# DISTANCE-DEPENDENT MODELING OF HEAD-RELATED TRANSFER FUNCTIONS

*Mengfan Zhang, Yue Qiao, Xihong Wu, Tianshu Qu*

Key Laboratory on Machine Perception (Ministry of Education), Speech and Hearing Research Center,
Peking University, Beijing, China, qutianshu@pku.edu.cn

## ABSTRACT

In this paper, a method for modeling distance dependent head-related transfer functions is presented. The HRTFs are first decomposed by spatial principal component analysis. Using deep neural networks, we model the spatial principal component weights of different distances. Then we realize the prediction of HRTFs in arbitrary spatial distances. The objective and subjective experiments are conducted to evaluate the proposed distance model and the distance variation function model, and the results have shown that the proposed model has less spectral distortions than distance variation function model, and the virtual sound generated by the proposed model has better performance in terms of distance localization.

***Index Terms*—** HRTF, SPCA, DNN, DVF, distance localization

## 1. INTRODUCTION

In recent years, spatial auditory display has gained attention in both research area and practical applications. To realize the fidelity and immersive experience in binaural audio reproduction, Head Related Transfer Functions (HRTF) are often used as filters describing the sound transmission from a sound source to the listeners' eardrum. For sound sources with different locations (azimuth, elevation, distance), the characteristics of HRTFs will vary accordingly; however, although the far-field condition of sound sources (i.e. distance larger than 1m) could provide a good approximation to omit the influence of distance in HRTFs, the near-field effects, especially the impacts of head, torso and pinnae, make HRTFs vary drastically [1]. Therefore, it is important to consider HRTFs as distance-dependent and further investigate the relation to varying distances for precise sound localization in near-field.

Based on complex behaviors in near-field, there have been measurements to obtain distance-dependent HRTFs [2], and algorithms and methods are proposed to model HRTFs in near-field. Researchers have examined binaural cues for near-field sound localization and use them to synthesize HRTFs, such as the interaural level difference (ILD) [3, 4]. An "auditory parallax model" was also proposed for HRTFs in near-field, and then applied for HRTF simulation [5, 6]. Besides, it was also noticed that the magnitude of near-field HRTF plays a role in distance perception [7]. For the computation model of near-field HRTF, a rigid sphere model was applied to simulate sound propagation towards listener's head [8], and the Distance Variation Function (DVF) was then derived as a distance filter to synthesize near-field HRTFs from far-field ones [9–12]. A more specific model with head, neck and torso was also implemented [13]. At the same time, effective interpolation algorithms from measured HRTF databases were also proposed. In [14], a framework using tetrahedral interpolation was given for interpolating HRTF measurements in 3-D locations; Huang et al. [15] proposed a tensor model to represent distance-dependent HRTFs and the interpolation of core tensors provided accurate HRTF prediction.

In our modeling for distance-dependent HRTFs in near-field, we apply the spatial principal component analysis (SPCA) to HRTFs, then the HRTFs can be represented by a weighted combination of spatial principal components (shortened as SPCs) [16]. Through the deep neural network (DNN) training, the spatial principal component weights (shortened as SPCA weights) of different distances are estimated. After that, we can combine the SPCs and the SPCA weights to reconstruct distance-dependent HRTFs.

The rest of the paper is organized as follows. In Section 2, the spatial principal component analysis is discussed. In Section 3, the distance modeling method based on SPCA is described. Section 4 gives both the objective and subjective evaluation of the proposed method. In section 5, the conclusion is presented.

## 2. SPATIAL PRINCIPAL COMPONENT ANALYSIS

The traditional PCA method is generally used in the time or frequency domain of HRTFs [17, 18]. In contrast to traditional PCA models, SPCA is applied to the spatial domain. The high spatial resolution HRTFs are decomposed into the combination of SPCs and the SPCA weights [16].

$$HRTF(\theta,\varphi,f,s)=\sum_{q}d_q(f,s)W_q(\theta,\varphi)+H_{av}(\theta,\varphi) \quad (1)$$

This work was supported by the National Natural Science Foundation of China (No. 11590773, No.61175043), the State Key Laboratory of Robotics (2018-O09), and the High-performance Computing Platform of Peking University.

where $W_q$ and $H_{av}$ are the SPCs and mean spatial function, which depend only on the source direction. $\varphi$ is the elevation angle, and $\theta$ is the azimuth angle, $d_q$ is SPCA weights which vary as functions of frequency $f$ and individual $s$.

To model distance-dependent HRTFs, we use $c_q(r)$ to predict the relationship between the SPCA weights of different distances. Then Eq. (1) can be rewritten as follows.

$$\begin{aligned} &HRTF(\theta,\varphi,r,f,s) \\ &\quad = \sum_q d_{q,r_0}(f,s)c_q(r)W_q(\theta,\varphi) + H_{av}(\theta,\varphi,r) \end{aligned} \tag{2}$$

where $d_{q,r_0}(f,s)$ is the SPCA weights of distance $r_0$, and $H_{av}$ is the function of source direction and distance.

The HRTFs used in this paper are derived from the PKU&IOA database [2], which contains HRTFs of $S = 2$ subjects, KEMAR with large and small ear, measured in $D = 793$ directions of $R = 8$ distances. The number of frequency points is 1024, and the sampling rate is 65536 Hz. After applying SPCA to the HRTFs, we select the preceding $Q$ SPCs. In this paper, the first 100 SPCs are selected, and 92.27% of the total variability can be attained after the reconstruction [19, 20].

## 3. DISTANCE-DEPENDENT MODELING OF HRTFS

### 3.1. Outline

Fig. 1 depicts the framework of distance-dependent modeling of HRTFs. The SPCA weights and the mean spatial function of different distances are modeled respectively. After that, the selected SPCs, the corresponding SPCA weights and the mean spatial function can be used to recover HRTF magnitude in each sampled direction of arbitrary distances.

The phase of HRTFs is reconstructed based on two assumptions that HRTFs are "minimum-phase" functions and the frequency dependence of ITD is of no perceptual relevance [17]. Therefore, the minimum-phase reconstruction method is used to generate mono HRIRs [17], and then the binaural HRIRs are predicted.

### 3.2. Preprocessing

The raw HRIRs are preprocessed as follows.

Firstly, transform the HRIRs into the frequency domain. Fourier transformation is applied to the raw HRIRs to obtain the HRTFs.

Secondly, transform the HRTFs into a logarithmic scale. The amplitude scale of the HRTFs are linear, and a logarithmic scale is much closer to our auditory perception [21]. Compute the base 10 log-magnitude responses of HRTFs, which is denoted as $HRTF_{log}$.

$$\begin{aligned} &HRTF_{log}(\theta,\varphi,r,f,s) \\ &\quad = 20log_{10}(|HRTF(\theta,\varphi,r,f,s)|) \end{aligned} \tag{3}$$

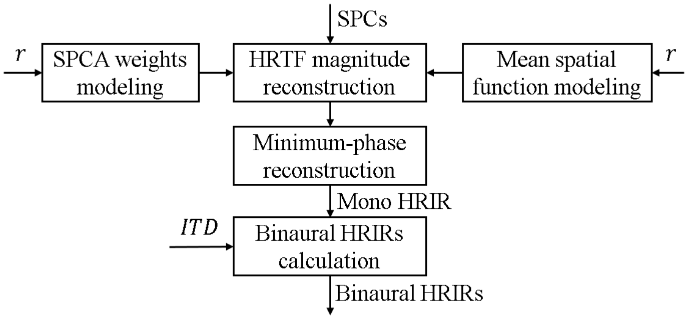
SPCs
r
SPCA weights modeling
HRTF magnitude reconstruction
Mean spatial function modeling
r
Minimum-phase reconstruction
Mono HRIR
ITD
Binaural HRIRs calculation
Binaural HRIRs

**Fig. 1**. The framework of distance-dependent HRTF modeling.

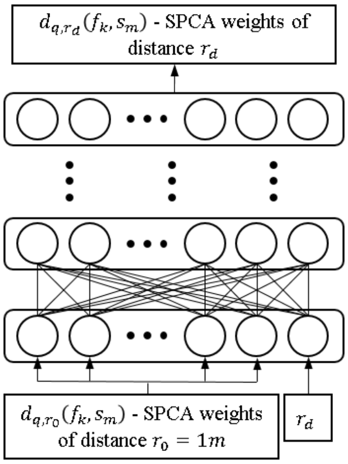
$d_{q,r_d}(f_k, s_m)$ - SPCA weights of distance $r_d$
$d_{q,r_0}(f_k, s_m)$ - SPCA weights of distance $r_0 = 1m$
$r_d$

**Fig. 2**. The architecture of DNN for predicting the SPCA weights of different distances.

Thirdly, the mean of all the logarithmic HRTFs is subtracted from each $HRTF_{log}$, and the result, i.e. $HRTF_{log\Delta}$, is decomposed by the SPCA.

$$\mu(r,f) = \frac{1}{D \times S}\sum_s\sum_\theta\sum_\varphi HRTF_{log}(\theta,\varphi,r,f,s) \tag{4}$$

$$\begin{aligned} &HRTF_{log\Delta}(\theta,\varphi,r,f,s) \\ &\quad = HRTF_{log}(\theta,\varphi,r,f,s) - \mu(r,f) \end{aligned} \tag{5}$$

Finally, calculate the mean spatial function, i.e. $H_{av}$, which is the mean of $HRTF_{log\Delta}$ over frequencies and subjects.

$$H_{av}(\theta,\varphi,r) = \frac{1}{NS}\sum_{s=1}^{S}\sum_{f=1}^{N} HRTF_{log\Delta}(\theta,\varphi,r,f,s) \tag{6}$$

### 3.3. Modeling of SPCA weights

Fig. 2 shows the architecture of DNN for the SPCA weights modeling. The input of DNN is the SPCA weights $d_{q,r_0}(f_k, s_m)$

of distance $r_0 = 1m$ and the target distance $r_d$. The ground truth is the SPCA weights $d_{q,r_d}(f_k, s_m)$ of the target distance $r_d$. $s_m(m = 1, 2)$ and $f_k(k = 1, 2, ..., 200)$ are the KEMARs and the frequency points, respectively. There are $S \times N \times R$ sets of SPCA weights, and we split it into the training set, validation set and test set. Each set contains the SPCA weights of all the distances. The mean and variance of the test set and validation set were normalized using the training set statistics to have zero mean and unit variance. Each DNN is set five layers which had a better prediction performance. Both the activation function and the output function are set hyperbolic tangent and the learning rate is 0.001.

The reconstruction error is used to test the DNN modeling:

$$e_{d,r_d}(f_k, s_m) = \frac{\sum\limits_q |\hat{d}_{q,r_d}(f_k, s_m) - d_{q,r_d}(f_k, s_m)|}{\sum\limits_q |d_{q,r_d}(f_k, s_m)|} \qquad (7)$$

where $\hat{d}_{q,r_d}(f_k, s_m)$ is estimated by DNN model, and $e_{d,r_d}$ $(f_k, s_m)$ is the reconstruction error of the SPCA weights for KEMAR $s_m$ at the frequency $f_k$ and distance $r_d$. The calculated $e_{d,r_d}$ in all the frequencies and individuals are smoothly distributed in 0.2 to 0.3.

### 3.4. Modeling of mean spatial function

The modeling of the mean spatial functions is similar to the prediction of the SPCA weights. This model is also based on DNN. The input of DNN is the $H_{av}(\theta, \varphi, r_0)$ of distance $r_0 = 1m$ and the target distance $r_d$. The ground truth is the $H_{av}(\theta, \varphi, r_d)$ of the target distance $r_d$. To guarantee the variability of the test data, we make all the training set, validation set and test set uniformly distributed in space. The mean and variance of the test set and validation set were normalized using the training set statistics to have zero mean and unit variance. The modeled DNN are set five layers, the activation function and the output function are set hyperbolic tangent, and the learning rate is 0.001.

After training the DNNs, we can predict $H_{av}(\theta, \varphi, r_d)$ in arbitrary spatial distance $r_d$. Mean square error (MSE) is used to calculate the reconstruction error of the mean spatial functions. And the result $e_H$ is equal to 0.195.

$$e_H = \frac{1}{D} \sum_\theta \sum_\phi (\hat{H}_{av}(\theta, \phi) - H_{av}(\theta, \phi))^2 \qquad (8)$$

### 3.5. Modeling of mean

Due to the magnitude of HRTF is attenuated when the spatial distance becomes larger, we build a model to estimate the mean of HRTF in different distances, which is calculated in Eq.(4). The modeling of the mean is similar to the prediction of the SPCA weights and the mean spatial functions. This model is based on DNN as well. The input of DNN is the $\mu(r_0, f_k)$ of distance $r_0 = 1m$ and the target distance $r_d$. The ground truth is the $\mu(r_d, f_k)$. There are $N \times R$ sets of means, and we split it into the training set, validation set and test set. Each set contains $\mu$ in all the distance. The mean and variance of the test set and validation set were normalized using the training set statistics to have zero mean and unit variance. The modeled DNN are set five layers, the activation function is hyperbolic tangent, and the output function is set linear function for high frequency attenuation. The learning rate is 0.006.

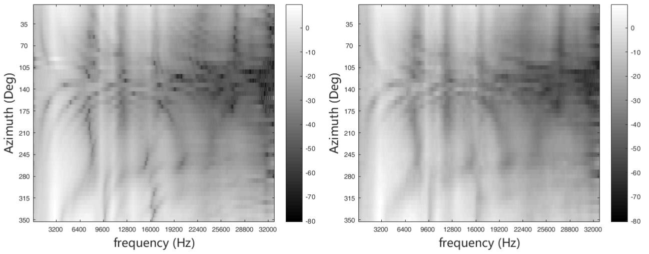

**Fig. 3**. Comparison of the log-amplitude of the real (left) and the reconstructed (right) HRTFs of KEMAR with small ear.

After training the DNNs, the mean in desired distance for each frequency bin can be predicted. The MSE of the estimated mean is 0.42.

### 3.6. Recovery of HRIRs

To reconstruct the HRIRs, we first model the SPCA weights, the mean spatial functions and the means, respectively. The distance is introduced into the input layer of the DNN models, so that we can predict these parameters in arbitrary spatial distances efficiently. The HRTF magnitude of different distances can be reconstructed by solving the Eq. 2. Fig. 3 shows comparison of the real and the reconstructed HRTFs of KEMAR with small ear. It depicts the horizontal plane at the distance of 50 cm. The similarity of their global shapes indicates that the reconstructed HRTF magnitudes are desired.

The minimum phase reconstruction method is then employed to the HRTF magnitudes to generate the mono HRIRs [17]. Since ITD only varies slightly when the source moves from far-field to near-field [1, 22], so we consider the ITDs in arbitrary spatial distances are equal to the ITDs in distance $r_0 = 1m$. Therefore, the binaural HRIRs in arbitrary spatial distances can be recovered.

## 4. EVALUATION EXPERIMENTS

To evaluate the effectiveness of our proposed method, we carried out objective and subjective experiments for HRTFs predicted by the proposed model and HRTFs estimated by DVF model [9].

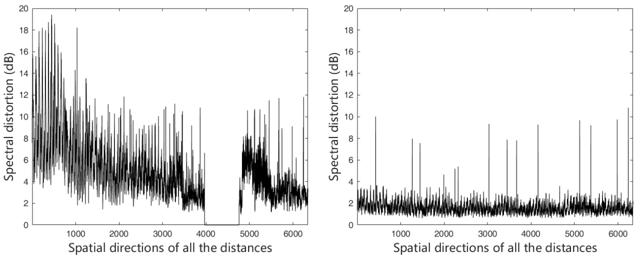

**Fig. 4**. Comparison of the SD between DVF model (left) and the proposed model (right).

### 4.1. Objective experiments

The spectral distortion (SD) is used as an objective evaluation metric between modeled and measured HRTF data.

$$SD = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(20lg\frac{|H(f_k)|}{|\hat{H}(f_k)|})^2} \tag{9}$$

where $H(f_k)$ is the magnitude response of the measured HRTF from the PKU&IOA database, $\hat{H}(f_k)$ is the magnitude response of the estimated HRTF.

The SD of the reconstructed HRTFs of KEMAR with small ear is shown as Fig. 4. Note that our proposed model selecting the first 100 SPCs results in an average SD of 1.56dB, so this is the reason that the SD of all spatial locations in our proposed model is larger than approximately 1 dB. SD equals zero in some spatial directions of DVF model is because we use HRTFs at 1 meter to estimate those at other spatial distances. The average SD of the proposed model in all the sampled directions and distances is 2.41dB, and the average SD of DVF model is 5.12 dB. This demonstrates that the proposed model is superior to the DVF model.

### 4.2. Subjective experiments

For the subjective evaluation, twelve subjects (10 male 2 female, age from 21 to 29) with normal hearing took part in the experiments. All experiments were performed in a sound booth (Background noise level : 20.9 dBA).

The stimuli in this experiment was a train of eight 250-ms bursts of Gaussian noise (20-ms cosine-squared onset-offset ramps), with 300 ms of silence between the bursts. The HRIRs of six distances, 20, 30, 50, 80, 120 and 160 cm, are generated by the proposed model and the DVF model, respectively. Note that the HRIRs of two distances, 80 and 120 cm, are not contained in the PKU&IOA database. Then the stimulus are filtered by the HRIRs produced by the two models to create two kinds of sounds. A total of two experiments are performed, and each experiment corresponds to a kind of sound generated by one method. The subject should tell the exact distance of each sound he/she heard during the experiment. Before each experiment, the subject is trained using the test sound of other six distances, 20, 40, 75, 100, 130 and 160 cm. Through listening these sounds, the subject can build up the distance perception for this kind of virtual sound. After that, eighteen binaural sounds are randomly played to the subject by a Sennheiser HD 650 headphone. The eighteen sounds contain sounds at six distances and each distance appears three times. The subject can listen to one sound for many times until he/she can tell the exact direction.

Figure 5 shows the results of the distance localization experiments of all the twelve subjects. The judgments are plotted as a function of the coordinates of the targets. The left column and the right column depict the judgments using the DVF model and the proposed model respectively. There are 216 judgments shown in each panel, corresponding to the thirty-six judgments made to each of six binaural sounds. Each solid circle in the image represents the amount of judgments for a target angle. As the legend illustrates, the size of the solid circle represents the number of judgments. For example, the biggest solid circle in the legend represents thirty judgments for a fixed distance. The correct answer is on the diagonal line. The localization performance of the proposed model is better than the DVF model. The proposed model has more solid circles with bigger sizes falling upon the diagonal line. This means a higher precision of localization and also indicates the proposed model can better reconstruct the HRTFs in arbitrary spatial distances.

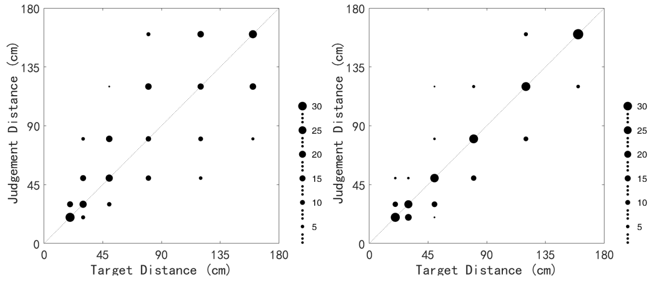

**Fig. 5**. Judged distance versus target distance of all subjects using the DVF model (left) and the proposed model (right). Each solid circle represents the amount of judgments for a target angle. Size of circle is increased with the judgments.

## 5. CONCLUSION

The paper proposed the distance-dependent HRTF modeling method based on spatial principal component analysis. By modeling the SPCA weights, mean spatial function and mean using DNN respectively, we reconstruct the HRTFs of arbitrary spatial distances. The objective and subjective experiments are carried out to evaluate the HRTFs generated by the proposed model and DVF model. Results show our proposed model is superior than the DVF model in both objective and subjective experiments.

## 6. REFERENCES

[1] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[2] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1124–1132, 2009.

[3] D. S. Brungart, "Preliminary model of auditory distance perception for nearby sources," *Computational models of auditory function*, pp. 83–96, 2001.

[4] T. Fukuda, T. Horiuchi, H. Hokari, and S. Shimada, "Relative distance perception by manipulating the ILD of HRTFs," *Acoustical Science and Technology*, vol. 24, no. 5, pp. 325–326, 2003.

[5] Y. Suzuki, S. Takane, H. Kim, and T. Sone, "A modeling of distance perception based on an auditory parallax model," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 3083–3083, 1998.

[6] H. Kim, Y. Suzuki, S. Takane, and T. Sone, "Control of auditory distance perception based on the auditory parallax model," *Applied Acoustics*, vol. 62, no. 3, pp. 245–270, 2001.

[7] Y. Liu and B. Xie, "Auditory discrimination on the distance dependence of near-field head-related transfer function magnitudes," in *Proceedings of Meetings on Acoustics ICA2013*. ASA, 2013, vol. 19, p. 050048.

[8] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[9] A. Kan, C. Jin, and A. van Schaik, "Distance variation function for simulation of near-field virtual auditory space," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006, vol. 5, pp. 325–328.

[10] A. Kan, C. Jin, and A. van Schaik, "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2233–2242, 2009.

[11] S. Spagnol, E. Tavazzi, and F. Avanzini, "Distance rendering and perception of nearby virtual sound sources with a near-field filter model," *Applied Acoustics*, vol. 115, pp. 61–73, 2017.

[12] Jiawang Xu, Xiaochen Wang, Maosheng Zhang, Cheng Yang, and Ge Gao, "Binaural sound source distance reproduction based on distance variation function and artificial reverberation," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 101–111.

[13] Z. Chen, G. Yu, B. Xie, and S. Guan, "Calculation and analysis of near-field head-related transfer functions from a simplified head-neck-torso model," *Chinese Physics Letters*, vol. 29, no. 3, pp. 034302, 2012.

[14] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *The Journal of the Acoustical Society of America*, vol. 134, no. 6, pp. EL547–EL553, 2013.

[15] Q. Huang, K. Liu, and Y. Fang, "Tensor modeling and interpolation for distance-dependent head-related transfer function," in *2014 12th International Conference on Signal Processing (ICSP)*. IEEE, 2014, pp. 1330–1334.

[16] B. Xie, "Recovery of individual head-related transfer functions from a small set of measurements.," *Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 282–294, 2012.

[17] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637, 1992.

[18] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in *The Workshop on Hands-Free Speech Communication & Microphone Arrays*, Edinburgh, UK, 2011, pp. 213–218.

[19] 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, 2008.

[20] X. Zeng, S. Wang, and L. Gao, "A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures," *Journal of Sound & Vibration*, vol. 329, no. 19, pp. 4093–4106, 2010.

[21] J. O. Smith, *Techniques for digital filtering design and system identification with the violin*, Ph.D. thesis, CCRMA, Stanford, 1983.

[22] M. Otani, T. Hirahara, and S. Ise, "Numerical study on source-distance dependency of head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3253–3261, 2009.