 10. Comparison of Yuxiang's left ear HRTF, measured v. predicted

Comparing the full HRTF predicted v. ground truth, we find that the model is able to predict the overall trend of the HRTF. However the pinna notches locations and detailed shapes did have some mismatch. This may be due to the measurement error in Yuxiang's pinna shape as well as head size. Also, as Yuxiang's HRTF is not measured by CIPIC group, the resulting measurements may have some differences in certain frequency region due to the microphone used, which may be the cause for these mismatch.

4.3 Attempts to learning HRTF in the midsagittal plane

In this step, we try to extend our mapping from a single frontal location to the whole midsagittal plane.

This has been a bit more difficult since the data we are mapping to is much larger. But using the conclusions from the previous section, we can approximate the lower frequency part with Snowman model, and just implement the pinna notches using sample functions. After generating the lower frequency part, we simply add the pinna notches manually to simulate the HRTF appearance.

5. DISCUSSIONS

5.1 Conclusions

In this project, we make our novel attempts to predict one's HRTF from physical appearance, by building the connections from the subject's anthropometric data to the features in their corresponding HRTFs. This is a comprehensive research problem that combines the backgrounds of both acoustics and machine learning. We built our physics model for acoustic features in the lower frequencies, and designed our machine learning model to evaluate features in higher frequencies. By combining these, we have our

preliminary results for both the database validation set, as well as for an actual human being.

At current stage, we are able to extract some information in the frontal direction. By comparing our predicted result to the ground truth, we find that our model is able to deliver the overall trend of the higher frequency part of the HRTF, as well as some pinna notches information. Still, there are some problems for the model to predict the frequency details near the pinna notches. Also, the model is not stable enough to guarantee good matching for all subjects, especially when the subject's physical shape greatly differs from the normal level.

5.2 Future works

Our current deep learning model only contains fully connected layers that learns the mapping function between anthropometric parameters and high frequency impulses. The correlation in the locations of in the target pinna notches from different directions are not considered in the model. In future, it would be meaningful to develop a more powerful model for learning the pinna notch locations in the medium plane. And also it would be meaningful to regard the mid-plane pinna notch locations as a whole for extracting features, which may be useful features for the model to learn. Lastly, mapping all directions of HRTF impulse responses remains as an interesting future task and need further investigations.

6. REFERENCES

[1] V Ralph Algazi, Richard O Duda, Ramani Duraiswami, Nail A Gumerov, and Zhihui Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064, 2002.

[2] Pierre Guillon, Rozenn Nicol, and Laurent Simon. Head-related transfer functions reconstruction from sparse measurements considering a priori knowledge from database analysis: A pattern recognition approach. In *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[3] Abhijit Kulkarni and H Steven Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747, 1998.

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

[5] Dmitry N Zotkin, Ramani Duraiswami, and Nail A Gumerov. Regularized hrtf fitting using spherical harmonics. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 257–260. IEEE, 2009.