HRTF MAGNITUDE SYNTHESIS VIA SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES

Piotr Bilinski1∗ Jens Ahrens2† Mark R. P. Thomas3 Ivan J. Tashev3 John C. Platt3

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.
It can be accurately and efficiently computed by $\ell_1$ 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 $11.25^\circ$ between $-45^\circ$ elevation in front of the subject to $-45^\circ$ 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: | \begin{align*}
\text{head height, width, depth, and circumference;} \\
\text{neck height, width, depth, and circumference;} \\
\text{distance between eyes / distance between ears;} \\
\text{maximum head width (including ears);} \\
\text{ear canals and eyes positions;} \\
\text{intertragal incisure width;} \\
\text{inter-pupillary distance.}
\end{align*} |
|------------------------|--------------------------------------------------|
| Ear-related features:   | \begin{align*}
\text{pinna: position offset (down/back);} \\
\text{height;} \\
\text{rotation angle;} \\
\text{cavum concha height and width;} \\
\text{cymba concha height;} \\
\text{fossa height.}
\end{align*} |
| Limbs and full body features: | \begin{align*}
\text{shoulder width, depth, and circumference;} \\
\text{torso height, width, depth, and circumference;} \\
\text{distances: foot– knee;} \\
\text{knee– hip;} \\
\text{elbow– wrist;} \\
\text{wrist– fingertip;} \\
\text{height.}
\end{align*} |
| Other features:         | \begin{align*}
\text{gender; age range; age; race;} \\
\text{hair color; eye color; weight; shirt size; shoe size.}
\end{align*} |
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 \( y \in [0, 1]^A \). 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 \( \beta^0 = [\beta_1, \beta_2, ..., \beta_N]^T \) that represents all the features of one person.

\[
\beta^0 = \arg\min_{\beta} \left( \sum_{n=1}^{A} (y_n - \sum_{n=1}^{N} \beta_n x_{n,n})^2 + \lambda \sum_{n=1}^{N} |\beta_n| \right). \tag{1}
\]

The first part of the above equation minimizes the differences between values of \( y \) and the new representation of \( y \). Note that the sparse vector \( \beta \in \mathbb{R}^N \) provides one weight value per person (and not per anthropometric feature). The second part of the above equation is the \( \ell_1 \) norm regularization term that imposes the sparsity constraints, and makes the vector \( \beta \) sparse. The shrinking parameter \( \lambda \) 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 \( \beta \) from the anthropometric features, we directly apply it to the HRTF tensor data and the subject’s HRTF values \( \hat{H} \) are defined as:

\[
\hat{H}_{d,k} = \sum_{n=1}^{N} \beta_n H_{n,d,k}, \tag{2}
\]

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]:

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

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

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

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

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 \( \lambda \)

Our approach has only one parameter \( \lambda \), 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 \( \lambda \) 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
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


