ON THE PREPROCESSING AND POSTPROCESSING OF HRTF INDIVIDUALIZATION BASED ON SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES

Jianjun He, Woon-Seng Gan, and *Ee-Leng Tan*

Digital Signal Processing Lab, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore
{jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg, etanel@ntu.edu.sg}

ABSTRACT

Individualization of head-related transfer functions (HRTFs) can be realized using the person's anthropometry with a pre-trained model. This model usually establishes a direct linear or non-linear mapping from anthropometry to HRTFs in the training database. Due to the complex relation between anthropometry and HRTFs, the accuracy of this model depends heavily on the correct selection of the anthropometric features. To alleviate this problem and improve the accuracy of HRTF individualization, an indirect HRTF individualization framework was proposed recently, where HRTFs are synthesized using a sparse representation trained from the anthropometric features. In this paper, we extend their study on this framework by investigating the effects of different preprocessing and postprocessing methods on HRTF individualization. Our experimental results showed that preprocessing and postprocessing methods are crucial for achieving accurate HRTF individualization.

Index Terms— Head-related transfer function (HRTF), anthropometry, 3D audio, HRTF individualization

1. INTRODUCTION

Humans' listening in the physical world is in three dimensions (3D). Seamless natural listening experience is a common pursuit in 3D audio for virtual auditory display (VAD) applications [1]. The cues that a human requires for sound localization can mostly be encapsulated in spatial filters called head-related transfer functions (HRTFs) [2], which are commonly used in 3D audio rendering for headphone and loudspeaker playback [3], [4].

However, HRTFs are highly individualized as they are the resultants of the interaction of sound waves with the human body in the form of propagation, reflection, and diffraction [5], [6]. As a consequence, use of non-individualized HRTFs usually results in spatial and timbral distortions in VAD [7]. To solve this problem, researchers have been working on HRTF individualization over the past two decades [8], [9]. In general, individualized HRTFs can be obtained from direct acoustic measurements [6] with interpolation [10], [11], perceptual feedbacks [12]-[14], special frontal projection of sound [14], and anthropometry [16]-[27].

Due to the inherent relation between HRTFs and anthropometry of a person, anthropometry data is widely used for HRTF individualization [16]-[27], where an underlying model is usually first trained from the anthropometry and HRTF database. For most existing methods, this training is often built on linear or non-linear relations, where dimensionality reduction of HRTF database and selection of anthropometric features are critical [17]. Recently, Tashev *el at* [26], [27] proposed an indirect anthropometry based HRTF individualization method. Instead of training the relation between HRTFs and anthropometry, their method obtains a sparse representation for the anthropometry of a new person using the anthropometry of the training subjects. This sparse representation is then used to synthesize the HRTFs of the new person using the HRTFs of the corresponding training subjects. In this paper, we introduce preprocessing and postprocessing methods in this HRTF individualization method and investigate their effects on the performance of HRTF individualization. In this work, we focus on the synthesis of the individualized HRTF magnitude spectra, although the methods discussed in this paper can also be applied to HRTF phase spectra.

The remainder of this paper is structured as follows. Section 2 introduces the anthropometry and HRTFs. Section 3 details the proposed method for anthropometry estimation. In Section 4, experimental results are presented and discussed. Finally, we conclude this paper in Section 5.

2. ANTHROPOMETRY AND HRTFS

The popular CIPIC HRTF database [6] consisting of HRTFs and anthropometry of human subjects is used in our study. The anthropometric features are made up of 17 head-and-torso related features and 20 pinna related features. However, there are only 35 subjects whose 37 anthropometric features are complete in the CIPIC database. In general, the anthropometric features measured follow a

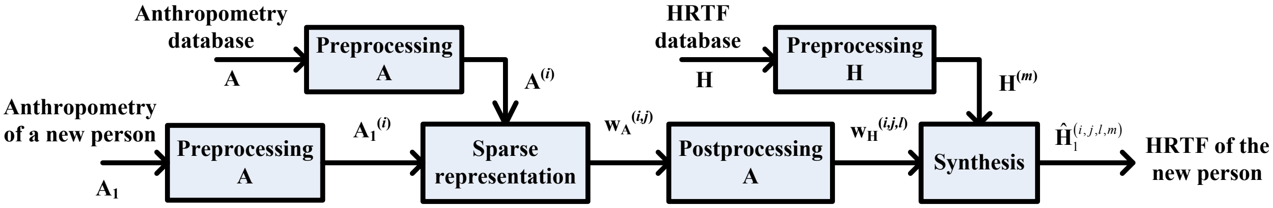

Fig. 1. Block diagram of HRTF individualization using sparse representation of anthropometric features with preprocessing and postprocessing.

normal distribution (as found in general population) with different means and variances. Interested readers can refer to [6] for more details. For the HRTFs in the CIPIC database, the measurement distance is fixed at one meter and a total number of 1250 directions (25 azimuths and 50 elevations) are measured. After free-field compensation and truncation, each head-related impulse response (HRIR, time-domain representation of HRTF) is 200-sample long with a sampling frequency at 44.1 kHz.

3. HRTF INDIVIDUALIZATION WITH PREPROCESSING AND POSTPROCESSING

In this section, we discuss in detail the preprocessing and postprocessing methods in the anthropometry based HRTF individualization, as illustrated in Fig. 1. Assuming we have the anthropometry and HRTF data of S subjects in the training database, and F features in one set of anthropometric features, we denote the anthropometry feature training database by $\mathbf{A} \in \mathbb{R}^{S \times F}$. The anthropometric features of the testing subject is denoted by $\mathbf{A}_1 \in \mathbb{R}^{1 \times F}$. Similarly, we denote the training database of HRTF magnitude as $\mathbf{H} \in \mathbb{R}^{S \times D \times K}$, where D and K refer to the number of directions and frequency bins, respectively. The actual and estimated HRTF magnitude of the testing subject is denoted by $\mathbf{H}_1$, $\hat{\mathbf{H}}_1 \in \mathbb{R}^{1 \times D \times K}$, respectively.

3.1. Preprocessing for anthropometry data

According to the CIPIC database [6], the anthropometry data measured has different scale, mean, and variance. Therefore, a preprocessing method to normalize the anthropometry data is necessary. A key consideration is that different anthropometry data would have approximately equal importance in determining the sparse representation. Denote the combined anthropometric features of the training and testing subjects as $\mathbf{A}_0 = \left[\mathbf{A} \quad \mathbf{A}_1\right]$. In the following, four anthropometry preprocessing methods are considered:

1). Direct: the anthropometry data is used directly without any processing, i.e., $\mathbf{A}^{(1)}(f) = \mathbf{A}(f) \quad \forall f = 1,2,...,F$.

2). Min-max: each anthropometry feature is subtracted by the sample minimum and subsequently divided by the difference between the sample maximum and sample minimum, i.e., $\mathbf{A}^{(2)}(f) = \dfrac{\mathbf{A}(f)\text{-min}\left[\mathbf{A}_0(f)\right]}{\max\left[\mathbf{A}_0(f)\right] - \min\left[\mathbf{A}_0(f)\right]}$.

3). Standard score: each anthropometry feature is subtracted by the sample mean and subsequently divided by the standard deviation of the sample, i.e., $\mathbf{A}^{(3)}(f) = \dfrac{\mathbf{A}(f)\text{-mean}\left[\mathbf{A}_0(f)\right]}{\text{std}\left[\mathbf{A}_0(f)\right]}$.

4). Standard deviation: each anthropometry feature is divided by the standard deviation of the sample, i.e., $\mathbf{A}^{(4)}(f) = \dfrac{\mathbf{A}(f)}{\text{std}\left[\mathbf{A}_0(f)\right]}$. Compared to the standard score, standard deviation normalization preserves the mean value of the anthropometric features by considering the weights in the sparse representation would sum up to one.

Note that the same preprocessing method is also applied to the anthropometry of the testing subject.

3.2. Preprocessing for HRTF data

We consider three types of preprocessing for HRTF magnitude, which result in (linear) magnitude, log magnitude, and power. These types of preprocessing are expressed as: $\mathbf{H}^{(m)}(d,k) = \begin{cases} \mathbf{H}(d,k) & ,m=1 \\ 20\log_{10}\left[\mathbf{H}(d,k)\right] & ,m=2 \\ \left[\mathbf{H}(d,k)\right]^2 & ,m=3 \end{cases}$

$\forall d = 1,2,...,D; \quad k = 1,2,...,K.$

3.3. Sparse representation

The key assumption in this HRTF individualization method is that HRTFs follow the same sparse representation as anthropometric features. Thus, we first learn a sparse representation between the anthropometric features of the training and testing subjects (both after anthropometry preprocessing i), i.e.,

$$\mathbf{A}_1^{(i)} \approx \mathbf{w}_\mathbf{A}^{(i)} \mathbf{A}^{(i)}, \tag{1}$$

where $\mathbf{w}_\mathbf{A}^{(i)} = [w_\mathbf{A}^{(i)}(1), w_\mathbf{A}^{(i)}(2), ..., w_\mathbf{A}^{(i)}(S)]$ provides one weight value per subject in the training database. Hence, the sparse representation $\mathbf{w}_{\mathbf{A}}^{(i)}$ can be obtained by solving the following minimization problem [26]

$$\mathbf{w}_{\mathbf{A}}^{(i,1)} = \arg\min_{\mathbf{w}_{\mathbf{A}}^{(i)}} \left\| \mathbf{A}_1^{(i)} - \mathbf{w}_{\mathbf{A}}^{(i)} \mathbf{A}^{(i)} \right\|_2^2 + \lambda \left\| \mathbf{w}_{\mathbf{A}}^{(i)} \right\|_1 , \qquad (2)$$

where λ is a regularization parameter that controls the sparsity of $\mathbf{w}_{\mathbf{A}}{}^{(i)}$. Larger values of λ lead to a more sparse representation. Furthermore, we also consider adding an additional nonnegative constraint to the sparse representation, and the final nonnegative sparse representation is expressed as

$$\mathbf{w}_{\mathbf{A}}^{(i,2)} = \arg\min_{\mathbf{w}_{\mathbf{A}}^{(i)}} \left\| \mathbf{A}_1^{(i)} - \mathbf{w}_{\mathbf{A}}^{(i)} \mathbf{A}^{(i)} \right\|_2^2 + \lambda \left\| \mathbf{w}_{\mathbf{A}}^{(i)} \right\|_1 , \quad \text{s.t. } \mathbf{w}_{\mathbf{A}}^{(i)} \geq 0. (3)$$

These two optimization problems (2) and (3) are solved using l_1-regularized least squares problem solver discussed in [28].

3.4. Postprocessing for anthropometry data

In the postprocessing for anthropometry data, we consider two approaches to deal with the weights obtained in sparse representation. The first approach is to use the weights directly, while the second approach normalizes the weights by the sum of the weights in sparse representation. This normalization would make the sum of the weights equal to one. Thus, we express the postprocessed sparse representation as $\mathbf{w}_{\mathbf{H}}^{(i,j,l)} = \begin{cases} \mathbf{w}_{\mathbf{A}}^{(i,j)} & ,l=1 \\ \dfrac{\mathbf{w}_{\mathbf{A}}^{(i,j)}}{\sum_{s=1}^{S} w_{\mathbf{A}}^{(i,j)}(s)} & ,l=2 \end{cases} \cdot$

3.5. HRTF synthesis

The postprocessed sparse representation $\mathbf{w}_{\mathbf{H}}^{(i,j,l)}$ is applied to the corresponding HRTF training database to estimate the HRTFs of the testing subject, which are subsequently converted back to the magnitude domain, i.e.,

$$\hat{\mathbf{H}}_1^{(i,j,l,m)}(d,k) = \begin{cases} \mathbf{w}_{\mathbf{H}}^{(i,j,l)} \mathbf{H}^{(m)}(d,k) & ,m=1 \\ 10^{\frac{\mathbf{w}_{\mathbf{H}}^{(i,j,l)} \mathbf{H}^{(m)}(d,k)}{20}} & ,m=2 \\ \sqrt{\mathbf{w}_{\mathbf{H}}^{(i,j,l)} \mathbf{H}^{(m)}(d,k)} & ,m=3 \end{cases} .$$

3.6. Evaluation

The objective evaluation of HRTF individualization accuracy is obtained with the commonly used distance measure spectral distortion (SD) [14], [17], [18], [26]. Considering S_{test} subjects in the test, we compute the SD (in dB) as

$$\mathrm{SD}^{(i,j,l,m)} = \sqrt{\frac{1}{S_{test}} \frac{1}{D} \frac{1}{K} \sum_{s=1}^{S_{test}} \sum_{d=1}^{D} \sum_{k=1}^{K} \left[20 \log_{10} \frac{\hat{\mathbf{H}}_s^{(i,j,l,m)}(d,k)}{\mathbf{H}_s(d,k)} \right]^2} , (4)$$

where $\mathbf{H}_s$, $\hat{\mathbf{H}}_s^{(i,j,l,m)}$ denote the actual and the estimated HRTF magnitude of the s^{th} testing subject, respectively. Note that SD is equivalent to the root-mean-square-error (RMSE) of log magnitude, and smaller SD indicates a better performance.

3.7. Selection of regularization parameter

In this paper, we adopt the cross validation technique [29] to determine the regularization parameter, with SD chosen as the criterion. That is to say, a number of regularization parameters will be tested and the value of λ which yields the lowest SD is to be determined. However, as seen from (2), the regularization parameter is very sensitive to the scale of the anthropometric features which varies among anthropometry preprocessing methods. To alleviate the selection difficulty, we normalize λ using $\lambda = \dfrac{\lambda_0}{1-\lambda_0} \left\| \mathbf{A}_1^{(i)} \right\|_2^2$. The normalization using the squared l_2-norm of the anthropometric features $\mathbf{A}_1^{(i)}$ ensures the scale of λ to fit any preprocessing methods. Furthermore, the introduction of $\dfrac{\lambda_0}{1-\lambda_0}$ will ease the selection of λ since any nonnegative value of λ can be obtained by adjusting λ_0 from 0 to 1. Some preliminary testing indicates that basically we only need to tune λ_0 up to 0.2.

4. EXPERIMENTS AND DISCUSSIONS

To maximally use the CIPIC database in our experiment, we sequentially select one subject as the testing subject, while the remaining subjects as the training subjects. As there are 35 subjects in the CIPIC database, we have $S_{\text{test}} = 35$ testing cases, and in each case, there are $S = 34$ subjects in the training database. Each preprocessing and postprocessing method is employed separately and hence, in total, we have $4 \times 3 \times 2 \times 2 = 48$ methods. Finally, we compute the SD for each method. The results of anthropometry estimation accuracy are illustrated in Fig. 2. Our observations on different methods are as follows.

First, we summarize the effect of preprocessing and postprocessing methods for direct sparse representation used in the training of anthropometry data, as shown in Fig. 2(a) and 2(b). Among the four anthropometry preprocessing methods, we found that the performance of standard score is the worst, whereas the best is obtained with standard deviation method. Among the three HRTF preprocessing methods, power is the worst, whereas the overall best performance is obtained with log magnitude. Considering the best anthropometry preprocessing method (standard deviation) and best HRTF preprocessing method (log magnitude), we found that the effect of postprocessing methods is very minimal.

Second, we summarize the effect of preprocessing and postprocessing methods for nonnegative sparse

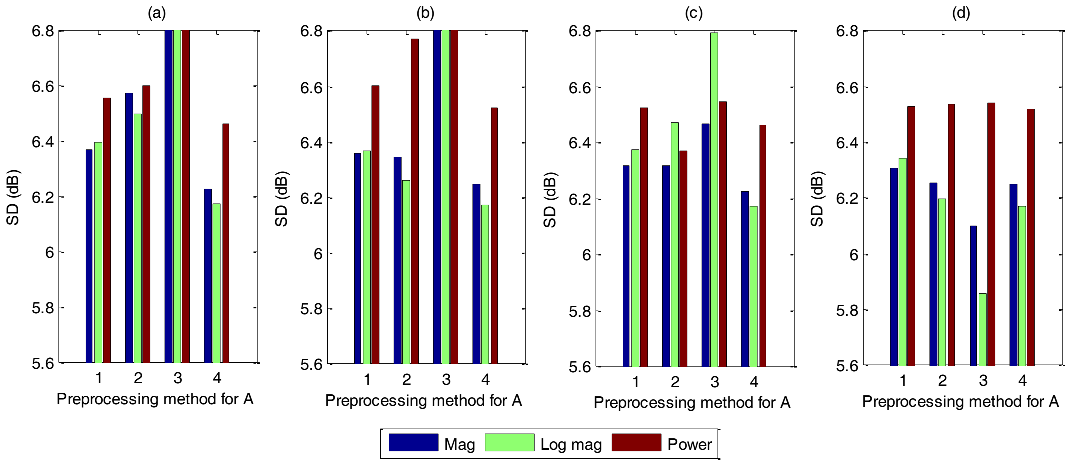

Fig. 2. Results of SD with respect to different processing methods. The sparse representation and postprocessing methods are: (a) direct and direct; (b) direct and normalized; (c) nonnegative and direct; (d) nonnegative and normalized, respectively. The four anthropometry preprocessing methods are shown in different x-axis value (1, 2, 3, 4), which represents direct, min-max, standard score, and standard deviation. Note that the range of y-axis SD values has been restricted to [5.6, 6.8] for a clearer view, which results in the SD values of the standard score anthropometry preprocessing method out of the range.

representation in Fig. 2(c) and 2(d). Compared with direct sparse representation, a better performance is observed in all nonnegative sparse representation methods that use the same preprocessing and postprocessing methods. The most significant improvement is found with the standard score preprocessing method for anthropometry and log magnitude for HRTFs, especially when applying normalized anthropometry postprocessing, as shown in Fig. 2(d). This finding further validates the importance of having nonnegative weights in sparse representation.

Our comparison of the 48 methods reveals that the best performance (i.e., lowest SD = 5.86 dB) is obtained with the following specifications: nonnegative sparse representation, standard score anthropometry preprocessing, and log magnitude of HRTFs with normalization applied to the weights. We have also considered an additional method which selects the closest set of anthropometric features from the training database and uses the HRTF of this subject as the individualized HRTFs for the new person. The SD of this method is 8.11 dB, which is much worse than the proposed method. Furthermore, we compute a lower bound for this type of linear regression based HRTF individualization methods. As SD can be considered as the RMSE of HRTFs in the log magnitude domain, the theoretically best weights can be obtained as $\mathbf{w}^{(\text{opt})} = \left[\mathbf{H}^{(2)}\right]^{+} \mathbf{H}_1^{(2)}$, where $\left[\mathbf{H}^{(2)}\right]^{+}$ represents the pseudo-inverse of $\mathbf{H}^{(2)}$. This method achieves the theoretical lower bound for SD, which is 5.12 dB with the CIPIC database. However, this performance is difficult to achieve in practice as the optimal weights $\mathbf{w}^{(\text{opt})}$ does not always satisfy nonnegative or sparse constraints. Besides the objective evaluation discussed in this paper, it would also be meaningful to evaluate HRTF individualization using subjective tests [30].

5. CONCLUSIONS

In this paper, we studied the effects of various preprocessing and postprocessing methods for HRTF individualization based on sparse representation of anthropometric data. Specifically, we investigated four anthropometry preprocessing methods, three HRTF preprocessing methods, two sparse representation methods, and two anthropometry postprocessing methods. Our experimental results with the CIPIC HRTF database indicate that the performance of HRTF individualization is generally affected by the preprocessing and postprocessing methods, and the preprocessing methods introduce more performance variations. Adding nonnegative constraints in sparse presentation improves the performance. The best performance is obtained with standard score in anthropometry normalization, log magnitude spectra of HRTFs, and nonnegative sparse representation with weights normalized. This method yields a SD of 5.86 dB, which is much better than the closest HRTF set method (8.11 dB) and relatively close to the theoretical lower bound (5.12 dB) of such linear regression based HRTF individualization methods. Future work includes subjective evaluation of the HRTF individualization methods.

ACKNOWLEDGMENT

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2010-T2-2-040.

REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Cambridge, MA: Academic Press, 1994.

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

[3] H. Møller, "Fundamentals of Binaural Technology," *Applied Acoustics*, vol. 36, 171-218, 1992.

[4] W. G. Gardner, and K. D. Martin, "HRTF Measurements of a KEMAR," *J. Acoust. Soc. Amer., vol.*, vol. 97, pp. 3907-3908, 1995. See also http:// www.sound.media.mit.edu/KEMAR.html.

[5] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Aud. Eng. Soc.*, vol. 43, pp. 300-321, 1995.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in Proc. *IEEE WASPAA*, New Paltz, NY, USA, Oct. 2001.

[7] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Non-individualized Head-Related Transfer Functions," *J. Acoust. Soc. Amer.*, vol. 94, pp. 111-123, 1993.

[8] S. Xu, Z. Li, and G. Salvendy, "Individualization of Head-related transfer function for three-dimensional virtual auditory display: a review," in R. Shumaker (Ed.): Virtual Reality, HCII 2007, LNCS 4563, pp. 397–407, 2007.

[9] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," in press, *IEEE Signal Processing Magazine*, DOI: 10.1109/MSP.2014.2372062, Mar. 2015.

[10] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Amer.*, vol. 134, pp. EL547–554, Dec. 2013.

[11] G. D. Romigh, "Individualized hread-related transfer functions: efficient modeling and estimation from small sets of spatial samples," PhD dissertation, School of Electrical and Computer Engineering, Carnegie Mellon University, 2012.

[12] C. J. Tan and W. S. Gan, "User-defined spectral manipulation of HRTF for improved localisation in 3 D sound systems," *Electronics letters,* vol. 34, no. 25, pp. 2387-2389, Dec. 1998.

[13] B. U. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," in *Proceedings of the 2003 International Conference on Auditory Display*, pp. 259-262, 2003.

[14] K. J. Fink, and L. Ray, "Individualization of head related transfer functions using principal component analysis," *Applied Acoustics*, vol. 87, pp. 162-173, 2015.

[15] K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 989-1000, Dec. 2013.

[16] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE WASPAA*, New York, pp. 157-160, Oct. 2003,.

[17] 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, Feb. 2008.

[18] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in Proc. *IEEE ICASSP*, Vancouver, British Columbia, Canada, May 2013.

[19] G. Grindlay and M. A. O. Vasilescu, "A multilinear approach to HRTF personalization," in Proc. *IEEE ICASSP*, Honolulu, Hawaii, USA, Apr. 2007.

[20] A. Mohan, R. Duraiswami, D. N. Zotkin, D. DeMenthon, and L. S. Davis, "Using computer vision to generate customized spatial audio," in Proc. *IEEE ICME*, Baltimore, Maryland, USA, Jul. 2003.

[21] 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, Mar. 2013.

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

[23] 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, Jul. 2010.

[24] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous HRTF datasets," in Proc. *IEEE WASPAA*, New Paltz, New York, USA, Oct. 2013.

[25] 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.

[26] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. Tashev, and J. C. Plata, "HRTF magnitude synthesis via sparse representation of anthropometric features," in Proc. *IEEE ICASSP*, Florence, Italy, pp. 4501-4505, May 2014.

[27] I. Tashev, "HRTF phase synthesis via sparse representation of anthropometric features," in Proc. *Information Theory and Applications Workshop (ITA)*, San Diego, CA, pp. 1-5, Feb. 2014.

[28] S. J. Kim, K. Koh, M. Lusig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *J. Selected topics in signal processing*, vol. 1, no. 4, pp. 606-617, Dec. 2007.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[30] J. Breebaart, F. Nater, and A. Kohlrausch, "Spectral and spatial parameter resolution requirements for parametric, filter-bank-based HRTF processing," *J. Audio Eng. Soc.*, vol. 58, no. 3, pp. 126-140, Mar. 2010.