PERSONALIZED HRTF MODELING USING DNN-AUGMENTED BEM

Mengfan Zhang[♭] *Jui-Hsien Wang*[♯] *Doug L. James*[♭]

[♭]Stanford University, Stanford, CA, USA
[♯]Adobe Research, Seattle, WA, USA

ABSTRACT

Accurate modeling of personalized head-related transfer functions (HRTFs) is difficult but critical for applications requiring spatial audio. However, this remains challenging as experimental measurements require specialized equipment, numerical simulations require accurate head geometries and robust solvers, and data-driven methods are hungry for data. In this paper, we propose a new deep learning method that combines measurements and numerical simulations to take the best of three worlds. By learning the residual difference and establishing a high quality spatial basis, our method achieves consistently $2\,\mathrm{dB}$ to $2.5\,\mathrm{dB}$ lower spectral distortion (SD) compared to the state-of-the-art methods.

***Index Terms*—** Head-related transfer functions, spatial principal component analysis, deep learning, boundary element methods

1. INTRODUCTION

Accurate modeling of the head-related transfer functions (HRTFs) is critical for applications requiring spatial audio. Because of morphological differences, every person has a unique HRTF, and capturing this personalized HRTF is important for accurate spatialization [1,2]. Previous works show that using a generic HRTF for spatialization leads to significant perceptual errors such as in-head localization, front-back confusion, and lowered ability to discriminate source elevations [3].

While HRTF capture via experimental measurements is a reliable method for generic HRTFs [4–6], it can be time-consuming (from minutes [7, 8] to hours, depending on the configurations), or even impossible to perform depending on accessibility to the specialized equipment and necessary expertise in order to obtain robust results. Furthermore, subjects have to remain still in the process, or risk introducing additional errors.

Numerical simulation of personalized HRTFs is possible after acquiring a 3D mesh of the subject's head and sometimes part of the body. Different formulations include the boundary element method (BEM) [9] and others [10, 11] can be used. However, they can introduce errors from various approximations such as the fast multipole expansion for BEM [12, 13] or perturbations in the numerical discretization [14]. In addition, these solvers can be slow although faster methods exist [15, 16]. Capturing accurate 3D head geometry is also challenging because of limited scanner resolution, difficulty in soft tissue modeling, occlusion, and various postprocessing requirements (e.g., hole filling, remeshing). The errors are worse at high frequencies.

Data-driven methods are increasingly popular for modeling HRTFs. This can be done, for example, by feeding a subject's anthropometric parameters through neural networks [17–21] to get HRTF predictions. However, the quality of reconstruction depends on the accuracy and size of the underlying dataset, which is often quite small [8, 22–25] compared to other machine learning applications. This limits how expressive the models can be and consequently the ability to model large variations in personalized HRTFs.

In this paper, we seek a way to combine these different methods for better HRTF reconstruction. We propose a new deep learning method that combines simulations with measurements. Our model is predictive like those typical of simulation-based reconstructions, yet versatile and more expressive for modeling the complex HRTF structures that might be missing in simulated results (see §5.2). We use a deep-learning framework based on spatial principal component analysis [19]. Instead of training the networks to directly predict HRTFs, we condition our neural network to learn the difference between simulations and measurements, thereby granting it a way to *correct* inaccurate numerical simulations. We found that our method achieves consistently lower spectral distortions compared to the state-of-the-art methods by $2\,\mathrm{dB}$ to $2.5\,\mathrm{dB}$ in all frequencies within hearing range.

2. BACKGROUND

2.1. Spatial principal component analysis

Spatial principal component analysis (SPCA) factorizes HRTFs using a spatial decomposition [19, 26]:

$$H(\theta, \varphi, f, s) \approx \sum_{q=1}^{Q} d_q(f, s) W_q(\theta, \varphi) + B(\theta, \varphi, f, s), \quad (1)$$

where H represents the ground-truth HRTF (e.g., measurements). Typically, only amplitudes are considered, where

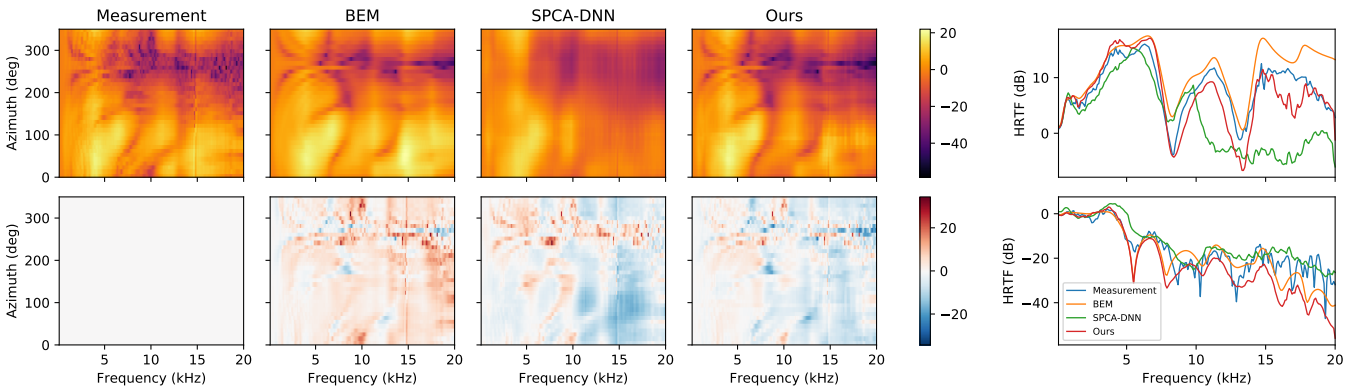

Fig. 1: Comparison of measured HRTF on the horizontal plane (0 degree elevation angle) with the reconstructed HRTFs from BEM, SPCA-DNN, and our method. Left Top: HRTFs in decibels; Left Bottom: residual differences between the reconstructed HRTFs and measurement. Right Top: frequency-dependent HRTF at 90 degree azimuthal angle (ipsilateral); Right Bottom: frequency-dependent HRTF at 270 degree azimuthal angle (contralateral).

phases can be reconstructed [19]. (θ, φ) represents the direction in azimuthal and elevation angles, f is the frequency, and s is the subject. $d_q(f, s)$ contains the weights for the spatial basis $W_q(\theta, \varphi)$. $B(\theta, \varphi, f, s)$ is what we call a bias. The SPCA method uses the sampled average as the bias, i.e.,

$$B(\theta, \varphi, f, s) \equiv H_{\text{av}}(\theta, \varphi) = \frac{1}{N \times S} \sum_s \sum_f H(\theta, \varphi, f, s) \tag{2}$$

for N frequencies and S subjects. Q is the number of basis functions in the chosen subspace. Let D be the total number of directions. If $Q < D$, then the factorization is approximate; if $Q = D$, it is exact. In matrix form, the decomposition (1) can be written as

$$\boldsymbol{H} = \boldsymbol{dW} + \boldsymbol{B}, \tag{3}$$

where HRTF $\boldsymbol{H}$ and bias $\boldsymbol{B}$ are NS-by-D matrices. $\boldsymbol{d}$ and $\boldsymbol{W}$ have dimensions NS-by-Q and Q-by-D, respectively. In the case of SPCA, the mean is used for the bias term, and thus $\boldsymbol{B}$ has identical rows for all frequencies and subjects.

Given a choice of Q, both the weights $\boldsymbol{d}$ and basis matrix $\boldsymbol{W}$ can be computed by extracting the first Q eigenvectors of the covariance matrix of the mean-centered errors,

$$\boldsymbol{E} \equiv \boldsymbol{H} - \boldsymbol{B}, \tag{4}$$

$$\text{Cov} = \boldsymbol{E}^\intercal \boldsymbol{E}. \tag{5}$$

For convenience, let us define $\boldsymbol{E}_{\text{hav}}$ to be the errors obtained with bias defined in (2). Although choosing up to D principal components is possible, this often leads to overfitting, and the general practice is to choose $Q < D$ such that around 90% of the variance is captured [17, 19, 27]. The variance is the sum of the eigenvalues of the covariance matrix. That is,

$$Var = \frac{\sum_{q=1}^{Q} \lambda_q}{\sum_{q=1}^{D} \lambda_q} \times 100\%. \tag{6}$$

2.2. Leveraging deep learning to model individual HRTFs

Zhang *et al.* introduced a deep learning method to further extend SPCA to model personalized HRTFs [19], henceforward referred to as SPCA-DNN. This is done by predicting the weight matrix $\boldsymbol{d}$, the only term in the SPCA factorization that depends on the subjects. SPCA-DNN trains N neural networks, one for each frequency, and each has an identical architecture of three densely connected layers, to predict the weight matrix using 8 anthropometric parameters of each subject, such as the head width and pinna height.

3. OUR APPROACH

In order to take the spectral changes and morphological variations in personalized HRTFs into account, we propose to replace the average bias in SPCA-DNN with a function computed from personalized numerical simulations for each subject, i.e., we use

$$B(\theta, \varphi, f, s) \equiv H_{\text{bem}}(\theta, \varphi, f, s), \tag{7}$$

where H_{bem} is an approximate HRTF of the subject using BEM simulation. Although this choice makes our SPCA no longer centered around the mean, the benefit is clear: H_{bem} utilizes all four independent variables (θ, φ, f, s), and can be much closer to the ground truth, H. Therefore, it reduces the complexity our basis needs to cover and lowers the error, $\boldsymbol{E}_{\text{bem}}$ in (4). A comparison of the two errors, $\boldsymbol{E}_{\text{hav}}$ and $\boldsymbol{E}_{\text{bem}}$, is shown in Fig. 2. As expected, $\boldsymbol{E}_{\text{bem}}$ is a better predictor, which results in overall lower magnitude and narrower error distribution. Of course, this comes with a price due to the 3D head geometry now required for each user of our method. Fortunately, 3D scanning is generally more accessible (e.g., photogrammetry requires only still images captured on consumer-grade cameras to reconstruct one's head) than more specialized HRTF measurement equipment. Further-

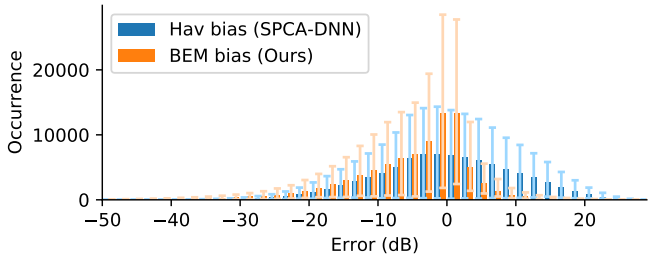

Fig. 2: Histogram of errors ($\boldsymbol{E}$) introduced by different biases (blue: $\boldsymbol{E}_{\text{hav}}$; orange: $\boldsymbol{E}_{\text{bem}}$). The histogram is averaged over all 58 subjects in our dataset; the error bars shown in thin lines represent 25 and 75 percentiles of every error bin.

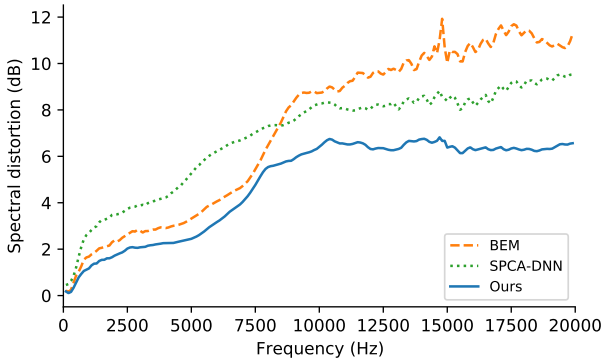

Fig. 3: Average spectral distortions for BEM, SPCA-DNN, and our method. The average SD across all subjects and frequencies are 7.23 dB, 6.89 dB, 4.80 dB for BEM, SPCA-DNN, and our method, respectively.

more, our method is insensitive to errors in BEM simulations, as shown in a later section §5.3.

4. EXPERIMENTAL DESIGN

Our method requires a dataset that contains experimental measurements of individualized HRTFs and the subjects' head meshes in order to perform BEM computations.

4.1. Dataset

We use the HUTUBS dataset [24] in this work as it is one of the few that contains all the required assets. The dataset contains 96 subjects' measured head-related impulse responses (HRIRs), 58 head meshes, and a set of 25 anthropometric parameters for each subject. We first Fourier transform HRIRs into the frequency domain to obtain HRTFs. Only the amplitudes are kept and stored as logarithms for $\boldsymbol{H}$ in (3). We also use 8 anthropometric parameters identical to the ones in the SPCA paper [19].

4.2. Numerical simulation

For numerical simulation, we use the open-source library `MESH2HRTF`, which is based on a fast-multipole BEM solver [28]. For each subject in the dataset, we run a BEM simulation for each frequency from 100 Hz to 20 kHz in increments of 100 Hz and sample the solution at 440 collocated spatial directions on a 1.47 m shell away from the subject's head center. All BEM simulations were performed on a single workstation (dual Intel Xeon E5-2690V3 processors); the principle of reciprocity is used to save computational time.

The 3D head models are adaptively remeshed to ensure solver performance. The procedure is identical to that described in [24, 28], which gradually coarsens the mesh from the ipsilateral ear at 1 mm to the contralateral ear at 10 mm. We were concerned that the resolution on the 10 mm side might be too coarse, so we perform studies on meshes at 1 mm-8 mm, 1 mm-6 mm, 1 mm-5 mm, and 1 mm-4 mm resolutions. We found no significant influence of this choice on our results, similar to previous work [29].

4.3. Neural networks and training

We closely follow the original SPCA-DNN method introduced in §2.2 to train our neural networks. Each weight network takes 8 anthropometric parameters from a user and predicts a 1-by-Q submatrix in the weight matrix $\boldsymbol{d}$ for a frequency sample. Since the Fourier transform is conjugate symmetric with respect to the frequencies for real-valued HRIRs, and only the amplitudes are modeled, there are in total $\lceil N/2 \rceil + 1$ networks. To avoid overfitting, we employ the same strategy as in SPCA-DNN to split subjects into training, validation, and testing sets, and perform k-fold cross validation during training. Early stopping is also used to prevent significantly increased validation errors. The mean squared error (MSE) of training, validation, and test sets are 14.59, 13.65, and 15.83 respectively. No overfitting was detected for our model.

Similar to SPCA-DNN [19], we train an additional neural network with 5 densely connected layers in order to predict HRTF in an arbitrary direction. We found that this added complexity only contributes an insignificant portion (around 0.13 dB) to the overall error.

5. RESULTS & ANALYSIS

Our method is analyzed in this section. The distance between any two HRTFs, H and $\hat{H}$, is measured by the frequency-dependent spectral distortion (SD) [14, 19]:

$$SD(f) = \frac{1}{DS} \sum_{s} \sum_{\theta} \sum_{\phi} |H(\theta, \varphi, f, s) - \hat{H}(\theta, \varphi, f, s)|. \tag{8}$$

H is typically the ground-truth measurement.

5.1. Quantitative analysis

In Fig. 3, we show that SDs of our method are lower than those of the baselines in all frequencies. On average our

Method	All directions	Hori. plane	Med. plane
BEM	7.23	6.18	8.65
SPCA-DNN	6.89	7.25	**3.89**
Ours	**4.80**	**4.62**	4.54

Table 1: Average SDs for different sets of directions.

method achieves around a $2.43\,\mathrm{dB}$ ($\approx$ 33%) improvement over BEM, and a $2.09\,\mathrm{dB}$ ($\approx 30\%$) improvement over SPCA-DNN. A similar level of improvement is achieved if we restrict the directional samples to those of the horizontal plane (0 elevation angle). On the median plane, SPCA-DNN method results in lower error, but our method is only $0.65\,\mathrm{dB}$ behind in performance. Despite the slightly higher error, our method captures clearer HRTF features compared to SPCA-DNN (e.g., see Fig. 4). We perform a T-test on every frequency and find that statistically our method produces significantly lower errors compared to BEM and SPCA-DNN ($p \ll 0.05$ for both baselines). Due diligence was performed to ensure assumptions of T-test (e.g., variance correction) are satisfied.

5.2. Qualitative analysis

HRTFs reconstructed using our method often have several qualitative characteristics that resemble ground-truth HRTFs that might be lacking in other methods. In particular, we find that our method can produce spectral notches and peaks that closely follow ones from the measurement (see Fig. 1, Fig. 4). This is missing from the SPCA-DNN results, where smoother variations are typically found. On the other hand, BEM completely misses the shadowing effects of the torso for HRTFs in lower elevations (e.g., see the values at $-90\,\mathrm{deg}$ in Fig. 4) due to missing torsos in the input mesh, where our method and SPCA-DNN reconstruct these regions successfully.

In the high frequency region (e.g., $>10\,\mathrm{kHz}$), BEM and SPCA-DNN methods both show increasing errors, while our method produces results that match with the measurements consistently and thus generates a flat error profile. This is useful because higher frequencies are notoriously harder to simulate and measure, yet by combining them our model learns to maintain error levels. Based on this observation, we expect HRTFs generated by our model to perform consistently in high frequencies and work better in certain scenarios (e.g., high-pitched source in complex scenes).

The level of SD reduction introduced by our method (§5.1) can lead to improvements in perception by reducing the front-back reversal rate and azimuth perception errors, similar to prior work where similar SD reductions were found [30].

5.3. Accuracy and performance

We analyze resilience and error tolerance of our model. The errors are parametrized by performing a series of BEM simulations on 4 different frequencies and for 4 different mesh

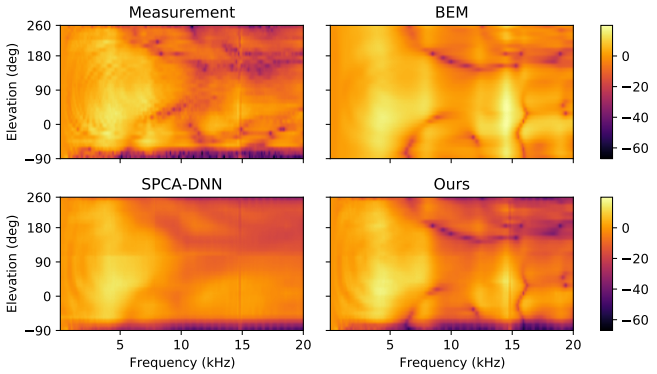

Fig. 4: Comparison of measured HRTF in decibels on the median plane with the reconstructed HRTFs from BEM, SPCA-DNN, and ours.

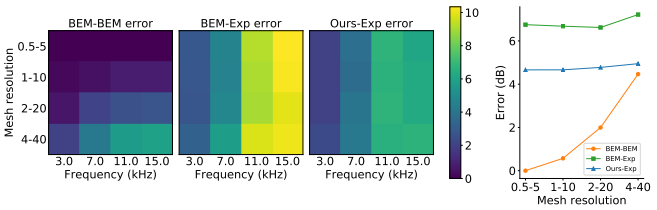

Fig. 5: Convergence analysis. BEM-BEM: BEM errors against the finest BEM resolution; BEM-Exp: BEM errors against measurement; Ours-Exp: errors of our method against measurement. Line plot on the right shows the same errors average over frequencies.

resolutions. The convergence results of this study that covers all 58 subjects in the dataset are shown in Fig. 5. As expected, the BEM solution converges when compared to BEM run on the highest mesh resolution (BEM-BEM). However, there is a persistent discrepancy when compared to measurement (BEM-Exp) because BEM actually converges to a wrong distribution. This explains why our results are not significantly improved even if highest quality BEM simulation is used (Ours-Exp) as $\boldsymbol{E}_{\mathrm{bem}}$ remains similar. Therefore, for applications more tolerant on HRTF accuracy, using our method with lower mesh resolution will save a lot of processing time, since BEM typically scales quadratically with the number of mesh elements. It will be interesting for future work to look into this performance-accuracy tradeoff further and leverage lower resolution BEM input with learning-based corrections.

6. CONCLUSION

We introduced a novel deep learning method that combines measurements and numerical simulations to obtain good-quality personalized HRTFs. Our method outperforms BEM and SPCA-DNN methods in SD for all frequencies. HRTFs generated by our model have salient spectral and spatial structures important to sound localization tasks. Our convergence analysis shows that our method is insensitive to the fidelity of BEM simulation.

7. REFERENCES

[1] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *ICVR*, 2007.

[2] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Applied Sciences*, 2020.

[3] E. M. Wenzel and et al., "Localization using nonindividualized head-related transfer functions," *JASA*, 1993.

[4] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *JASA*, 1995.

[5] T. Qu and et al., "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *TASLP*, 2009.

[6] H. Wierstorf and et al., "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *AES Convention*, 2011.

[7] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *JAES*, 2007.

[8] R. Sridhar, J. Tylka, and E. Choueiri, "A database of head-related transfer functions and morphological measurements," in *AES Convention*, 2017.

[9] W. Kreuzer and et al., "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *JASA*, 2009.

[10] F. Ma and et al., "Finite element determination of the head-related transfer function," *Journal of Mechanics in Medicine and Biology*, 2015.

[11] T. Xiao and H. Qing, "Finite difference computation of head-related transfer function for human hearing," *JASA*, 2003.

[12] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*, Elsevier Science, 2005.

[13] H. Cheng and et al., "A wideband fast multipole method for the helmholtz equation in three dimensions," *Journal of Computational Physics*, 2006.

[14] S. Prepeliță and et al., "Influence of voxelization on finite difference time domain simulations of head-related transfer functions," *JASA*, 2016.

[15] Nikunj Raghuvanshi, Rahul Narain, and Ming C Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *TVCG*, 2009.

[16] J. Wang and D. L. James, "Kleinpat: Optimal mode conflation for time-domain precomputation of acoustic transfer," *TOG*, 2019.

[17] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, 2008.

[18] C. J. Chun and et al., "Deep neural network based HRTF personalization using anthropometric measurements," in *AES Convention*, 2017.

[19] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual HRTFs based on spatial principal component analysis," *TASLP*, 2020.

[20] M. Zhang and et al., "Distance-dependent modeling of head-related transfer functions," in *ICASSP*, 2019.

[21] M. Zhang, X. Wu, and T. Qu, "Individual distance-dependent HRTFs modeling through a few anthropometric measurements," in *ICASSP*, 2020.

[22] V. R. Algazi and et al., "The CIPIC HRTF database," in *WASPAA*, 2001.

[23] R. Bomhardt and et al., "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proc. Mtgs. Acoust.*, 2016.

[24] F. Brinkmann and et al., "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *JAES*, 2019.

[25] G. Yu, R. Wu, Y. Liu, and B. Xie, "Near-field head-related transfer-function measurement and database of human subjects," *JASA*, 2018.

[26] B. Xie, "Recovery of individual head-related transfer functions from a small set of measurements.," *JASA*, 2012.

[27] X. Zeng, S. Wang, and L. Gao, "A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures," *JSV*, 2010.

[28] H. Ziegelwanger and et al., "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," in *ICSV*, 2015.

[29] P. Mokhtari and et al., "Computer simulation of KEMAR's head-related transfer functions: Verification with measurements and acoustic effects of modifying head shape and pinna concavity," in *Principles and Applications of Spatial Hearing*. 2011.

[30] K. J. Fink and L. Ray, "Tuning principal component weights to individualize HRTFs," in *ICASSP*, 2012.