HRTF Magnitude Synthesis via Sparse Representation of Anthropometric Features

Piotr Bilinski, Jens Ahrens, Mark R. P. Thomas, Ivan J. Tashev, John C. Platt

HRTF MAGNITUDE SYNTHESIS VIA SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES

Piotr Bilinski[1*] *Jens Ahrens*[2†] *Mark R. P. Thomas*[3] *Ivan J. Tashev*[3] *John C. Platt*[3]

[1] INRIA, 2004 Route des Lucioles, 06902 Sophia Antipolis, France
[2] University of Technology Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
[3] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
piotr.bilinski@inria.fr, jens.ahrens@tu-berlin.de, {markth,ivantash,jplatt}@microsoft.com

ABSTRACT

We propose a method for the synthesis of the magnitudes 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 subject's anthropometric features w.r.t. the anthropometric features in the training set. The fundamental assumption is that the magnitudes 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 subject's 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 tensor 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 instrumental evaluation we use log-spectral distortion. 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

1. 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 [2–11] 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 this paper, we propose a method for HRTF synthesis using sparse representation [12, 13]. Sparse representation has recently become a very popular technique in many domains.

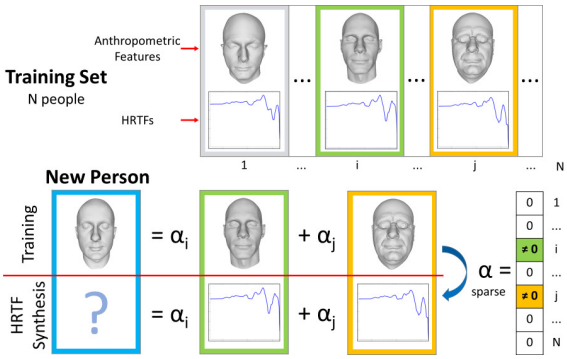

Fig. 1. Block diagram of the proposed approach: The sparse representation is determined for the anthropometric features and then applied to the acoustic data.

[*] Work done during a research internship at Microsoft Research.
[†] Work done while being a post-doctoral researcher at Microsoft Research.

It can be accurately and efficiently computed by ℓ_1 minimization. Moreover, in Computer Vision, it has been shown that if the sparsity is properly harnessed, the choice of input features is no longer critical [14, 15]. It is still important that the number of features is sufficient and that the sparse representation can be correctly found.

The main idea of the presented approach is to treat the HRTF synthesis problem 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 directly on the HRTF tensor data to synthesize the subject's HRTFs as illustrated in Fig. 1. For simplicity, we consider only the magnitudes of the HRTFs. Preliminary experiments performed by the authors on the complex HRTF data using different methods of time alignment yielded comparable results.

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, 16].

The remainder of the paper is organized as follows. Section 2 presents the collected dataset. In Section 3, we describe our sparse representation based approach. In Section 4, we present experimental results. Finally, we conclude in Section 5.

2. DATA COLLECTION

We created a new dataset for the presented study that consists of measured HRTFs and 96 anthropometric features of 36 subjects with an age range from 16 to 61 years (age mean of 33).

2.1. HRTF Measurement

Fig. 2 illustrates the setup for the HRTF measurement. It consists of an arc equipped with 16 evenly distributed loudspeakers that moves to 25 different measurement positions at steps of $11.25°$ between $-45°$ elevation in front of the subject to $-45°$ elevation behind the subject. The subject sits in a chair with the head fixed in the center of the arc. The chirp signals played by the loudspeakers are recorded with omnidirectional microphones that are placed at the entrances of the subject's blocked ear canals. The HRTFs are measured at $16 \times 25 = 400$ positions.

The mechanical setup does not allow the measurement of HRTFs at positions underneath the subject. The data for these positions is obtained by interpolating the measured data using the approach from [17] to the virtual complement of the measurement grid. This results in $32 \times 16 = 512$ sound source locations that are each represented by 512 frequency bins (from 0 Hz to 24 kHz) for the left and the right ear separately.

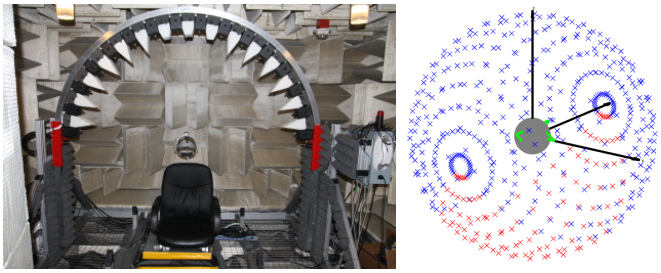

Fig. 2. HRTF measurement setup (left image), and measured and interpolated HRTF directions (right image; blue and red marks, respectively).

For simplicity we will omit differentiating of the left and the right ears further in this paper. The HRTF synthesis is identical for both ears.

2.2. 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 third method.

The extracted anthropometric features are superset of the CIPIC HRTF Database [16] and are listed in Table 1.

Table 1. List of used anthropometric features.

Head-related features:
head height, width, depth, and circumference;
neck height, width, depth, and circumference;
distance between eyes / distance between ears;
maximum head width (including ears);
ear canals and eyes positions;
intertragal incisure width; inter-pupillary distance.
Ear-related features:
pinna: position offset (down/back); height; width; rotation angle;
cavum concha height and width;
cymba concha height; fossa height.
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.

3. PROPOSED APPROACH

3.1. Training Data Representation

Assume that we have N subjects in the training set.

HRTFs. The HRTFs for each subject are described by a tensor of size $D \times K$, where D is the number of HRTF directions and K is the number of frequency bins. All the HRTFs of the training set are stacked in a new tensor $\mathbf{H} \in \mathbb{R}^{N \times D \times K}$, so the value $H_{n,d,k}$ corresponds to the k-th frequency bin for d-th HRTF direction of the n-th person.

Anthropometric features. In the preparation stage all of the categorical features 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.

3.2. Sparse Representation for HRTF Synthesis

We propose to synthesize HRTFs for a new subject given its anthropometric features $\mathbf{y} \in [0,1]^A$. The main idea is to treat the HRTF synthesis problem as finding a sparse representation of the subject's anthropometric features, with the assumption that the HRTFs are in the same relation. We 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 ($\mathbf{y} = \boldsymbol{\beta}^T \mathbf{X}$), and then apply the same sparse vector directly on the HRTF tensor data $\mathbf{H}$. We can write this task as a minimization problem, for a non-negative shrinking parameter λ:

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

The first part of the above equation minimizes the differences between values of $\mathbf{y}$ and the new representation of $\mathbf{y}$. 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 3.4.

We solve the above minimization problem using Least Absolute Shrinkage and Selection Operator (LASSO) [18, 19]. We assume that the HRTFs 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 HRTF tensor data and the subject's HRTF values $\hat{H}$ are defined as:

$$\hat{H}_{d,k} = \sum_{n=1}^{N} \beta_n H_{n,d,k}\ , \quad (2)$$

where $\hat{H}_{d,k}$ corresponds to k-th frequency bin for d-th HRTF direction of the synthesized HRTF.

3.3. HRTF Metrics

To determine the accuracy of the synthesized HRTFs, we compare them with the true (measured) HRTFs of the subject under consideration.

For objective evaluation, we use the log-spectral distortion (LSD) as a distance measure between two HRTFs for a given sound source direction d and all frequency bins from the range k_1 to k_2, as commonly used in the recent literature, e.g. [2, 3, 20]:

$$\mathrm{LSD}_d(\mathbf{H}, \hat{\mathbf{H}}) = \sqrt{\frac{\sum_{k=k_1}^{k_2} \left(20 \log_{10} \frac{|H_d(k)|}{|\hat{H}_d(k)|} \right)^2}{k_2 - k_1 + 1}}\ [\mathrm{dB}], \quad (3)$$

where $H_d(k)$ is the measured HRTF for the d-th direction, $\hat{H}_d(k)$ is the synthesized HRTF for the same (d-th) direction, and $k_2 - k_1 + 1$ is the number of considered frequencies. Note that the perceptual meaning of LSD_d is unclear.

To compare two HRTF sets for all available directions, we use the root mean square error (RMSE):

$$\mathrm{LSD}(\mathbf{H}, \hat{\mathbf{H}}) = \sqrt{\frac{1}{D} \sum_{d=1}^{D} \left(\mathrm{LSD}_d(\mathbf{H}, \hat{\mathbf{H}}) \right)^2}\ [\mathrm{dB}], \quad (4)$$

where $D = 512$ is the number of available HRTF directions.

Note that when we concatenate HRTF values of all the HRTF directions into one dimensional data tensor, (4) is equivalent to (3). The perceptual meaning of LSD is equally unclear.

3.4. Regularization Parameter λ

Our approach has only one parameter λ, which is a non-negative regularization parameter. To prevent over-fitting, we tune this parameter on the training set using leave-one-person-out cross-validation approach [19, 21]. We select the parameter λ which gives the smallest cross-validation error. The cross-validation error is calculated as the root mean square error, using (4).

4. EXPERIMENTS

4.1. Evaluation Protocol

To estimate the accuracy of the proposed approach, we sequentially use the data of one person for testing and the remaining data of $N - 1$ people for training. The HRTFs of

Table 2. Evaluation results in [dB].

Direction	Frequencies [Hz]	The Best Classifier	Sparse Representation	Ridge Regression	HATS	The Worst Classifier
Straight	50 - 8000	2.4633	3.5286	5.8927	6.1292	7.856
	20 - 20000	4.2049	5.5754	8.7495	7.9714	10.2496
All	50 - 8000	4.3176	4.4883	6.1377	7.3503	7.8506
	20 - 20000	9.3792	9.878	12.1868	13.7724	14.9344

each person from the dataset are predicted once. We optimize the parameter λ for every training set separately (see Sec. 3.4). The evaluation metric is RMSE. We evaluate the proposed approach and the baselines in two frequency bands: full audible bandwidth (20 Hz - 20 kHz), and wideband (50 Hz - 8 kHz), assuming that the latter frequency band contains most of the critical information. The evaluation is conducted for one direction (straight ahead) as well as for all available 512 directions, for the left and right ears combined.

4.2. "The Best" and "The Worst" Classifiers baselines

To assess how well our technique performs and 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 HRTF from the training set in the LSD_d and LSD sense using only HRTF data.

4.3. Ridge Regression baseline

We also compare our approach with the ridge regression model [19, 22, 23], where the ℓ_1 norm regularization term is replaced with the ℓ_2 norm regularization term. Therefore, we no longer impose the sparsity in the model. We can write this as a minimization problem, for a non-negative parameter λ:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{a=1}^{A} \left(y_a - \sum_{n=1}^{N} \beta_n X_{n,a} \right)^2 + \lambda \sum_{n=1}^{N} \beta_n^2 \right), \quad (5)$$

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 3.4. This minimization problem is convex and hence has a unique solution.

4.4. HATS baseline

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

4.5. Results

The experimental results are presented in Table 2 for the full audible bandwidth (20 Hz to 20 kHz) and for the wideband (50 Hz to 8 kHz).

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 HRTF classifier. Additional experiments, not presented here, show that removing anthropometric features from any of the four categories (Sec. 2.2) does not significantly affect the results.

The ridge regression model shows much worse results than the sparse representation, which confirms the importance of sparsity in our approach.

The HRTFs of the HATS typically show RMSEs between the ridge regression model and the worst HRTF classifier.

5. CONCLUSIONS

We proposed a method for HRTF synthesis using anthropometric features and sparse representation. 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 combine the HRTF magnitudes in the dataset and thereby synthesize a personalized set of HRTF magnitudes. The log-spectral distortion between the synthesized and actual measured HRTFs for the subject under consideration confirm the effectiveness of the sparse representation based approach. Our method shows lower distortions than all other evaluated techniques and obtains results close to the best possible HRTF classifier (*i.e.* the nearest HRTF in the training set).

Future work includes determining a perceptually motivated distance measure and validating the synthesized HRTFs in a perceptual experiment.

6. REFERENCES

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

[2] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, February 2008.

[3] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada, May 2013.

[4] D. N. Zotkin, J. Hwang, R. Duraiswami, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2003.

[5] G. Grindlay and M. A. O. Vasilescu, "A multilinear approach to HRTF personalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, April 2007.

[6] 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 (ICME)*, Baltimore, Maryland, USA, July 2003.

[7] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 3, pp. 508–519, March 2013.

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

[9] Z. Haraszy, D.-G. Cristea, V. Tiponut, and T. Slavici, "Improved head related transfer function generation and testing for acoustic virtual reality development," in *World Scientific and Engineering Academy and Society Circuits, Systems, Communications and Computers (WSEAS CSCC) Multiconference - WSEAS International Conference on Systems (ICS)*, Corfu Island, Greece, July 2010.

[10] 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)*, New Paltz, New York, USA, October 2013.

[11] W. W. Hugeng and D. Gunawan, "Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements," *Journal of Telecommunications*, vol. 2, no. 2, pp. 31–41, May 2010.

[12] Ke Huang and S. Aviyente, "Sparse representation for signal classification," in *Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, December 2006.

[13] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics (CPAM)*, vol. 59, no. 6, pp. 797–829, June 2006.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 2, pp. 210–227, February 2009.

[15] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Yi Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 2, pp. 372–386, February 2012.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2001.

[17] J. Ahrens, M. R. P. Thomas, and I. Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hollywood, California, USA, December 2012.

[18] S. Kukreja, J. Lofberg, and M. J. Brenner, "A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification," in *International Federation of Automatic Control Symposium on System Identification (IFAC SYSID)*, Newcastle, Australia, March 2006.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, second edition, 2009.

[20] K. J. Fink and L. E. Ray, "Tuning principal component weights to individualize HRTFs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Quebec, Canada, August 1995.

[22] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, February 1970.

[23] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, February 1970.