HRTF PHASE SYNTHESIS VIA SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES

Ivan Tashev

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
ivantash@microsoft.com

ABSTRACT

We propose a method for the synthesis of the phases of Head-Related Transfer Functions (HRTFs) using a sparse representation of anthropometric features. Our approach treats the HRTF synthesis problem as finding a sparse representation of the subjects anthropometric features w.r.t. the anthropometric features in the training set. The fundamental assumption is that the group delay of a given HRTF set can be described by the same sparse combination as the anthropometric data. Thus, we learn a sparse vector that represents the subjects anthropometric features as a linear superposition of the anthropometric features of a small subset of subjects from the training data. Then, we apply the same sparse vector directly on the HRTF group delay data. For evaluation purpose we use a new dataset, containing both anthropometric features and HRTFs. We compare the proposed sparse representation based approach with ridge regression and with the data of a manikin (which was designed based on average anthropometric data), and we simulate the best and the worst possible classifiers to select one of the HRTFs from the dataset. For objective evaluation we use the mean square error of the group delay scaling factor. Experiments show that our sparse representation outperforms all other evaluated techniques, and that the synthesized HRTFs are almost as good as the best possible HRTF classifier.

Index Terms— Head-related Transfer Function, HRTF Personalization, HRTF Synthesis, Sparse Representation, Anthropometric Features

I. INTRODUCTION

Head-related transfer functions (HRTFs) represent the acoustic transfer function from a sound source position to the entrance of the blocked ear canal of a human subject [1]. HRTFs are typically measured under anechoic conditions at a sufficient distance and describe the complex frequency response as a function of the sound source position (i.e. azimuth and elevation). Imposing HRTFs onto a non-spatial audio signal and playing back the result over headphones allows for positioning virtual sound sources at arbitrary locations. There are many potential applications of HRTFs, such as 3D audio for games, live streaming of events, music performances, virtual reality, training, and entertainment.

Since the measurement of HRTFs requires specialized equipment, the automatic personalization (selection or synthesis) of the listener's HRTFs based on a limited dataset is desirable whereby measuring a small set of anthropometric features of a given subject might be tolerable. Many techniques have been recently proposed for HRTF personalization [2], [3], [4], [5], [6], [7], [8], [9], [10] based on a selected set of anthropometric features. Their effectiveness heavily depends on the choice of anthropometric features. For this purpose, most of the existing techniques try to find linear relationships between anthropometric features and HRTFs. Other techniques try to find simple, approximated, non-linear relationships. Feature selection is still an open issue as it has been shown to be an NP-hard problem.

In our previous work [11] we proposed a method for synthesis of HRTF magnitudes using sparse representation. The main idea of this approach is to treat the synthesis of the HRTF magnitudes as finding a sparse representation of the subject's anthropometric features as a linear superposition of the anthropometric features of a small subset of subjects from the training data. We assume that the HRTF data is in the same relation as these anthropometric features. Then, we apply the same sparse vector on the HRTF magnitudes to synthesize the subject's HRTFs magnitude response. In this paper we extend the same approach for synthesis of the HRTF phases, using the averaged group delay as function of the direction and elevation.

To ensure that we employ an extensive set of features, we created a new dataset with an extended amount of anthropometric features compared to the existing literature [4], [12]. The remainder of the paper is organized as follows. Section II presents the collected dataset, while in section III we describe our approach for modeling of the HRTF phases. In section IV, we describe briefly our sparse representation based approach. In section V, we present experimental results. Finally, we conclude in Section VI.

II. DATA COLLECTION

We created a new dataset for the presented study that consists of measured HRTFs and 47 anthropometric features of 104 subjects with an age range from 13 to 62 years (age mean of 34). More details on the dataset can be found

in [11].

II-A. HRTF Representation

The HRTFs for each subject are represented as a set of frequency domain filters in pairs, one for the left and one for the right ear. The sampling rate is 48 kHz and each of the filters contains 512 taps (from 0 Hz to 24 kHz). The directions grid contains 512 directions spread on the entire sphere.

II-B. Anthropometric Features

The anthropometric features can be grouped into four categories: ear-related features, head-related features, limbs and full body features, and other features (gender, race, age, *etc*.). These four groups were obtained in three ways: direct measurements, questionnaire, and automatic deduction from 3D scans of the subject's head. Most of the ear- and head-related anthropometric features are obtained through the latter method.

The collected anthropometric features are superset of the CIPIC HRTF Database [12], but in this study we use the listed in Table I subset of the available in the database anthropometric features.

Table I. List of used anthropometric features.

Head-related features:
Head height, width, and depth;
Neck height, width, depth, and circumference;
Distance between eyes / distance between ears;
Maximum head width (including ears);
Inter-pupillary distance.
Limbs and full body features:
Shoulder width, depth, and circumference;
Torso height, width, depth, and circumference;
Distances: foot– knee; knee– hip; elbow– wrist; wrist– fingertip;
Height.
Other features:
Gender; age range; age; race;
Hair color; eye color; weight; shirt size; shoe size.

III. HRTF PHASE MODELING

A typical HRTF phase response for one direction is shown in Fig. 1. The phase responses pretty much linearly depend on the frequency, which also leads to a linear phase difference, as it is shown in Fig. 2. Note that in this figure the frequency is in linear scale. There is less reliable phase estimation in the very low part of the frequency band, and in the upper frequencies the phase response is affected by the features of the pinna. Earlier studies also show that the HRTF phase response is mostly linear [13] and that listeners are insensitive to the details of the interaural phase spectrum as long as the interaural time delay (ITD) of the combined low-frequency part of the waveform is maintained [14]. This

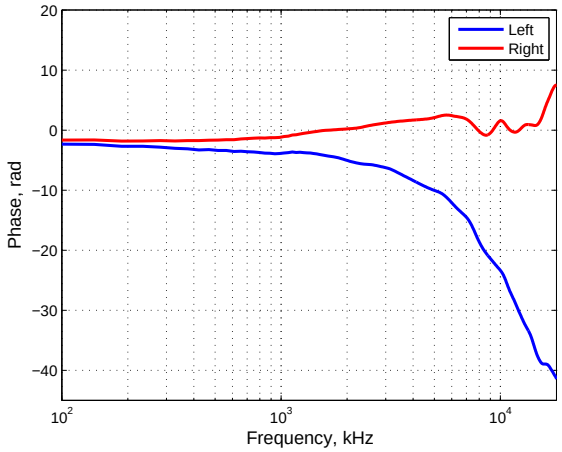

Fig. 1. HRTF phase response for subject n=5, direction $-40°$ and elevation $0°$.

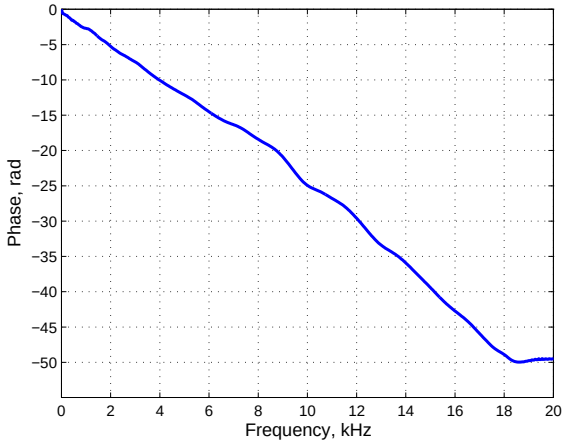

Fig. 2. HRTF left-right phase difference for subject n=5, direction $-40°$ and elevation $0°$.

is why we model the phase response of the subject HRTFs as a time delay, dependent on the direction and elevation. Fig. 3 shows the time delay for the same subject and direction and its interpolation as a constant delay. This represents a linear phase response, and the time delay can be measured as linear interpolation of the measured HRTF phases in the medium frequency range $500-1500$ Hz, where it also is most reliably measured.

We go one step further and make the assumption that the ITD as function of the direction and elevation has similar shape across all human subjects [15]. The only differences is in the scaling factor, which depends on the anthropometric features, mostly size of the head and the position of the ears. The average ITD of the properly scaled individual ITDs of the 104 subjects in the data set is shown in Fig. 4. Then the only individual feature of the HRTF phase response is the scaling factor - the number we have to multiply the average

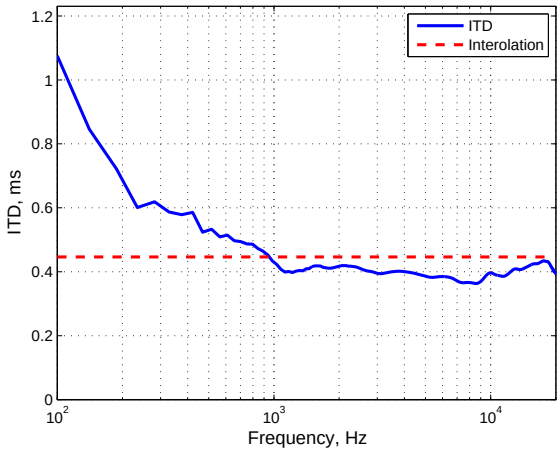

Fig. 3. Iteraural time delay for subject n=5, direction $-40°$ and elevation $0°$.

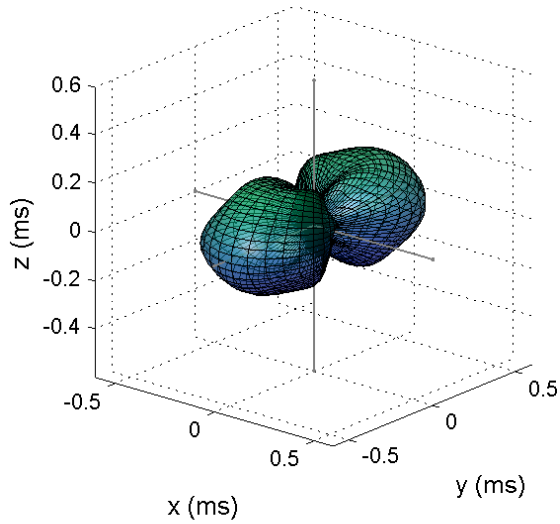

Fig. 4. Average ITD contour as function of the direction and elevation.

ITD to fit it into the individual ITD. This converts the problem of personalization of the HRTF phases to learning a single scaling factor as function of the anthropometric features.

IV. PROPOSED APPROACH

IV-A. Training Data Representation

Let assume that we have N subjects in the training set.

ITD scaling factors. The HRTF phases for each subject are described by a single ITD scaling factor for the average group delay. The ITD scaling factors for all persons in the dataset are stacked in a vector $\mathbf{H} \in \mathbb{R}^N$, so the value H_n corresponds to the scaling factor of the n-th person.

Anthropometric features. In the preparation stage all of the categorical features (hair color, race, eye color) are converted to binary indicator variables. For the rest of the anthropometric features a min-max normalization is applied to each of the features separately to make the feature values more uniform. Each person is described by A anthropometric features and can be viewed as a point in the space $[0,1]^A$. All anthropometric features of the training set are arranged in a matrix $\mathbf{X} \in [0,1]^{N \times A}$, where one row of $\mathbf{X}$ represents all the features of one person.

IV-B. Sparse Representation for ITD scaling factors

We propose to estimate the ITD scaling factor for a new subject given its anthropometric features $\mathbf{y} \in [0,1]^A$. The main idea is to treat the scaling factor estimation problem as finding a sparse representation of the subject's anthropometric features, with the assumption that the scaling factors are in the same relation. We also assume that our training set is sufficient to span a new person's anthropometric features. We learn a sparse vector $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_N]^T$ that represents the subject's anthropometric features as a linear superposition of the anthropometric features from the training data ($\hat{\mathbf{y}} = \boldsymbol{\beta}^T \mathbf{X}$), and then apply the same sparse vector directly on the scaling vector $\mathbf{H}$. We can write this task as a minimization problem, for a non-negative shrinking parameter λ:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{a=1}^{A} \left(y_a - \sum_{n=1}^{N} \beta_n X_{n,a} \right)^2 + \lambda \sum_{n=1}^{N} |\beta_n| \right) \tag{1}$$

The first part of the above equation is minimizing the differences between values of $\mathbf{y}$ and the new representation of $\hat{\mathbf{y}}$. Note that the sparse vector $\boldsymbol{\beta} \in \mathbb{R}^N$ provides one weight value per person (and not per anthropometric feature). The second part of the above equation is the ℓ_1 norm regularization term that imposes the sparsity constraints, and makes the vector $\boldsymbol{\beta}$ sparse. The shrinking parameter λ in the regularization term controls the sparsity level of the model and the amount of the regularization. It will be discussed further in Section IV-D.

We assume that the ITD scaling factors are represented by the same relation as the anthropometric features. Therefore, once we learn the sparse vector $\boldsymbol{\beta}$ from the anthropometric features, we directly apply it to the ITD scaling factors vector and the subject's ITD scaling factor value $\hat{H}$ is estimated as:

$$\hat{H} = \sum_{n=1}^{N} \beta_n H_n. \tag{2}$$

IV-C. ITD Scaling Factor Metric

To determine the accuracy of the estimated ITD scaling factors, we compare them with the true (measured) ITD scaling factor of the subject under consideration. For objective

evaluation we use the root mean square error (RMSE):

$$\epsilon = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{H}_n - H_n\right)^2} \quad (3)$$

where $\hat{H}_n$ is the estimated scaling factor for the n-th subject and H_n is the measured scaling factor for the same subject. Note that the perceptual meaning of this metric is unclear.

IV-D. Regularization Parameter λ

In the preparation stage we tune the nonnegative regularization parameter λ to prevent over-fitting using the leave-one person-out cross-validation approach [16], [17]. We take out each person from the dataset and estimate the sparse weighting vector $\boldsymbol{\beta}$ using the equation (1) for a series of λ values. Then we select the value of λ which gives minimal error according to equation (3). This process is repeated for all persons and the optimal λ for the dataset is computed as mean of the optimal values for each person.

IV-E. Computing the scaling factor

Let assume that we have a training dataset, as described in Sec. IV-A, and an optimal value for the regularization parameter λ, computed using the procedure in Sec. IV-D. Then, given vector $\mathbf{y}$ of the anthropometric features of a person with unknown ITD scaling factor, we can compute the sparse weighting vector $\boldsymbol{\beta}$ using the equation (1) with the optimal for this training set value of λ. The computed sparse vector then is used to estimate the person's ITD scaling factor $\hat{H}$ according to (2). The computed scaling factor multiplies the average ITD and we have estimated the time delay as function of the direction and elevation for this person. Converting the time delay to phase response for the left and the right ears is trivial.

V. EXPERIMENTS

V-A. Evaluation Protocol

To evaluate the accuracy of the proposed approach, we used the same leave-one-person-out cross-validation approach as in Sec. IV-D. We sequentially used the data for one person for testing and treated the remaining data of $N-1$ people as a training set. Before each use the training set went trough the procedure described in the same Sec. IV-D for determining the optimal value of the regularization parameter λ. Then we computed the ITD scaling factor for the test person as described in Sec. IV-E. After evaluating in this way all of the persons in the dataset the error is estimated according to (3).

V-B. Baselines

To assess how well our technique performs we established several baselines.

"The Best" and "The Worst" Classifiers. To create reference results, we simulate the best and the worst possible classifiers. We follow the proposed evaluation protocol and for each subject we find the nearest and farthest ITD scaling factor from the training set.

Ridge Regression. We also compare our approach with the ridge regression model [16], [18], [19], where the ℓ_1 norm regularization term is replaced with the ℓ_2 norm regularization term. This is a similar minimization problem, for a non-negative parameter λ:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{a=1}^{A} \left(y_a - \sum_{n=1}^{N} \beta_n X_{n,a} \right)^2 + \lambda \sum_{n=1}^{N} \beta_n^2 \right) \quad (4)$$

where the shrinkage parameter λ controls the size of the coefficients and the amount of the regularization, and it is optimized as explained in the Section IV-D.

HATS. We also use as reference the ITD scaling factor measured from the Brüel & Kjær's Head and Torso Simulator (HATS). The HATS is a manikin that is designed based on average anthropometric features.

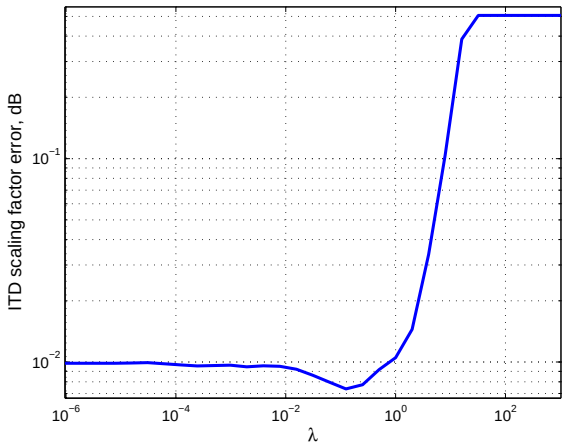

Fig. 5. ITD scalling factor estimation error as function of λ.

V-C. Results

The experimental results are presented in Table II. The proposed sparse representation based approach outperforms all other evaluated techniques. It obtains low RMSE, which is often close to the RMSE of the best classifier. The ridge regression model shows worse than the sparse representation results, which confirms the importance of sparsity in our approach. The ITD scaling estimation error of the HATS show RMSEs higher than the sparse representation model and close to the worst classifier, which justifies the HRTF personalization.

Fig. 5 shows the ITD scaling factor estimation error as function of the regularization parameter λ. On the left, where the values of λ are small, we have less sparse representation of the test subject. When the λ value increases - the sparsity also increases and the ITD scaling factor estimation error

Table II. ITD scaling factor estimation error

The Best Classifier	Sparse Representation	Ridge Regression	HATS	The Worst Classifier
0.005178	0.08338	0.0942	0.1408	0.277

decreases. We have well a defined minimum. The effect of the spare representation in this case is 12% relative reduction of the estimation error. Note the logarithmic vertical scale.

VI. CONCLUSIONS

We proposed a method for HRTF phase frequency response synthesis using anthropometric features and sparse representation. The phase frequency response is modeled as average ITD, function of the direction and elevation, scaled accordingly to the subject's personal ITD, which depends on the subject's anthropometric features. The anthropometric features of a given subject are presented as a sparse linear combination of the anthropometric features of the subjects in the dataset, and then the same relation is used to estimate the ITD scaling factor and thereby synthesize a personalized set of HRTF phases. The root mean square error confirms the effectiveness of the sparse representation based approach. Our method shows lower error than all other evaluated techniques and obtains results closest to the best classifier (i.e. the nearest ITD scaling factor in the training set).

Future work includes combining the magnitude and phase estimation methods, determining a perceptually motivated distance measure, and validating the synthesized HRTFs in a perceptual experiment.

VII. REFERENCES

[1] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, revised edition, 1996.

[2] Hongmei Hu, Lin Zhou, Hao Ma, and Zhenyang Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, 2008.

[3] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in *ICASSP*, 2013.

[4] Dmitry N. Zotkin, Jane Hwang, Ramani Duraiswami, and Larry S. Davis, "HRTF personalization using anthropometric measurements," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[5] A. Mohan, R. Duraiswami, D. N. Zotkin, D. DeMenthon, and L. S. Davis, "Using computer vision to generate customized spatial audio," in *IEEE International Conference on Multimedia and Expo*, 2003.

[6] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Transactions on Audio, Speech & Language Processing*, 2013.

[7] D. Schonstein and B. F. G. Katz, "HRTF selection for binaural synthesis from a database using morphological parameters," in *20th International Congress on Acoustics (ICA)*, 2010.

[8] Z. Haraszy, D.-G. Cristea, V. Tiponut, and T. Slavici, "Improved head related transfer function generation and testing for acoustic virtual reality development," in *14th WSEAS international conference on Systems: part of the 14th WSEAS CSCC multiconference - Volume II*, 2010.

[9] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous HRTF datasets," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[10] W. W. Hugeng and D. Gunawan, "Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements," in *CoRR*, 2010.

[11] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *ICASSP*, 2014.

[12] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano, "The CIPIC HRTF database," 2001.

[13] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *Journal of Acoustical Society of America*, vol. 61, pp. 1567–1576, 1977.

[14] A. Kulkarni, S. K. Isabelle, and H.S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America*, vol. 105, pp. 2821–2840, 1999.

[15] Jens Ahrens, "HRTF ITD calculation and scaling," Tech. Rep., Microsoft Reseacrh, 2012.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2008.

[17] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 1995.

[18] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 1970.

[19] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications for nonorthogonal problems," *Technometrics*, 1970.