ESTIMATION OF PINNA NOTCH FREQUENCY FROM ANTHROPOMETRY: AN IMPROVED LINEAR MODEL BASED ON PRINCIPAL COMPONENT ANALYSIS AND FEATURE SELECTION

Riccardo Micciini
Aalborg University
rmicci18@student.aau.dk

Simone Spagnol
Aalborg University
ssp@create.aau.dk

ABSTRACT

In this paper, anthropometric data from a database of Head-Related Transfer Functions (HRTFs) is used to estimate the frequency of the first pinna notch in the frontal part of the median plane. Given the presence of high correlations between some of the anthropometric features, as well as repeated values for the same subject observations, we propose the introduction of Principal Component Analysis (PCA) to project the features onto a space where they are more separated. We then construct a regression model employing forward step-wise feature selection to choose the principal components most capable of predicting notch frequencies. Our results show that by using a linear regression model with as few as three principal components, we can predict notch frequencies with a cross-validation mean absolute error of just about 600 Hz.

1. INTRODUCTION

Binaural sound rendering can be achieved by incorporating the acoustic effects of the human head into a given sound, so as to simulate the pressure at the entrance of the ear canals. The set of functions used to perform this are called Head-Related Transfer Functions (HRTFs), and consist of digital filters characterizing sounds coming from a specific point in space. Unfortunately, obtaining personal HRTFs is only possible with expensive equipment and invasive recording procedures. For this reason, non-individual HRTFs are often preferred in practice, with the drawback of being prone to systematic localization errors such as front/back reversals, wrong elevation perception, and inside-the-head localization [1].

The most relevant differences between the HRTFs of two subjects are due to the different shapes, sizes, and orientations of the pinnae. The pinna has a key role in shaping HRTFs because of the reflections and resonances occurring in its rims and cavities, which can be seen in the HRTF as sequences of notches and peaks, respectively. The spectral location of peaks and notches represents a pivotal cue to the characterization of the sound source's spatial position, and in particular of its elevation. Despite the availability of various recent research works targeted at predicting the HRTF or some of its features from pinna anthropometry (see for instance [2–4]), we are still far from a complete understanding of the underlying relationships.

The relationship between the center frequencies of the three main pinna notches (known as N_1, N_2, and N_3) in a set of frontal median-plane HRTFs and 13 different anthropometric features of the pinna was explored in [5] with linear regression models. The anthropometric feature set included global pinna measurements (e.g. pinna height, concha width) as well as measurements that vary with the elevation angle of the sound source (distances between the ear canal and pinna edges). The results of that work showed that while the considered features are not able to approximate with sufficient accuracy neither the N_2 nor the N_3 frequency, eight of them are sufficient for modeling the frequency of N_1 within an acceptable margin of error, and that distances between the ear canal and the outer helix border are the most important features for predicting N_1.

In this work, we take a step forward by further investigating the model presented in [5] on the same input data and considering linear transformations and selection within the feature space in order to improve N_1 prediction. Given the presence of high correlation among features as well as repeated anthropometric parameters for each record pertaining a certain subject, we propose the introduction of Principal Component Analysis (PCA) to project the features onto a space where they are more separated. Subsequently, we apply feature selection on the regression model in order to preserve the components with higher predictive power, thereby reducing overfitting.

2. METHODS

2.1 HRTF feature extraction

The raw dataset consists of measured Head-Related Impulse Responses (HRIRs) for the 33 subjects from the CIPIC database [6] for which full anthropometric data (records and single-ear pictures) is available. We consider HRIRs measured in the frontal half of the median plane, with elevation ranging from $\phi = -45°$ to $\phi = 45°$ at 5.625-degree steps (17 HRIRs per subject). Elevations higher than 45° were discarded because of the general lack of spectral notches in the corresponding HRTFs [7].

Pinna notch frequencies in each HRIR are extracted with the *ad-hoc* signal processing algorithm by Raykar *et al.* [8].

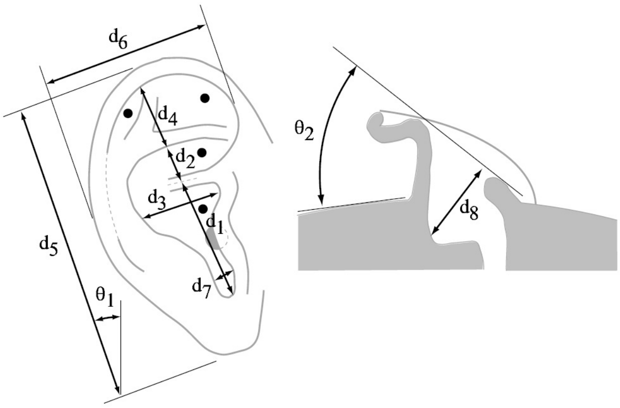

Figure 1. The 10 pinna parameters included in the CIPIC database (figure reproduced from [6]).

Then, for each available elevation, the extracted notches are grouped in frequency tracks along adjacent elevations with the McAulay-Quatieri partial tracking algorithm [9], with a matching interval of $\Delta = 1\,\text{kHz}$ [10]. When available, the three longest tracks are labeled as N_1, N_2, and N_3 in increasing order of average frequency; if a subject lacks a notch track, labels are assigned according to the closest notch track frequency median [5]. Given the previously reported low correlation between anthropometric parameters and the two notches N_2 and N_3, we focus on N_1 as prediction target, for which we have a total of 367 different observations belonging to 29 different subjects.

2.2 Anthropometric feature extraction

In addition to the 10 global pinna features contained in the CIPIC database and reported in Fig. 1, we extract 3 elevation-dependent features from scaled individual pinna images according to the following ray-tracing procedure. The three contours corresponding to the outer helix border, the inner helix border, and the concha border/antitragus (C_1, C_2, and C_3 respectively) are traced by hand and stored as sequences of pixels. Then, as can be seen in Fig. 2, the point of maximum protrusion of the tragus is chosen as the reference ear-canal point for the computation of distances. For each elevation $\phi \in [-45, 45]$, we compute distances in centimeters between the reference point and the point intersecting each pinna contour along the ray originating from the reference point with slope $-\phi$, and store them as $r_k(\phi)$, where $k \in \{1, 2, 3\}$ refers to contour C_k.

We assume that the $r_k(\phi)$ features are, together with the 10 individual global pinna features, good predictors of the N_1 frequency in the HRTF measured at elevation ϕ. This assumption is due to the results of a previous work [11] that highlighted a qualitatively reciprocal linear relationship between distance from the ear canal to the hypothesized pinna reflection points and pinna notch frequencies.

2.3 Dimensionality reduction

The so created dataset is composed of 13 features (d_i, $i = 1 \ldots 8$, θ_j, $j = 1 \ldots 2$, $r_k(\phi)$, $k = 1 \ldots 3$, $\phi \in [-45, 45]$) and 367 observations for N_1. As the focus of this work is on anthropometric features, the elevation angle ϕ is not considered as regressor. However, since our sample comprises only 29 unique subjects, the global pinna features — which do not depend on elevation — are repeated for each specimen. Moreover, the features are mutually correlated, with an average Pearson correlation coefficient of 0.24 and a maximum of 0.95 across the 78 feature pairs.

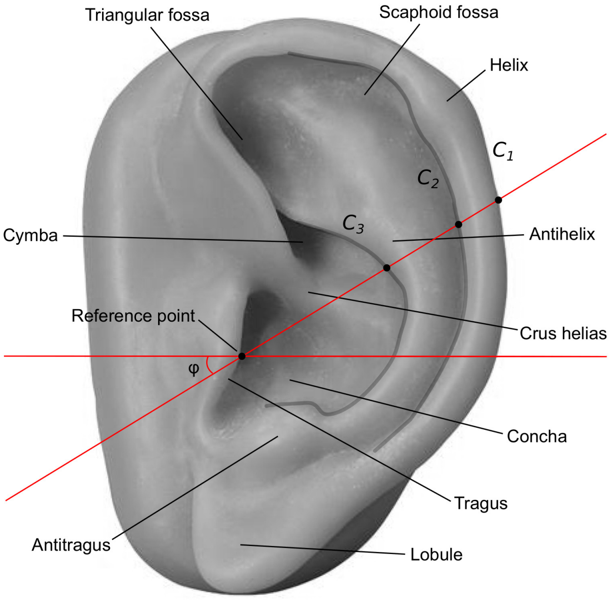

Figure 2. Pinna anatomy and extraction of the three elevation-dependent features.

In order to untangle the data and reduce its dimensionality, we apply a PCA to its features. This technique is used to find a new orthogonal coordinate system in the original data space, which best represents the variance expressed by the data. We therefore obtain 13 new features, called *principal components*, which are largely uncorrelated (average Pearson coefficient of 1.9e−16) and ordered by decreasing eigenvalue. It is important to point out that the original features have been preemptively normalized into 0-mean and unit variance, so as to avoid features with large magnitudes dominating the results.

2.4 Regression model

Finally, multiple linear regression with forward step-wise feature selection is performed on the principal components using all the 367 data records. The feature selection step improves the generalization capabilities of the model by discarding predictors which are irrelevant and may instead cause overfitting. The procedure consists in instantiating a regression model for each of the available predictors, evaluating their performances using the most appropriate metric, then selecting the predictor resulting in the best performances and repeating the previous steps with models composed of the previously selected features along with any of the remaining ones, until the desired number of predictors is reached. In this case, we settled on 3 principal components, which is the minimum amount required to explain

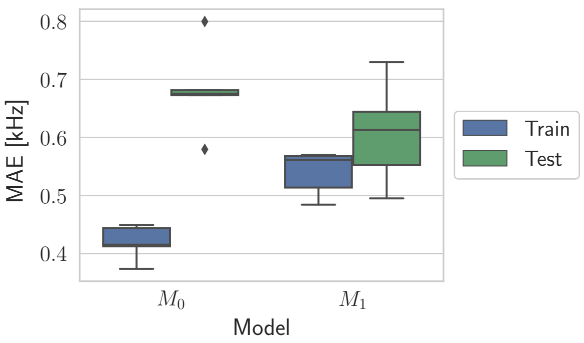

Figure 3. Box-and-whisker plot showing the mean absolute error (expressed in kHz) of models M_0 (baseline) and M_1, for training and test sets respectively.

more than 50 % of the variance in the data.

The metric used to determine the best-performing predictors is the *mean absolute error*, calculated as

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j| \qquad (1)$$

where n is the number of observations, y_j is the true value, and $\hat{y}_j$ is the predicted value. This metric was preferred over root-mean-square error because it provides an intuitive representation of the average residual, and because it is robust to outliers and large errors.

The resulting model, M_1, is validated using a 5-fold cross-validation scheme. The pool of 29 subjects is divided into 5 approximately equal-sized subsets. During each iteration, one of the subsets is set aside and used to validate the model, whereas the remaining ones constitute the training set. Therefore, the ratio between training and test data is approximately 20 %, depending on the exact number of observations available per subject. While this scheme does not guarantee a constant training-test ratio, it ensures that no test subject appears in the training set, which would otherwise greatly simplify the prediction task.

The model M_1 described above is then compared against the baseline M_0, a multiple linear regression model comprising all the original features.

3. RESULTS

Figure 3 shows the performances of models M_0 (baseline) and M_1 in terms of their mean absolute error, aggregated over all the cross-validation folds. It is interesting to notice how the baseline model performs better than the custom one on the training set (average MAE equal to 419 Hz for M_0 and 539 Hz for M_1), while the opposite is true for validation data (average MAE equal to 682 Hz for M_0 and 607 Hz for M_1, representing an 11 % average decrease in error). This means that some of the variance expressed by the anthropometric features is not useful for generalizing on unseen data. Therefore, our feature selection process renders M_1 more resilient to overfitting.

Figure 4 shows two example instances of true and predicted notch frequency from the two models under consideration, computed during the validation step. Despite both models being capable of modeling the mostly monotonic relation between frequency and elevation (and by proxy, $r_k(\phi)$ features), the baseline model presents a larger bias, a clear consequence of the aforementioned overfitting.

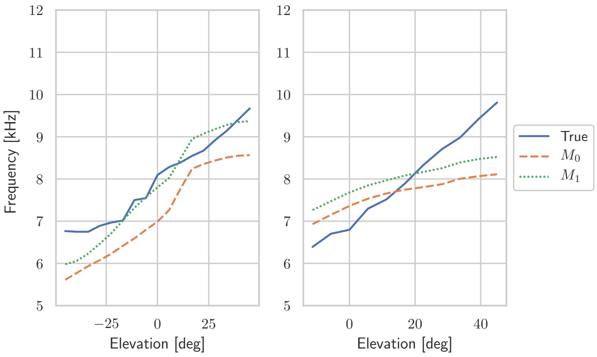

Figure 4. Sample plots of notch frequency over elevation, with both true and predicted values, taken from two subjects in the test set.

PC_i	PC_1	PC_2	PC_3	PC_4	PC_5
ρ	-0.43	0.50	0.13	-0.43	-0.13

Table 1. Pearson correlation coefficients between each of the first 5 principal components and target frequencies.

We also found that principal components PC_i with $i \in \{1, 2, 4\}$ are consistently selected across folds, revealing their predictive potential. Despite explaining 11 % of the data variance on average, PC_3 is never selected as a predictor; while this may seem counterintuitive, it can be explained by looking at the Pearson correlation coefficients between said principal components and target notch frequencies, as shown in Table 1. In this case, it is clear how PC_3 does not manifest enough correlation for it to positively impact the performances of the model. In terms of the role of the selected principal components, the matrix of loadings reveals how the component with the most predictive power mainly codes elevation-dependent features $r_k(\phi)$, whereas the second and third ones present a mixture of elevation-dependent and global features.

When evaluating the performances of the models in terms of their psychoacoustical implications, it is desirable to consider whether the predicted notch frequency lies within 10 % of the real one. Indeed, for spectral notches in the high-frequency range, differences lower than said threshold are, on average, indistinguishable [12]. Since every observation is used once and only once throughout cross-validation, it is possible to determine the percentage of *psychoacoustically valid* predictions by counting how many fall within the threshold, and normalizing by the overall number of observations. Therefore, when considering test data only, the percentage of psychoacoustically valid predictions for the baseline and the custom models is 56.3 % and 61.2 % respectively, constituting a modest 8.75 % improvement in perceptually noticeable performances.

4. CONCLUSIONS

The results of this work show that N_1 frequencies can be predicted from anthropometric data within a certain degree of accuracy. However, our regression model was built using a limited amount of training data from a single HRTF database. More recent and documented databases such as HUTUBS [13] will be used in future works in order to carry out larger data analyses, possibly using state-of-the-art feature extraction and nonlinear regression algorithms.

It has to be noted that HRTF data collected on a human population implies issues related to microphone position and head movements that pose critical challenges when merging different datasets. The authors will shortly expand the recently collected Viking HRTF dataset [14], designed to guarantee reproducible measurements on a mannequin with different interchangeable ears, through new acquisitions on a larger ear sample in a controlled environment. These measurements will serve as a solid basis for accurate investigations on the relation between HRTFs and anthropometric data, the final objective being an effective tuning of low-order structural HRTF models [11,15,16]. Applications of these models are expected to range from personal entertainment to assistive technologies [17, 18].

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 797850, and from NordForsk's Nordic University Hubs programme under grant agreement No. 86892.

5. REFERENCES

[1] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

[2] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 317–333, 2014.

[3] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 690–701, 2015.

[4] K. Iida, H. Shimazaki, and M. Oota, "Generation of the amplitude spectra of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener's pinnae," *Appl. Acoust.*, vol. 155, pp. 280–285, 2019.

[5] S. Spagnol and F. Avanzini, "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model," in *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, Trondheim, Norway, 2015, pp. 231–236.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, 2001, pp. 1–4.

[7] S. Spagnol, M. Hiipakka, and V. Pulkki, "A single-azimuth pinna-related transfer function database," in *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, Paris, France, 2011, pp. 209–212.

[8] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, 2005.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[10] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. EL58–EL64, 2015.

[11] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 508–519, 2013.

[12] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 820–836, 1989.

[13] F. Brinkmann, M. Dinakaran, R. Pelzer, J. J. Wohlgemuth, F. Seipl, and S. Weinzierl, "The HUTUBS HRTF database," 2019, DOI: 10.14279/depositonce-8487.

[14] S. Spagnol, K. B. Purkhús, S. K. Björnsson, and R. Unnthórsson, "The Viking HRTF dataset," in *Proc. 16th Int. Conf. Sound and Music Computing (SMC 2019)*, Malaga, Spain, 2019, pp. 55–60.

[15] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, 1998.

[16] S. Spagnol, E. Tavazzi, and F. Avanzini, "Distance rendering and perception of nearby virtual sound sources with a near-field filter model," *Appl. Acoust.*, vol. 115, pp. 61–73, 2017.

[17] F. Avanzini, S. Spagnol, A. Rodá, and A. De Götzen, "Designing interactive sound for motor rehabilitation tasks," in *Sonic Interaction Design*, K. Franinovic and S. Serafin, Eds. Cambridge, MA, USA: MIT Press, 2013, ch. 12, pp. 273–283.

[18] S. Spagnol, G. Wersényi, M. Bujacz, O. Balan, M. Herrera Martínez, A. Moldoveanu, and R. Unnthórsson, "Current use and future perspectives of spatial audio technologies in electronic travel aids," *Wireless Comm. Mob. Comput.*, vol. 2018, p. 17 pp., 2018.