Audio Engineering Society

Convention Paper

Presented at the 150th Convention
2021 May 25–28, Online

Individualized HRTF-based Binaural Renderer for Higher-Order Ambisonics

Mengfan Zhang[1], Tianyi Guan[1], Lianwu Chen[2], Tianxiao Fu[2], Dan Su[2], and Tianshu Qu[1]

[1]*Key Laboratory on Machine Perception (Ministry of Education), Speech and Hearing Research Center, Peking University, China*
[2]*Tencent AI Lab, Shenzhen, China*

Correspondence should be addressed to Tianshu Qu (qutianshu@pku.edu.cn)

ABSTRACT

Ambisonics is a promising spatial sound technique in augmented and virtual reality. In our previous study, we modeled the individual head-related transfer functions (HRTFs) using deep neural networks based on spatial principal component analysis. This paper proposes an individualized HRTF-based binaural renderer for the higher-order Ambisonics. The binaural renderer is implemented by filtering the virtual loudspeaker signals using individualized HRTFs. We perform subjective experiments to evaluate generic and individualized binaural renderers. Results show that the individualized binaural renderer has front-back confusion rates that are significantly lower than those of the generic binaural renderer. Therefore, we validate that using individualized HRTFs to convolve with those virtual loudspeaker signals to generate virtual sound at an arbitrary spatial direction still performs better than those using generic HRTFs. In addition, by measuring or modeling individual's HRTFs in a small set of directions, our proposed binaural renderer system effectively predict individual's HRTFs in arbitrary spatial directions.

1 Introduction

Ambisonics is a versatile surround sound recording and reproduction technique. The binaural rendering process assigns Ambisonics signals to headphones, which produce the target sound field perceived by listeners. Some works contribute to this topic. Noisternig et al. [1] presented a computationally efficient 3D real time rendering engine for binaural sound reproduction via headphones. Rumori [2] designed a Girafe, which is a versatile, modular software system for projects using Ambisonics or the binaural virtual Ambisonics approach. Tylka and Choueiri [3] implemented a toolkit for ambisonics-to-binaural renderers for the ambiX binaural plug-in, which enables the user to generate custom binaural rendering configurations. Hold et al. [4] analyzed the effects of truncating the spherical harmonics representation of a binaurally rendered sound field and proposed a method that reduces coloration. Zhang et al. [5] improved the performances of higher-order Ambisonics (HOA) by learning features of sound fields with the generative adversarial network (GAN).

Head-related transfer function (HRTF) is an acoustic function for sound waves modified by the human structure, such as the head, ears, and torso. A prevalent method to implement binaural rendering uses virtual loudspeakers to combine the Ambisonics and HRTFs to construct the sound field. However, localization error

exists during the perception due to the accuracy limitation of Ambisonics and the low performance of the non-individualized HRTFs. Nevertheless, directly measuring HRTFs for each listener is complicated. Thus, modeling individual HRTFs is very important and enhances the performances of the binaural rendering.

More researchers have modeled individual HRTFs. Brown et al. [6] separated the effects of different physiological structures on HRTF, modeling each part with a low-order sub-filter and combining all sub-filters to represent HRTF. Middlebrooks [7] used the frequency scaling method, assuming that the HRTF spectral characteristics of diverse listeners are similar but the corresponding frequencies of spectral characteristics are different. Through the frequency scaling method, Middlebrooks obtained new subject's HRTFs. Zotkin et al. [8] selected the HRTF data from a subject whose anthropometric parameters were closest to the new subject. Jin et al. [9] separately applied the principal component analysis (PCA) to both the HRTF amplitude spectrum and the anthropometric parameters and then constructed a linear mapping from the PCA weights of the anthropometric parameters to the PCA weights of HRTFs. Hu et al. [10] used back-propagation artificial neural networks to map the PCA weights of HRTFs to the selected anthropometric parameters. Chun [11] used the deep neural network (DNN) to map the anthropometric parameters to the head-related impulse response (HRIR). Zhang [12, 13, 14] used DNN models based on spatial principal component analysis (SPCA) to predict HRTFs in arbitrary spatial directions and distances.

In this paper, we propose an individualized HRTF based binaural renderer using our previous work in modeling individual HRTFs [12]. We aim to validate the individualized binaural renderer performs better than generic binaural renderer after decoding the HOA signals to loudspeaker signals and then convolving loudspeaker signals with HRTFs. In addition, though generating individualized HRTFs in a small set of directions, we build a system to generate individualized virtual sound in arbitrary spatial directions. Our paper is organized as follows. In Section 2, we describe the basic HOA theory and the binaural rendering process. In Section 3, we model individual HRTFs using DNN based on SPCA. In Section 4, we present our subjective experiments and analyze the results. In section 5, we present our conclusion.

2 HOA and Binaural Renderer Basis

HOA is a spatial sound technique based on the superposition of spherical harmonic functions. The sound field will be encoded into HOA signals with a certain order, and then the decoding process transforms HOA signals to loudspeaker signals, which could be directly fed into real loudspeakers, or be treated as sound sources spatially distributed in virtual scenes. In the latter, a binaural renderer combines these signals into a two-channel binaural signal for headphones.

2.1 HOA Principles

Considering a specific HOA signal B_n^m of order n and degree m, the spherical harmonic function Y_n^m, [15, 16], is described as

$$Y_n^m(\theta,\varphi) = \sqrt{(2n+1)(2-\delta_{0,m})\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(sin\varphi) \times \begin{cases} \sin(-m\theta), & if\ m<0 \\ \cos(m\theta), & if\ m\geq 0 \end{cases} \quad (1)$$

where $P_n^m(\sin\varphi)$ is the associated Legendre functions, and $\delta_{0,m}$ is the Kronecker delta function. θ and φ refer to the azimuth and elevation in a head-related spherical coordinate system [17]. The origin of the coordinate system is the midpoint of a line drawn between the upper margins of the entrances to the two ear canals.

The sound field with no interior sources is expressed in the Fourier-Bessel series:

$$p(kr,\theta,\varphi) = \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} B_n^m Y_n^m(\theta,\varphi), \quad (2)$$

where $j_n(kr)$ is the spherical Bessel function of the first kind, and B_n^m is the spherical harmonic component signal (HOA signal). k is the wave number, r is the radius, and p is the sound pressure. Since the practical HOA has a finite order, n is limited to the maximum order N.

The HOA signal B_n^m is derived from the encoding process, and the usual ways of constructing sound fields to be encoded include virtual sound field simulation and real HOA recordings. Then, the decoding of HOA signals is processed. Moreau et al. gave a more specific description of HOA encoding and decoding processes [16].

2.2 HOA Decoder

The decoding process converts HOA signals into multiple loudspeaker signals. Consider a sound source q emitting a plane wave g from the direction (θ_q, φ_q), which is the far field condition of the sound source. The sound pressure p is approximated in the N-order spherical harmonic expansion as follows:

$$p = g \sum_{n=0}^{N} i^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta_q, \varphi_q) Y_n^m(\theta, \varphi). \quad (3)$$

Hence, the HOA signal of a plane wave source conveying signal g with order n and degree m is expressed as follows:

$$B_n^m = g Y_n^m(\theta_q, \varphi_q). \quad (4)$$

Then, the target HOA signals $\mathbf{B} = \left[B_0^0, B_1^{-1}, \ldots, B_n^m\right]$ are decoded into several loudspeaker signals. Suppose there are L loudspeakers in space with directions $(\theta_1, \varphi_1), \ldots, (\theta_L, \varphi_L)$, each conveying a plane wave signal s_i. $\mathbf{B}$ is estimated with the superposition of the HOA components of signals $\mathbf{g} = [g_1, g_2, \ldots, g_L]$:

$$\boldsymbol{B} = \boldsymbol{g} \cdot \boldsymbol{D} \quad (5)$$

From Eq.4, the matrix $\mathbf{D}$ has the following form:

$$\boldsymbol{D} = \begin{bmatrix} Y_0^0(\theta_1, \varphi_1) & \ldots & Y_N^N(\theta_1, \varphi_1) \\ Y_0^0(\theta_2, \varphi_2) & \ldots & Y_N^N(\theta_2, \varphi_2) \\ \vdots & \ddots & \vdots \\ Y_0^0(\theta_L, \varphi_L) & \ldots & Y_N^N(\theta_L, \varphi_L) \end{bmatrix}. \quad (6)$$

The signals of loudspeakers are derived as

$$\boldsymbol{g} = \boldsymbol{B} \cdot \boldsymbol{C}, \quad (7)$$

where the decoding matrix $\boldsymbol{C}$ is the pseudo-inverse of the matrix $\boldsymbol{D}$:

$$\boldsymbol{C} = (\boldsymbol{D}^t \boldsymbol{D})^{-1} \boldsymbol{Y}^t. \quad (8)$$

Especially noteworthy is that this solution is often imprecise, and its approximation error from the pseudo-inverse option is dependent on the spatial distribution of the loudspeakers, more specifically, on the discrete orthogonality of spherical harmonics. The algorithm performs better when the loudspeakers are evenly distributed, for example, in a polyhedron pattern. Other patterns such as a pentakis-dodecahedron with 32 positions were also examined [16]. Also, the maximum order is determined by the number of loudspeakers. To achieve an N-order HOA decoding process, the number of loudspeakers should be larger than $(N+1)^2$ [16].

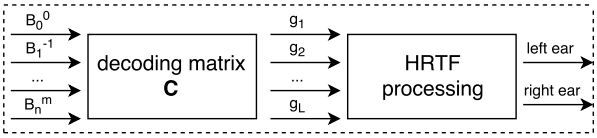

Fig. 1: The scheme of conversion from HOA signals to binaural signals.

2.3 Binaural Renderer

In the renderer, the loudspeaker signals from decoding process are fed into virtual loudspeakers evenly placed on a spherical surface, and then the binaural signals are created by accumulating the virtual loudspeakers' signals, which are filtered by the HRTFs corresponding to the virtual loudspeakers' spatial directions. The binaural signals to the left and right channels of headphone are obtained:

$$g_{left} = \sum_{l=1}^{L} g_l * HRIR_{left}, \quad (9)$$

$$g_{right} = \sum_{l=1}^{L} g_l * HRIR_{right}, \quad (10)$$

where g_{left} and g_{right} identify the binaural signals. HRIR is the time domain expression of HRTF, and the symbol $*$ is the convolution operator.

Fig.1 describes the conversion from HOA signals to binaural signals.

3 Modeling of Individual HRTFs

HRTF is an acoustic function of a sound signal's frequency, direction, and distance and an individual's morphology. The HRTF represents the acoustic cues for a sound at a certain position, which are processed by the auditory system. In binaural rendering, non-individual HRTFs lead to some perception errors such as in-head localization, front-back confusion, or breakdown of elevation discrimination ability [18]. Thus, we use the individualized HRTFs to improve the performance of the binaural renderer. We derive the raw HRIRs from the CIPIC database [19]. The HRTFs are modeled based on SPCA using deep neural networks [12].

Fig. 2 depicts the framework of individual HRTF modeling. The SPCA weights, the spatial principal components (SPCs), and the H_{av} are obtained by decomposing HRTFs using SPCA. Then, those parameters and ITDs

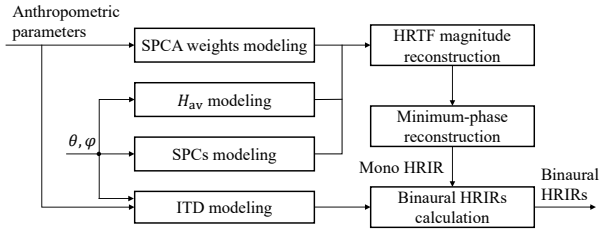

Fig. 2: The framework of individual HRTF modeling [12].

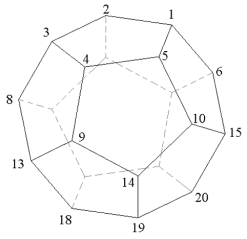

Fig. 3: Twenty vertices of a regular dodecahedron.

are respectively modeled . The SPCA weights are obtained by modeling the individual's morphology since the SPCA weights vary as functions of anthropometric parameters and frequency. Eight anthropometric parameters, head width, head depth, shoulder width, cavum concha height, cavum concha width, fossa height, pinna height, and pinna width, are selected to represent the human morphology. Thus, the SPCA weights for any individual outside the database are estimated from the individual's eight anthropometric parameters. Those anthropometric parameters are captured by taking pictures. Since the SPCs and the H_{av} depend only on the source direction, we model them using DNNs to predict new values in arbitrary spatial directions. Since the ITDs are influenced by both anthropometric parameters and the spatial directions, we use head dimensions to model a new individual's ITDs with arbitrary spatial directions by training DNNs. Accordingly, HRTF magnitudes of arbitrary spatial directions are recovered using the predicted SPCs, SPCA weights, and H_{av}. Then the minimum-phase reconstruction method is used to generate mono HRIRs [20]. Finally, binaural HRIRs are obtained using estimated ITDs and the corresponding left and right mono HRIRs.

In summation, through taking some photos of an individual, we capture a quantity of anthropometric parameters to reconstruct an individual's binaural HRIRs with arbitrary spatial directions.

4 Subjective Experiments

4.1 Settings

We perform subjective experiments to evaluate both the generic renderer and the individualized renderer. The generic HRIR set is chosen to be the CIPIC KEMAR with small ears. For the HOA decoding process, the 3-order HOA is performed, and the virtual loudspeakers are positioned at the twenty vertices of a regular dodecahedron, as is shown in Fig. 3. The HOA signals are decoded to virtual loudspeaker signals with the plane-wave assumption, and then each loudspeaker signal is convoluted with generic HRIR and individualized HRIR with the same spatial direction. The 20 directions of HRIR are chosen from the CIPIC database to be the nearest with the directions of 20 vertices in the dodecahedron. Note that the azimuth angle and the elevation angle in the CIPIC database are measured in a head-centered interaural-polar coordinate system. Thus, we convert the directions of 20 vertices in the dodecahedron into an interaural-polar coordinate system and then choose the closest directions. The transformation formulas are as follows:

$$\begin{aligned} \sin(\theta') &= \sin(\theta)\cos(\varphi), \\ \tan(\varphi') &= \tan(\varphi)\big/\cos(\theta)\ , \end{aligned} \tag{11}$$

where θ and φ refer to the azimuth angle and the elevation angle in the head-related spherical coordinate system respectively, and θ' and φ' are the azimuth angle and the elevation angle in the interaural-polar coordinate system respectively. Next , we obtain two different binaural renderers, the generic renderer and the individualized renderer.

4.2 Procedure

Our experiments compare the azimuth localization performance of the individualized renderer and the generic renderer for headphones. The stimulus in this experiment is 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, and the sampling rate is 44.1kHz. The stimulus, given a certain direction, is first encoded to 3-order HOA signals, on which the decoding and rendering processes generate binaural signals that are played to the subjects.

Two azimuth localization experiments are performed at two elevations (0 and 15 degrees). Each experiment randomly includes two tests for spatial localization. One tests the individualized renderer; the other tests the generic renderer. For each test, 36 binaural signals (twelve azimuths each appearing three times) are randomly played by a Sennheiser HD 650 headphone through a RME FireFace UCX sound card. The twelve azimuths are 0, 30 60, 90, 120, 150, 180, 210, 240, 270, 300, and 330 degrees. Subjects give the exact direction based on the sound they heard through a graphical user interface (GUI) on a computer. Each experiment includes a five-minute break. Twelve subjects (9 male, 3 female, ages 25 to 37) with normal hearing participate. All the experiments are performed in a sound booth.

4.3 Results

Fig. 4 shows the results of the localization experiments for all twelve subjects at two elevation angles. The judgments are plotted as a function of the targets' coordinates. The left and right columns depict the judgments using the generic and SPCA renderers. Each panel shows 432 judgments, corresponding to the 36 judgments for the twelve binaural sounds. The localization performance of the SPCA renderer is better than that of the generic renderer since the judgments are more closely gathered near the diagonal line.

The averages of all the subjects' localization experiments are shown in Table 1. The average front-back confusion rates of the individualized renderer are 8.8% and 10.4% smaller than the front-back confusion rates of the generic renderer at elevations of 0 and 15 degrees respectively. The average angles of error of the two renderers at both elevations are similar. Bartlett's test shows the variances are equal in all the conditions ($p > 0.05$). T-tests show the front-back confusion rate of the individualized renderer is significantly lower than that of the generic renderer at both the elevation of 0 ($p < 0.05$) and 15 degrees ($p < 0.05$), and the difference in the angle of error of the two renderers is insignificant at both the elevation of 0 ($p = 0.37$) and 15 degrees ($p = 0.58$). Therefore, the performance of the individualized renderer is better than that of the generic renderer.

5 Conclusion

In this paper, we propose an individualized HRTF-based binaural renderer for HOA. First, individualized

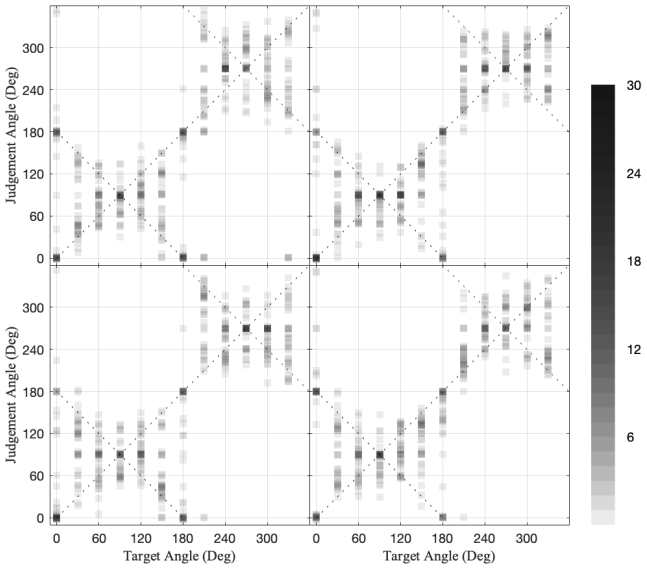

Fig. 4: Judged direction versus target direction of all subjects with the generic renderer (left column) and the individualized renderer (right column) in elevation of 0 degrees (top row) and 15 degrees (bottom row). Two oblique lines with a slope of 135 degrees correspond to the front-back confusions.

HRTFs modeled with DNN based on SPCA are calculated. Second, we use generic HRTFs and individualized HRTFs to filter the virtual loudspeaker signals, then we obtain two kinds of binaural renderers. The subjective experiments are performed to evaluate both the individualized and generic renderers. The subjective experiments' results show that the front-back confusion rates of the individualized renderer are significantly lower than the generic renderer. Therefore, our paper effectively validates that the individualized binaural renderer performs better than generic binaural renderer after decoding the HOA signals to loudspeaker signals and then convolving loudspeaker signals with HRTFs. Our system shows that by decoding HOA signals to loudspeaker signals, we use individualized HRTFs in a small set of directions to effectively generate virtual sound in arbitrary spatial directions.

6 Acknowledgment

This work is supported by the National Key Research and Development Program (No.2019YFC1408501), the National Natural Science Foundation of China

Table 1: The averages of the localization experiments.

Elevation (Deg)	Renderer type	Front-back confusion rate (%)	Angle of error (Deg)
0	Generic	34.5	17.8
	Individualized	25.7	19.4
15	Generic	37.5	19.9
	Individualized	27.1	18.8

(No.61175043, No.61421062), and the High-performance Computing Platform of Peking University.

References

[1] Noisternig, M., Sontacchi, A., Musil, T., and Holdrich, R., "A 3D ambisonic based binaural sound reproduction system," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, p. 21, Alberta, Canada, 2003.

[2] Rumori, M., "Girafe–a versatile ambisonics and binaural system," in *Proceedings of the Ambisonics Symposium*, pp. 1–5, Graz, Austria, 2009.

[3] Tylka, J. G. and Choueiri, E., "A Toolkit for Customizing the ambiX Ambisonics-to-Binaural Renderer," in *Audio Engineering Society Convention 143*, p. 403, New York, USA, 2017.

[4] Hold, C., Gamper, H., Pulkki, V., Raghuvanshi, N., and Tashev, I. J., "Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265, IEEE, 2019.

[5] Zhang, L., Wang, X., Hu, R., Li, D., and Tu, W., "Optimization of sound fields reproduction based Higher-Order Ambisonics (HOA) using the Generative Adversarial Network (GAN)," *Multimedia Tools and Applications*, pp. 1–16, 2020.

[6] Brown, C. P. and Duda, R. O., "A structural model for binaural sound synthesis," *IEEE transactions on speech and audio processing*, 6(5), pp. 476–488, 1998.

[7] Middlebrooks, J. C., "Individual differences in external-ear transfer functions reduced by scaling in frequency." *Journal of the Acoustical Society of America*, 106(3), pp. 1480–1492, 1999.

[8] Zotkin, D. N., Hwang, J., Duraiswaini, R., and Davis, L. S., "HRTF personalization using anthropometric measurements," in *2003 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics*, pp. 157–160, New Paltz, NY, USA, 2003.

[9] Jin, C., Leong, P., Leung, J., Corderoy, A., and Carlile, S., "Enabling individualized virtual auditory space using morphological measurements," in *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pp. 235–238, Sydney, Australia, 2000.

[10] Hu, H., Zhou, L., Ma, H., and Wu, Z., "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, 69(2), pp. 163–172, 2008.

[11] Chun, C. J., Moon, J. M., Lee, G. W., Kim, N. K., and Kim, H. K., "Deep neural network based hrtf personalization using anthropometric measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[12] Zhang, M., Ge, Z., Liu, T., Wu, X., and Qu, T., "Modeling of Individual HRTFs based on Spatial Principal Component Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1), pp. 785–797, 2020.

[13] Zhang, M., Qiao, Y., Wu, X., and Qu, T., "Distance-dependent modeling of head-related transfer functions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 276–280, IEEE, 2019.

[14] Zhang, M., Wu, X., and Qu, T., "Individual Distance-Dependent HRTFS Modeling Through A Few Anthropometric Measurements," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 401–405, IEEE, 2020.

[15] Nachbar, C., Zotter, F., Deleflie, E., and Sontacchi, A., "Ambix-a suggested ambisonics format," in *Ambisonics Symposium, Lexington*, p. 11, 2011.

[16] Moreau, S., Daniel, J., and Bertet, S., "3D sound field recording with higher order ambisonics–Objective measurements and validation of a 4th order spherical microphone," in *Audio Engineering Society Convention 120*, p. 6857, Paris, France, 2006.

[17] Blauert, J., "Spatial Hearing : The Psychophysics of Human Sound Source Localization, Revised Edition," *Mit Press*, 77(9), pp. 926–927, 1997.

[18] Lindau, A., Brinkmann, F., and Weinzierl, S., "Sensory Profiling of Individual and Non-individual Dynamic Binaural Synthesis Using the Spatial Audio Quality Inventory," in *Forum Acusticum*, Krakow, Poland, 2014.

[19] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The cipic hrtf database," in *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, New Platz, NY, USA,, 2001.

[20] Kistler, D. J. and Wightman, F. L., "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, 91(3), p. 1637, 1992.